Monetary Policy Comes of Age: A 20th Century Odyssey

Marvin Goodfriend

In the early 1960s the Federal Reserve (Fed) was little known outside of the financial services industry and university economics departments. Twenty years later Fed Chairman Paul Volcker was one of the most recognized names in American public life. Now hardly a week goes by when the Fed is not featured prominently in the business news. The Fed was thrust into the limelight in the intervening years when the public came to associate it with inflation-fighting policy actions that raised interest rates and weakened economic activity. Even though inflation has been held in check since the mid-1980s, the public remains acutely aware of Fed policy today.

Monetary economists and central bankers alike now understand that effective monetary policy must be built on a consistent commitment to low inflation. That is why in recent years the Fed has made low inflation a particularly high priority. The large fraction of the public having first-hand experience with high inflation naturally supports the view that inflation must be contained. As the collective memory of inflation fades, however, public support for low inflation will become increasingly difficult to sustain. A permanent national commitment to price stability requires that citizens personally unfamiliar with the trauma of high inflation understand the rationale for price stability and the tactical policy actions needed to maintain it.

The author is senior vice president and director of research. This article, which originally appeared in this Bank's *1996 Annual Report*, benefited greatly from the comments of Doug Diamond, Mike Dotsey, Bob Hetzel, Tom Humphrey, Bob King, Ben McCallum, Alan Stockman, and Alex Wolman. It should be emphasized that the views expressed are the author's alone and not necessarily those of the Federal Reserve System.

This article reviews the history of U.S. monetary policy in the 20th century with the aim of providing that understanding. It identifies mistakes that led to high and volatile inflation, lessons learned from the experience, and principles applied in the pursuit of low inflation today. U.S. monetary policy came of age in the 20th century in the sense that the country left the strict rules of the gold standard for the freedom of an inconvertible paper standard, which the Fed only slowly and painfully learned to manage. What follows is the story of that 20th century odyssey.

Section 1 discusses monetary policy under the gold standard and the circumstances that led to the founding of the Fed. Section 2 outlines the main conceptual obstacles that had to be overcome in order to manage monetary policy under a paper standard. The causes and disruptive consequences of inflationary policy at mid-century are discussed in Section 3.

Certain key theoretical and practical developments paved the way for the Fed to take responsibility for controlling inflation in the early 1980s. Section 4 covers these developments. Progress in the theory of the demand for and the supply of money as well as empirical evidence supporting the theory played key roles here. The failure of nonmonetary approaches to controlling inflation was also important. The recognition that a credible Fed commitment to price stability could minimize the unemployment cost of achieving low inflation also played a role.

Section 5 recommends that the Fed be given a legislative mandate for low inflation. The case is based on lessons learned in the inflationary 1960s, ’70s, and early ’80s, and on the principles that have been applied successfully to maintain low inflation since then. The closing section summarizes the monetary policy lessons learned on the 20th century odyssey.

1. MONETARY POLICY UNDER THE GOLD STANDARD

When the Federal Reserve was established in 1913, inflation was not the problem it was to become in the latter part of the century. The nation was on a gold standard and the purchasing power of money in 1913 was about what it had been 30 years before, or for that matter, 100 years before. The gold standard sharply restricted inflation by requiring that money created by the U.S. Treasury be backed by gold.[1]

The classical gold standard yielded price stability only to the extent that the Treasury’s stock of monetary gold happened to expand at a rate sufficient

[1] Technically, the United States was on a bimetallic (gold and silver) standard until 1900. Though it is true that the money supply was limited by the size of the Treasury’s gold and silver holdings, there was considerable short-run variability in the money multiplier. See Cagan (1965) and Freidman and Schwartz (1963).

to satisfy the economy's demand for money at stable prices. For instance, slow growth in the gold supply caused the price level to decline at over 1 percent per year from 1879 to 1897, and gold discoveries and new mining techniques caused inflation to average over 2 percent per year between 1897 and 1914. Nevertheless, by the standard of what was to come, the variation of inflation was very small.

Although the economy grew rapidly throughout the gold standard years, the period was marked by a number of recessions associated with temporary deflations and substantial interest rate movements. Sudden sustained short-term interest rate spikes of over 10 percentage points occurred on eight occasions between the Civil War and the founding of the Federal Reserve. Five of these spikes were associated with bank runs characterized by a demand to convert bank deposits into currency that could not be satisfied by the fractional cash reserves held by banks.[2]

Finally, in response to the banking panic of 1907 and the ensuing recession, the nation was no longer willing to run monetary policy entirely according to the classical gold standard rules. The Federal Reserve was established in the United States with the power to create currency and bank reserves at least somewhat independently of the nation's monetary gold. The Fed was given authority to create currency and reserves by making loans to banks through its discount window or by acquiring securities in the money market. The Fed's mission was to provide an elastic supply of money to smooth short-term interest rates against liquidity disturbances, while preserving the link between money and gold in the long run in order to restrain inflation.[3]

Through its dominant presence in the market for currency and bank reserves, the Fed easily gained control of short-term interest rates and eliminated the kind of interest rate spikes seen earlier.[4] By smoothing short-term interest rates, however, the Fed was obliged to substitute its discretionary management of short rates for the impersonal market forces that had determined rates previously. The Bank of England had successfully managed short rates for decades in the context of the classical gold standard.[5] And the Fed could have followed similar gold standard operating procedures. However, the classical gold standard collapsed with World War I, and the nation was never willing to support Fed procedures geared to defending the gold standard. The Fed was left without clear operational procedures for positioning short-term interest rates to stabilize economic activity around full employment with stable prices.

---

[2] Major banking panics occurred in 1873, 1884, 1890, 1893, and 1907.

[3] This latter understanding was viewed as part of the Fed's mission, although it is only implicitly, not explicitly, stated in the Federal Reserve Act of 1913 itself.

[4] See Goodfriend (1988).

[5] See Hawtrey (1938).

2. THE OBSTACLES TO UNDERSTANDING MONETARY POLICY

Improvements in monetary policy that seemed within reach after the founding of the Fed proved elusive. The 1930s saw the sharpest deflation, the worst banking crisis, and the longest and deepest economic depression in American history.[6] Then, beginning in the mid-1960s there were two decades of unprecedented peacetime inflation that tripled the general price level by the early 1980s.[7]

Why has it taken so long for the Fed to give price stability pride of place?[8] Initially, there was a tendency to underestimate the disruptive potential of inflation and a willingness to be tolerant of each new burst of inflation in the hope that it would soon die down. Such hope seemed reasonable since protracted peacetime inflation had never before been a problem in the United States. Another difficulty was that it took some time for economists to develop a framework capable of understanding monetary policy in the absence of a link to gold. Prior to the 20th century the world had little practical experience with monetary regimes in which money was unbacked by a commodity such as gold or silver. With some exceptions, mainly during wartime, there was little empirical evidence on such regimes and little interest in analyzing them.

The main problem was confusion within the economics profession about the determination of the general price level and the control of inflation in a regime of inconvertible paper money.[9] There was also little understanding of the role played by inflation expectations in the wage- and price-setting process and in the determination of interest rates. Finally, the relationship between unemployment and inflation was seriously misunderstood. The resolution of these disputes provided the foundation for today's monetary policy success.

3. INFLATIONARY MONETARY POLICY AT MID-CENTURY

Largely as a result of the nation's unfortunate experience with inflation in the period from the mid-1960s through the early 1980s, monetary economists

[6] According to Friedman and Schwartz (1963) U.S. real net national product fell by more than one-third from 1929 to 1933, implicit prices of goods and services fell by more than one-quarter, and wholesale prices by more than one-third. More than one-fifth of the commercial banks in the United States holding nearly one-tenth of the deposits closed because of financial difficulties. As a result of the sharp contraction in economic activity, the unemployment rate peaked at over 20 percent in 1932–33, and remained above 10 percent for the remainder of the decade.

[7] The Fed had already recognized inflation as a problem on three occasions prior to the mid-1960s: in the aftermath of World War II, during the Korean War and the period of the 1951 Fed-Treasury Accord, and again in the mid-1950s. See footnote 12.

[8] Under the leadership of Benjamin Strong, Governor of the Federal Reserve Bank of New York, the Fed made price stability a priority briefly in the 1920s. See Hetzel (1985).

[9] See, for example, Bronfenbrenner and Holzman (1963) and Friedman (1987).

and central bankers now understand that the costs of inflationary monetary policy are significant and varied. First are the costs that even a steady, perfectly anticipated inflation imposes on society. Then there are the disruptive and destabilizing costs of unstable inflation, more difficult to quantify but substantial nonetheless. These latter costs stemmed from alternating expansionary and contractionary policy actions. Specifically, there was a tendency—known as go-stop monetary policy—for the Fed to exacerbate the cyclical volatility of inflation and unemployment. And there was a related tendency to produce rising inflation and increasingly volatile inflation expectations over time. The forces giving rise to these tendencies are identified and described below together with their disruptive consequences.

The Cost of Steady Inflation

The cost of steady inflation begins with the fact that a steadily falling purchasing power of money causes people to hold less cash than they would if prices were stable. Attempts to economize on money holdings manifest themselves in several ways. Banks invest in teller machines, people visit banks or teller machines more frequently, businesses devote more time and effort to managing their cash balances, etc.[10] Even more important, individuals and firms take steps to protect the value of their savings and investments against loss due to inflation. The effort and resources devoted to dealing with inflation are wasted from society's point of view in the sense that they could be better employed in producing goods and services.

Another major cost of steady inflation stems from the incomplete indexation of the tax system. The biggest problem in this regard results because taxes are assessed on *nominal* interest earnings and *nominal* capital gains, that is, on investment returns in dollars. Inflation causes nominal returns to rise because investors demand compensation for the declining purchasing power of money. For instance, long-term bond rates contain a premium for expected inflation over the life of the bond. Since nominal returns are taxed as income, however, inflation reduces the after-tax return to saving and investment and thereby tends to inhibit capital accumulation and economic growth.[11]

[10] Estimates in Lucas (1994) imply that the economization of money balances that occurred at a rate of inflation of 5 percent per year (associated with a short-term nominal interest rate of about 6 percent) wasted about 1 percent of U.S. GDP. The payment of interest on transactions deposits in recent years raises money balances and reduces this welfare cost somewhat. The bulk of the welfare gain to reducing inflation is probably realized at a slightly positive inflation rate. See Wolman (1997).

[11] Feldstein (1996) reports that the net present value of the welfare gain of shifting from 2 percent inflation per year to price stability forever is about 30 percent of the current level of GDP.

Go-Stop Monetary Policy[12]

A central bank such as the Fed that is charged with conducting monetary policy on a discretionary basis is naturally inclined to give considerable weight to the public's mood. Go-stop monetary policy was, in good part, a consequence of the Fed's inclination to be responsive to the shifting balance of public concerns between inflation and unemployment. Of course, difficulties in judging the strength of the economy and in gauging inflationary pressures compounded the problem, as did ignorance of the lags in the effect of policy.

For the most part the public tolerated inflation as long as it was low, steady, and predictable. When labor markets were slack, the public was even willing to risk higher inflation in order to stimulate additional economic activity. Only when economic activity was strong and inflation moved well above the prevailing trend did inflation top the list of public concerns.

It is easy to understand why inflation need not greatly concern the public when it is steady and predictable. Individuals and firms are inconvenienced only slightly by steady inflation. As long as wages, prices, and asset values move up in tandem, there are no big financial consequences, especially when inflation is low. Likewise, a temporary and modest increase of inflation around a low, well-established trend need not immediately arouse concerns.

However, a persistent departure of inflation above trend causes anxieties because people wonder where a new trend might be established. Investors worry about how much of an inflation premium to demand in interest rates; businesses worry about how aggressively to price in order to cover rising costs; and workers worry about maintaining the purchasing power of their wages.

In marked contrast to inflation, which affects all, unemployment actually affects relatively few at a given time. The unemployment rate in recent decades has risen at most to only about 10 percent of the labor force. The public is concerned about unemployment not so much because of those who are currently unemployed but because people are afraid of becoming unemployed. It follows that the public is generally more concerned about unemployment when the unemployment rate is rising, even if it is still low, than when it is falling, even if it is already high.

The above-mentioned reasoning helps explain why the Fed produced go-stop monetary policy in the 1960s and '70s. In retrospect, one observes the following pattern of events.

[12] Friedman (1964, 1972, and 1984) discusses go-stop policy. Romer and Romer (1989) document that since World War II the Fed tightened monetary policy decisively to fight inflation on six occasions beginning, respectively, in October 1947, September 1955, December 1968, April 1974, August 1978, and October 1979. The unemployment rate rose sharply after each policy shock. Only two significant increases in unemployment were not preceded by Fed action to fight inflation. One occurred in 1954 after the Korean War and the second occurred in 1961, after the Fed tightened monetary policy to improve the international balance of payments.

First, because inflation became a major concern only after it clearly moved above its previous trend, the Fed did not tighten policy early enough to preempt inflationary outbursts before they became a problem.

Second, by the time the public became sufficiently concerned about inflation for the Fed to act, pricing decisions had already begun to embody higher inflation expectations. Thus delayed, a given degree of restraint on inflation required a more aggressive increase in short-term interest rates with greater risk of recession.

Third, in any cyclical episode there was a relatively narrow window of broad public support for the Fed to tighten monetary policy. The window opened after inflation was widely recognized as the major concern and closed when tighter monetary policy caused the unemployment rate to begin to rise. Often the Fed did not take full advantage of a window of opportunity to raise short rates, because it wanted more confirmation that higher short-term rates were required.

Fourth, it was probably easier for the Fed to maintain public support for fighting inflation with prolonged rather than preemptive tightening. A more gradual lowering of interest rates in the later stage of a recession was a less visible means of fighting inflation than raising rates more sharply earlier. Once unemployment peaked and began to fall, the public's anxiety about it diminished. Prolonged tightening was attractive as an inflation-fighting measure in spite of the fact that it probably lengthened the "stop" phase of the policy cycle.

Rising Inflation and Unstable Inflation Expectations

Over time, deliberately expansionary monetary policy in the "go" phase of the policy cycle came to be anticipated by workers and firms. Workers learned to take advantage of tight labor markets to make higher wage demands, and firms took advantage of tight product markets to pass along higher costs in higher prices. Increasingly aggressive wage- and price-setting behavior tended to neutralize the favorable employment effects of expansionary policy. And the Fed became evermore expansionary on average in its pursuit of low unemployment, causing correspondingly higher inflation and inflation expectations. Lenders demanded unprecedented inflation premia in long-term bond rates. And the absence of a long-run anchor for inflation caused inflation expectations and long bond rates to fluctuate widely.[13]

The breakdown of mutual understanding between the markets and the Fed greatly inhibited the conduct of monetary policy. The Fed continued to manage closely short-term nominal interest rates.[14] But the result of an interest rate

[13] The monthly average 30-year bond rate rose from around 8 percent in early 1978 to peak above 14 percent in the fall of 1981. The long bond rate was near 13 percent as late as the summer of 1984.

[14] See Cook (1989).

policy action is largely determined by its effect on the real interest rate, which is the nominal rate minus the public's expected rate of inflation. And the Fed found it increasingly difficult to estimate the public's inflation expectations and to predict how its policy actions might influence those expectations. Compounding the problem, enormous increases in short-term interest rates were required by the early 1980s to stabilize the economy. Stabilization policy became more difficult because the public could not predict what a given policy action implied for the future, and consequently, the Fed could not predict how the economy would respond to its policy actions.

4. THE CONTROL OF INFLATION: DISINFLATION IN THE 1980s

By the late 1970s, policymakers and monetary economists had come to understand the costly and disruptive features of inflation discussed above. With considerable public support, the Fed under the leadership of Chairman Paul Volcker initiated the great disinflation in October 1979, marking the beginning of the period in which the Fed would make lowering inflation a priority. What followed was a tightening of monetary policy that succeeded in bringing the inflation rate down permanently for the first time in the post–Korean War period, first from over 10 percent to around 4 percent by 1983, and then to around 3 percent by the mid-1990s.

This section reviews three developments that paved the way for the Fed to take responsibility for price stability. Most important was the progress that economists made in understanding money demand and supply. Next was the failure of nonmonetary approaches to controlling inflation. Finally, and to a lesser extent, was the idea advanced by monetary economists that the unemployment cost of disinflation might be minimized if the disinflation were credible.

The Central Bank's Responsibility for Inflation

The consensus among monetary economists that central banks are responsible for inflation is built on both theory and evidence. Above all, there is the substantial body of evidence from the inflationary experiences of a great many nations, including the widespread inflation in the industrialized world during the 1960s and '70s, showing that sustained inflation is always associated with excessive money growth. The evidence also clearly indicates that inflation is stopped by slowing the growth of the money supply.[15]

[15] See, for instance, Friedman (1987), Poole (1978), and Sargent (1986).

The theory of money demand and supply supports the cross-country evidence by illuminating the mechanics of the link between monetary policy and inflation. The theory of money demand implies that control of the money supply is necessary and sufficient to control the trend rate of inflation. And the theory of money supply implies that a central bank can control the trend rate of money growth. As will become clear below, money demand may be thought of as the fulcrum by which a central bank controls inflation, and the money supply may be thought of as the lever by which it does so.

Money Demand

The theory of money demand asserts that individuals and businesses choose to hold a target stock of money that is proportional to their expenditures, a target that balances the convenience of holding money against the foregone interest earnings.[16] The key implication of money demand theory for monetary policy is that there is a reasonably stable long-run relationship between a nation's demand for money and its production and exchange of goods and services.

It follows that sustained inflation results when the growth of the nation's money stock exceeds the rate of growth of the nation's physical product.[17] Prices must rise in this case because otherwise individuals and firms would spend their growing excess money balances. Since one person's expenditure is another person's receipts, the spending would put upward pressure on prices until the inflation rate matched the rate of money growth in excess of the growth of output. Only then would the ongoing increase in the stock of money be willingly absorbed by the public.

The theory of money demand also implies that the overall price level cannot move very much over the long run if the stock of money grows in tandem with the growth of output.[18] If an inflation were to start, it would reduce the purchasing power of a given nominal stock of money and cause individuals and businesses to cut their spending in an effort to maintain their inventory of monetary purchasing power. With no additional money balances forthcoming in the aggregate, the downward pressure on spending would stop the inflation.

Money Supply

The nation's basic money supply consists of currency and checkable deposits held by households and businesses. A central bank can control the former

[16] See McCallum and Goodfriend (1987).

[17] The public's target ratio of money to expenditure may exhibit a trend at times in response to, say, rising interest rates or technical progress in the payments system. For instance, the ratio of money to expenditure will trend downward if money provides transaction services more efficiently over time. In that case, the money growth rate consistent with price stability will be below the growth of physical product.

[18] See the preceding note.

because it has a monopoly on the creation of currency.[19] Checkable deposits are created by banks. A central bank also has the power to control checkable deposits because banks must hold reserves to service their deposits, and a central bank controls the aggregate stock of bank reserves.[20]

The financial services industry has long been creating new instruments in which the public can hold liquid balances, e.g., certificates of deposit and money market mutual funds. New financial instruments usually do not add to the basic money supply since they are only imperfect substitutes for currency or checkable deposits.[21] Nevertheless, the introduction of money substitutes has adversely affected the predictability of money demand in the short run. In practice, however, money demand is sufficiently stable and money supply sufficiently controllable over time, so that financial innovations do not fundamentally alter a central bank's power over inflation.[22]

Failed Approaches to Controlling Inflation

A variety of nonmonetary approaches to controlling inflation were tried in the 1960s and '70s. In the United States, for example, the federal government published voluntary wage-price guidelines at various times to persuade firms and workers to forego price and wage increases deemed excessive.[23] Actual controls were imposed for a few years in the early '70s but for the most part they were lifted by the mid-'70s.[24] By the end of the period, both controls and guidelines came to be regarded as arbitrary, unfair, and ineffective. Moreover, where they were effective they often created allocative disruptions, e.g., price controls in the energy sector created shortages and long lines at gas stations.

In the early 1960s economists believed that budget policy might play a key role in fighting inflation. In the United States, however, it quickly became clear in the Vietnam War period that political concerns would immobilize fiscal policy as a practical economic policy tool. Moreover, it later became clear that

[19] Electronic private substitutes for government currency have become feasible recently. See Lacker (1996).

[20] See Cagan (1965).

[21] There have been exceptions, however. For instance, a new deposit type known as the negotiable order of withdrawal (NOW) account was introduced in the late '70s and early '80s as part of the deregulation of the prohibition of interest on checkable deposits. NOW accounts were interest-earning substitutes for demand deposits and so were immediately included in the Fed's M1 measure of the basic money supply for purposes of targeting and control. See Broaddus and Goodfriend (1984).

[22] For instance, see Lucas (1988) and Meltzer (1963) on the long-run stability of the demand for M1.

[23] See Heller (1966) and Shultz and Aliber (1966).

[24] See Kosters (1975).

the inflation of the 1970s was not closely related to the government's fiscal situation.[25]

Even after the Fed under Chairman Volcker had begun its momentous disinflation, the Carter administration imposed credit controls in early 1980 in an effort to foster the process. The credit control program caused a sharp recession with little impact on inflation and was phased out at midyear.[26]

Thus did policymakers learn the hard way that policies for stopping inflation other than monetary control didn't work. As much as anything else, the failure of nonmonetary approaches to disinflation set the stage for the Fed to take responsibility for bringing inflation down.

Credibility for Low Inflation and the Unemployment Cost of Disinflation

In the early 1960s policymakers were inclined to accept the inflationary consequences of policy actions taken to stimulate aggregate demand and employment. That inclination was based to a great extent on evidence of a century-long negative Phillips curve correlation between unemployment and (wage) inflation in the United Kingdom that appeared to offer a trade-off in which the benefits of lower inflation would have to be balanced against the costs of higher unemployment.[27]

When stimulative policy succeeded in driving down the unemployment rate in the '60s, the resulting increase in inflation at first seemed consistent with a stable Phillips curve trade-off; and the rising inflation was tolerated as a necessary evil.[28] In the 1970s, however, the Phillips curve correlation broke down as inflation and unemployment both moved higher, and it became clear that high inflation could not buy permanently low unemployment.[29]

Even though protracted inflation was widely understood by the late 1970s to have costs with no offsetting benefits, it was recognized that bringing inflation down would be costly too. Previous experience with go-stop policy made it clear that there was a short-run trade-off between unemployment and inflation.[30] Policymakers expected the temporary unemployment cost of a large permanent disinflation to exceed the costs of earlier disinflations that the Fed had produced in the "stop" phase of its policy cycles.

---

[25] Government fiscal concerns are the driving force behind high inflations. See Sargent (1986).

[26] See Schreft (1990).

[27] See Phillips (1958).

[28] See Heller (1966) and Tobin (1972).

[29] See Fischer (1994), pp. 267–68.

[30] King and Watson (1994), for example, report a significant negative correlation between unemployment and inflation over the business cycle.

To some degree a view then emerging in the academic community might have encouraged the Fed to pursue the disinflation. The view holds that the unemployment cost of disinflation can be minimized if a disinflation policy is credible. The idea that credibility would govern the costliness of disinflation has since become widely accepted in theory.[31] The acquisition and maintenance of credibility for low inflation have become major practical concerns of Fed policymakers and central bankers around the world.

The idea underlying the role of credibility is that wage- and price-setting behavior is geared to expectations of money growth. The Fed supports the ongoing inflation as long as money grows in excess of output. If the Fed's disinflation is credible, the Fed slows money growth and wage and price inflation come down, too, with little effect on employment. On the other hand, if the disinflation is not credible, then wage and price inflation continues as before. If the Fed persists in slowing money growth anyway, a deficiency of aggregate demand causes unemployment as households and businesses cut spending in an attempt to maintain their targeted monetary purchasing power.[32]

In practice, however, disinflation is nearly always costly because credibility for low inflation is hard to acquire after it has been compromised. Moreover, a central bank's commitment to low inflation is only as credible as the public's support for it. The Fed probably embarked on the disinflation in 1979, in part, because the public finally seemed ready to accept it.

Although its discount rate changes often made the headlines prior to 1979, the Fed rarely sought publicity for its monetary policy actions. Chairman Volcker broke sharply with tradition by initiating the period of disinflationary policy with a high-profile announcement signaling that the Fed would take responsibility for inflation and bring it down.[33] In so doing, Chairman Volcker built credibility by staking his own reputation and the Fed's on

[31] Barro and Gordon (1983), Fellner (1976), Sargent (1986), and Taylor (1982) contain early discussions of credibility as it relates to monetary policy. Persson and Tabellini (1994) contains a recent survey of research on the role of credibility in monetary and fiscal policy. The new large-scale Federal Reserve Board macroeconomic model is designed to take account of different degrees of credibility in policy simulations. See "A Guide to FRB/US" (1996).

[32] What happens is this: In the first instance households and businesses attempt to exchange financial assets for money. Such actions, however, cannot satisfy the aggregate excess demand for money directly. They drive asset prices down and interest rates up until the interest sensitive components of aggregate expenditure grow slowly enough to eliminate the excess demand for money. As the disinflation gains credibility, wage and price inflation slows, and real aggregate demand rebounds until the higher unemployment is eliminated.

Ball (1994) shows that a perfectly credible disinflation need have no adverse effects on employment even in a model with considerable contractual inertia in the price level.

[33] The Fed did not explicitly assert its responsibility for inflation in the initial announcements of its disinflationary policy. However, by emphasizing the key role played by money growth in the inflation process, and by announcing a change in operating procedures to emphasize the control of money, the Fed *implicitly* acknowledged its responsibility for inflation. See *Federal Reserve Bulletin* (November 1979), pp. 830–32.

achieving the low inflation objective. The unprecedented increases in short-term interest rates that followed further demonstrated the Fed's commitment to reducing inflation.[34]

Nevertheless after two decades of rising inflation, a widespread skepticism worked against Fed credibility.[35] Wage and price setters doubted that there would be sufficiently widespread public support for the Fed's disinflation. Indeed, the inflation was not broken until a sustained slowing of money growth beginning in 1981 created a serious recession that tested the Fed's determination and the public's support.[36] Although the recession was the worst since the 1930s, it was less severe than might have been expected considering the size of the accompanying disinflation. Most remarkable is that the roughly 6 percentage point disinflation occurred in just two years: 1981 and 1982. The size and speed of the disinflation suggests that the acquisition of credibility played a key role in making it happen.

5. MONETARY POLICY AT THE CLOSE OF THE CENTURY: MAINTAINING LOW INFLATION

The Fed has succeeded in maintaining low inflation for almost 15 years now. With luck the United States should enter the 21st century with inflation near what it was under the gold standard at the opening of the 20th century. Macroeconomic performance during the low inflation period has been good,

[34] The Fed took short-term rates from around 11 percent in September 1979 to around 17 percent in April 1980. This was the most aggressive series of actions the Fed has ever taken in so short a time, although the roughly 5 percent increase in short rates from January to September of 1973 was almost as large. See Goodfriend (1993).

[35] The collapse of confidence in U.S. monetary policy in 1979 and 1980 was extraordinary. The price of gold rose from around $275 per ounce in June 1979 to peak at about $850 per ounce in January 1980, and it averaged over $600 per ounce as late as November 1980. Evidence of a weakening economy caused the Fed to pause in its aggressive tightening in early 1980. But with short rates relatively steady, the 30-year rate jumped sharply by around 2 percentage points between December and February, signaling a huge jump in long-term inflation expectations. The collapse of confidence in early 1980 was caused in part by the ongoing oil price shock and the Soviet invasion of Afghanistan in December 1979. But the Fed's hesitation to proceed with its tightening at the first sign of a weakening economy probably also played a role. In any case, the Fed responded with an unprecedented 3 percentage point increase in short rates in March, taking them to around 17 percent. See Goodfriend (1993).

[36] After making its disinflationary policy commitment in October 1979, the Fed let the growth of effective M1 overshoot its target range in 1980 and the inflation rate continued to rise, peaking at over 10 percent in the fourth quarter. Then, in sharp contrast to the preceding four years, effective M1 actually undershot its target range in 1981. Effective M1 grew around 4.6 percentage points slower in 1981 than its average annual growth over the preceding five years. Further, the actual 2 percent shortfall in M1 from the midpoint of its 1981 target was built into the 1982 target path. See Broaddus and Goodfriend (1984).

The unemployment rate rose from around 6 percent in 1978 to average nearly 10 percent in the recession year of 1982.

especially when compared to the inflationary period preceding it. The only recession during the period, in 1990–91, was mild by recent standards. Over the period as a whole, employment growth has been strong and productivity growth may have picked up somewhat. Moreover, both short- and long-term interest rates are around one-third of what they averaged in the early 1980s and are much less volatile too.

The promise of low inflation is being fulfilled. The challenge today is for the Fed to understand the secret of its success. In that regard the low inflation period has as much to teach as the traumatic period that preceded it. In reviewing below the lessons learned and principles applied, we shall see that the best way of assuring our continued monetary policy success would be for Congress to give the Fed a legislative mandate for low inflation.

Lessons Learned and Principles Applied

One of the most important lessons learned from the last four decades is that credibility for low inflation is the foundation of effective monetary policy. The Fed has acquired credibility since the early 1980s by consistently taking policy actions to hold inflation in check. In effect, the Fed has reestablished a mutual understanding between itself and the markets. From this perspective, wage and price setters keep their part of an implicit bargain by not inflating as long as the Fed demonstrates its commitment to low inflation. Ironically, the Fed has learned from nearly a century of experience to pursue rule-like behavior in order to fully achieve the gains from moving away from the gold standard.

Experience shows that the guiding principle for monetary policy is to preempt rising inflation. The go-stop policy experience teaches that waiting until the public acknowledges rising inflation to be a problem is to wait too long. At that point, the higher inflation becomes entrenched and must be counteracted by corrective policy actions more likely to depress economic activity.

The main tactical problem for the Fed is to decide when preemptive policy actions are necessary and how aggressive they should be. In this regard, the Fed must be careful to consider any adverse effect that a poorly timed policy tightening could have on employment and output. For that matter, the Fed must be prepared to ease monetary policy when a weakening economy calls for it. The central bank's credibility depends not only on its inflation-fighting credentials but also on its perceived competence.

A natural starting point to balance these concerns is to use a policy rule-of-thumb based on historical data to benchmark Fed policy. The stance and direction of monetary policy can then be chosen in light of historical experience conditioned on any special current circumstances. The most relevant historical experience is, of course, the relatively brief low inflation period since the

mid-1980s. As the Fed extends low inflation over time, the nation will build up a richer relevant history against which to benchmark policy.[37]

However, even our brief experience with low inflation contains useful insights such as this. In some years, such as 1994, inflationary pressures might be judged to call for a particularly aggressive preemptive tightening. At other times, such as in 1996, there might be some concern about the potential for rising inflation but enough doubt to adopt a wait-and-see attitude. The Fed's success in 1994 and 1996 suggests that the key to effective management of short-term interest rates over the business cycle is to move rates up decisively and preemptively when warranted in order to build credibility for low inflation. With credibility "in the bank," so to speak, the Fed can hold rates steady or move them down out of concern for unemployment at other times.[38] The lesson is that credibility enhances flexibility.

A Legislative Mandate for Price Stability

Largely as a result of the common understanding of the theory and history of monetary policy reviewed above, there is today a consensus among monetary economists and central bankers that maintaining low inflation is the foundation of effective monetary policy. Moreover, there is an *emerging* consensus that a central bank's commitment to price stability should be strengthened by legislation making low inflation the primary goal of monetary policy.[39]

---

[37] Simple policy rule specifications studied with models estimated on historical data can be of great practical value in benchmarking actual policy decisions. McCallum (1988) and Taylor (1993) present two rules, respectively, that are particularly useful in this regard. McCallum models the monetary base (currency plus bank reserves) as the Fed's policy instrument, and has it responding to a moving average of base velocity and departures of nominal GDP from a target path. Taylor models the real short-term interest rate (the market interest rate minus expected inflation) as the policy instrument, and has it responding to inflation and the gap between actual and potential GDP.

Each specification has advantages and disadvantages. Taylor's rule matches more closely the way the Fed thinks of itself as operating. But McCallum's rule makes clear that the ultimate power of the Fed over the economy derives from its monopoly on the monetary base. McCallum's rule has the advantage that it could still be used if disinflation happened to push the market short rate to zero, or if inflation expectations became excessively volatile. In either situation the Fed might be unable to use the real short rate as its policy instrument.

[38] See Board of Governors "Monetary Policy Report to Congress" (1994, 1995, and 1996).

[39] In 1995, Senator Connie Mack introduced a bill that would make low inflation the primary goal of monetary policy. In 1989, Fed Chairman Alan Greenspan testified in favor of a prior resolution that would have mandated a price stability objective for the Fed. Academics as diverse as Blinder (1995), Fischer (1994), and Friedman (1962) all agree that the Fed should be given some sort of mandate for low inflation. The remarkable convergence of professional thinking in favor of a mandate was evident at the Federal Reserve Bank of Kansas City's August 1996 conference on price stability. See *Achieving Price Stability* (1996).

Inflation targeting is employed by a number of central banks around the world. See Leiderman and Svensson (1995).

The recommended priority for price stability derives not from any belief in its intrinsic value relative to other goals such as full employment and economic growth. Price stability should take priority for two reasons: first, the Fed actually has the power to guarantee it over the long run, and second, monetary policy encourages employment and economic growth in the long run mostly by controlling inflation.[40] Also, and this is very important, a mandate for price stability would not prevent the Fed from taking the kinds of policy actions it takes today to stabilize employment and output in the short run. What it *would* do is discipline the Fed to justify these actions against a commitment to protect the purchasing power of money.

Two often-repeated objections to a mandate for low inflation deserve mention here. One is the notion that low inflation targeting is largely irrelevant because two enormous oil price increases in the 1970s—in 1973/74 and 1979/80—were responsible for the worst inflation of that period.[41] The claim continues that our success in controlling inflation will be determined by whether we have large oil price shocks in the future or not. Clearly, oil price increases create a problem for the economy: the higher price of oil diverts expenditure to oil products and raises real costs throughout the economy, with adverse consequences for demand and employment in non-oil sectors.

The economy must adjust to the higher real cost of oil in any case. The problem for a central bank is to make sure that the adjustment problem is not compounded with monetary instability. A central bank with a mandate for low inflation is more likely to resist excessive monetary accommodation than one with a weaker commitment to price stability. This is because an oil price shock will be less likely to set in motion wage and price increases that the central bank will be inclined to accommodate. The Fed was in just this predicament when the 1970s oil price shocks hit, since rising inflation trends were already well established before each oil shock. The destabilizing effects on inflation, inflation expectations, and employment and output would almost surely have been less troublesome in a climate of stable inflation.

A second objection to a mandate for low inflation is that it would hold back economic growth. In fact, the opposite is more nearly true. In terms of the earlier discussion of money demand and supply, trend growth of national output continually raises the demand for money, and the Fed accommodates the growing demand for money at stable prices.

Would monetary policy prevent the economy from growing faster if labor productivity unexpectedly surged? Not for long, because unemployment would begin to rise as businesses found that they could meet demand with less labor

[40] Rudebusch and Wilcox (1994) report empirical evidence on inflation and productivity growth. Dotsey and Ireland (1996) study the question in a quantitative, theoretical model.

[41] Oil prices rose from around $3 to $12 a barrel during the 1973/74 oil price shock, and from about $15 to over $35 in 1979/80.

input. And the Fed would resist rising unemployment by easing monetary policy to encourage faster growth in aggregate demand. In short, the Fed's policy procedures do not "target growth." A mandate for price stability would allow the Fed to naturally and automatically accommodate an increase in productivity growth over time.

Ultimately the Fed can only secure full credibility for low inflation with the backing of the public. The public's misunderstanding of the tactics of monetary policy is particularly troublesome. For instance, accusations that the Fed was "busting ghosts" when it ran short-term interest rates up in 1994 threatened to undermine support for policy actions that were clearly called for.[42] Preemptive policy actions in 1994 laid the foundation for continued economic expansion. The task ahead must be to broaden and deepen the public's understanding and support for the strategy and tactics of monetary policy and to lock in credibility for low inflation with a legislative mandate.

6. CONCLUSION

American monetary policy has come full circle in the 20th century. Early in the century the nation overcame a long-standing distrust of government intervention in the monetary system to establish a central bank. The Federal Reserve embodied the idea that discretionary monetary policy could improve on the rules of the gold standard, rules that were seen as unduly restrictive. We now know that the faith then placed in discretion over rules was somewhat misplaced. Today, monetary economists and central bankers alike understand that effective monetary policy must be built on a consistent commitment to low inflation.

Numerous lessons were learned on the 20th century odyssey. The most important is that the Federal Reserve, through its management of monetary policy, is responsible for inflation. This became clear partly as a result of advances in monetary theory and partly as a result of evidence on money demand and money supply. It was also the result of a learning process in which nonmonetary approaches to controlling inflation were seen to fail, and the monetary approach succeeded.

Discretionary monetary policy actions can be invaluable in fighting a financial crisis or a weak economy. But we learned that the promise of discretion can be realized fully only in the context of a monetary policy that makes price stability a priority. Otherwise discretion leads inexorably to go-stop policy that brings rising and unstable inflation and inflation expectations, with adverse consequences for interest rates and employment.

[42] See Thurow (1994). By successfully keeping inflation in check, preemptive policy actions *necessarily* appear to be busting ghosts. So the appearance of ghost busting is a consequence of good monetary policy.

The go-stop experience taught that the Fed should fight inflation by tightening monetary policy before price pressures break out into the open. Waiting until inflation has begun to rise may better assure public support for higher short-term interest rates. But delayed tightening allows higher inflation to become more firmly established, requiring even higher rates to choke it off, with a greater risk of recession.

An emerging consensus among monetary economists and central bankers supports the need for a legislative mandate to make low inflation the primary goal of monetary policy. That recommendation has broad backing for three reasons. A central bank can guarantee low inflation over time. Monetary policy most effectively stabilizes employment over the business cycle when it has credibility for low inflation. And full credibility for low inflation needs the support of a legislative mandate.

Monetary policy has come of age in the 20th century in the sense that monetary economists and central bankers have come to terms with the past—lessons have been learned and principles have been applied successfully. The country should build on that professional consensus to broaden the public's understanding and support for price stability and the preemptive policy procedures to sustain low inflation. The nation has the opportunity to bring a tumultuous chapter in its monetary history to a close. It should grasp that opportunity and enjoy the benefits that sustained price stability would bring.

REFERENCES

Achieving Price Stability. A Symposium Sponsored by the Federal Reserve Bank of Kansas City. Jackson Hole, Wyoming, 1996.

"A Guide to FRB/US: A Macroeconomic Model of the United States." Manuscript. Macroeconomic and Quantitative Studies, Division of Research and Statistics, Federal Reserve Board: July 1996.

Ball, Laurence. "Credible Disinflation with Staggered Price-Setting," *American Economic Review,* vol. 84 (March 1994), pp. 282–89.

Barro, Robert J., and David B. Gordon. "Rules, Discretion and Reputation in a Model of Monetary Policy," *Journal of Monetary Economics,* vol. 12 (July 1983), pp. 101–21.

Blinder, Alan. "Central Banking in Theory and Practice," The Marshall Lecture, Cambridge University, Federal Reserve Board, processed May 1995.

Board of Governors of the Federal Reserve System. "Monetary Policy Report to Congress," *Federal Reserve Bulletin,* July 1994, February 1995, and August 1996.

_____. "Announcements," *Federal Reserve Bulletin,* vol. 65 (November 1979), pp. 830–32.

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

Bronfenbrenner, Martin, and Franklyn D. Holzman. "Survey of Inflation Theory," *American Economic Review,* vol. 53 (September 1963), pp. 593–661.

Cagan, Phillip. *Determinants and Effects of Changes in the Stock of Money, 1875–1960*. New York: Columbia University Press, 1965.

Cook, Timothy. "Determinants of the Federal Funds Rate: 1979–1982," Federal Reserve Bank of Richmond *Economic Review,* vol. 75 (January/February 1989), pp. 3–19.

Dotsey, Michael, and Peter Ireland. "The Welfare Cost of Inflation in General Equilibrium," *Journal of Monetary Economics,* vol. 37 (February 1996), pp. 29–47.

Economic Report of the President. Washington: Government Printing Office, 1996.

Feldstein, Martin. "The Costs and Benefits of Going from Low Inflation to Price Stability," *NBER* Working Paper 5469. Forthcoming in C. Romer and D. Romer, eds., *Monetary Policy and Inflation*. Chicago: University of Chicago Press, 1997.

Fellner, William. *Towards a Reconstruction of Macroeconomics: Problems of Theory and Policy.* Washington, D.C.: American Enterprise Institute for Public Policy Research, 1976.

Fischer, Stanley. "Modern Central Banking," in Forrest Capie, Stanley Fischer, Charles Goodhart, and Norbert Schnadt, eds., *The Future of Central Banking: The Tercentenary Symposium of the Bank of England*. Cambridge: Cambridge University Press, 1994.

Friedman, Milton. "The Quantity Theory of Money," in John Eatwell, Peter Newman, and Murray Milgate, eds., *The New Palgrave Dictionary of Money and Finance*. New York: The Stockton Press, 1987.

_____. "Monetary Policy for the 1980s," in John H. Moore, ed., *To Promote Prosperity: U.S. Domestic Policy in the Mid-1980s*. Stanford: Hoover Institution Press, 1984.

_____. *An Economist's Protest: Columns in Political Economy.* New Jersey: Thomas Horton and Company, 1972.

_____. Statement before U.S. Congress, House of Representatives, Committee on Banking and Currency in *The Federal Reserve System After Fifty Years*. Subcommittee on Domestic Finance. Hearings, 88 Cong. 2 Sess. Washington: Government Printing Office, 1964.

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

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

Goodfriend, Marvin. "Interest Rate Policy and the Inflation Scare Problem: 1979–1992," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 79 (Winter 1993), pp. 1–24.

_____. "Central Banking Under the Gold Standard," *Carnegie-Rochester Conference Series on Public Policy,* vol. 29 (Autumn 1988), pp. 85–124.

_____, and William Whelpley. "Federal Funds: Instrument of Federal Reserve Policy," Federal Reserve Bank of Richmond *Economic Review,* vol. 72 (September/October 1986), pp. 3–11.

Greenspan, Alan. Statement before the U.S. Congress, House of Representatives, Subcommittee on Domestic Monetary Policy of the Committee on Banking, Finance and Urban Affairs, *Zero Inflation*. Hearing, 101 Cong. 1 Sess. Washington: Government Printing Office, 1989.

Hawtrey, R. G. *A Century of Bank Rate*. London: Longran, 1938.

Heller, Walter W. *New Dimensions of Political Economy.* Cambridge, Mass.: Harvard University Press, 1966.

Hetzel, Robert L. "The Rules versus Discretion Debate over Monetary Policy in the 1920s," Federal Reserve Bank of Richmond *Economic Review,* vol. 71 (November/December 1985), pp. 3–14.

King, Robert G., and Mark W. Watson. "The Post-War U.S. Phillips Curve: A Revisionist Econometric History," *Carnegie-Rochester Conference Series on Public Policy,* vol. 41 (December 1994), pp. 157–219.

Kosters, Marvin H. *Controls and Inflation: The Economic Stabilization Program in Retrospect.* Washington, D.C.: American Enterprise Institute for Public Policy Research, 1975.

Lacker, Jeffrey M. "Stored Value Cards: Costly Private Substitutes for Government Currency," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 82 (Summer 1996), pp. 1–25.

Leiderman, Leonardo, and Lars Svensson, eds. *Inflation Targets*. London: Center for Economic Policy Research, 1995.

Lucas, Robert E., Jr. "On the Welfare Cost of Inflation," Center for Economic Policy Research, Stanford University, February 1994.

_____. "Money Demand in the United States: A Quantitative Review," *Carnegie-Rochester Conference Series on Public Policy*, vol. 29 (Autumn 1988), pp. 137–68.

McCallum, Bennett T. "Robustness Properties of a Rule for Monetary Policy," *Carnegie-Rochester Conference Series on Public Policy*, vol. 29 (Autumn 1988), pp. 173–204.

_____, and Marvin S. Goodfriend. "Demand for Money: Theoretical Analysis," in *The New Palgrave: A Dictionary of Economics*. London: Macmillan Press, 1987, reprinted in Federal Reserve Bank of Richmond *Economic Review*, vol. 74 (January/February 1988), pp. 16–24.

Meltzer, Allan. "The Demand for Money: The Evidence from the Time Series," *Journal of Political Economy*, vol. 71 (June 1963), pp. 219–46.

Persson, Torsten, and Guido Tabellini, eds. *Monetary and Fiscal Policy, Vol. 1: Credibility*. Cambridge, Mass.: MIT Press, 1994.

Phillips, A. W. "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957," *Economica*, vol. 25 (November 1958), pp. 283–99.

Poole, William. *Money and the Economy: A Monetarist View*. United States: Addison-Wesley Publishing Company, 1978.

Romer, Christina D., and David H. Romer. "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz," *NBER Macroeconomics Annual*, vol. 4 (1989), pp. 121–70.

Rudebusch, Glenn, and David Wilcox. "Productivity and Inflation: Evidence and Interpretations," Federal Reserve Board, Washington, D.C., 1994.

Sargent, Thomas J. *Rational Expectations and Inflation*. New York: Harper and Row, 1986.

Schreft, Stacey L. "Credit Controls: 1980," Federal Reserve Bank of Richmond *Economic Review*, vol. 76 (November/December 1990), pp. 25–55.

Schultz, George P., and Robert Z. Aliber, eds. *Guidelines: Informal Controls and the Market Place*. Chicago: University of Chicago Press, 1966.

Taylor, John B. "Discretion Versus Policy Rules in Practice," *Carnegie-Rochester Conference Series on Public Policy*, vol. 39 (December 1993), pp. 195–214.

_____. "Establishing Credibility: A Rational Expectations Viewpoint," *American Economic Review Papers and Proceedings*, vol. 72 (May 1982), pp. 81–85.

Thurow, Lester C. "The Fed Goes Ghostbusting," *The New York Times*, May 6, 1994.

Tobin, James. "The Cruel Dilemma," in Arthur M. Okun, ed., *Problems of the Modern Economy: The Battle Against Unemployment.* Revised. New York: W. W. Norton and Company, Inc., 1972.

Wolman, Alexander. "Zero Inflation and the Friedman Rule: A Welfare Analysis," Federal Reserve Bank of Richmond *Economic Quarterly* (forthcoming 1997).

The Pre-Commitment Approach in a Model of Regulatory Banking Capital

Edward S. Prescott

The pre-commitment approach to bank capital regulation is a radical departure from existing bank regulatory methods. First proposed in Kupiec and O'Brien (1995c), the approach advocates letting banks choose their capital levels and fining them if losses exceed this level. The essence of the proposal is to use fines (or other penalties) to encourage risky banks to hold more capital than safer ones.

Since a change in regulatory method will affect the banking sector, it is crucial to ascertain what will happen if the proposal is implemented. Because the approach is so new, there exists only a small literature explaining and evaluating it. Accordingly, the goal of this paper is to produce understanding of the pre-commitment approach and to determine its effectiveness as a regulatory tool.

Regulators care about banks' capital levels because the deposit insurance fund is liable in the event a bank is unable to repay its depositors. For a given portfolio, a higher ratio of capital to assets reduces the insurance fund's exposure to losses because there are proportionally fewer deposits to repay in the event of a loss. Along with the monitoring of banks and deposit insurance premiums, capital requirements are an essential part of the mechanism used by regulators to insure deposits.

Since 1988, regulators have used capital requirements to protect against credit risk, that is, against the event of borrower default. They have done so by

The author thanks Doug Diamond, Marvin Goodfriend, Ed Green, Tony Kuprianov, Jeff Lacker, David Marshall, Jim O'Brien, Subu Venkataraman, John Walter, and John Weinberg for helpful comments and discussions. An earlier version of this paper was presented at the Federal Reserve Bank of Chicago. The views expressed in this paper are those of the author and do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

categorizing bank assets into different risk categories, taking the risk-weighted sum of the assets, and then requiring capital to be roughly 8 percent of the total.[1] These rules, however, did not consider other sources of risk such as those from movements in market prices. Changes in market prices are particularly important sources of risk to banks that have large trading portfolios of derivatives and other financial securities.

Concern that these sources of risk are a hazard to the banking system and to the insurance fund produced three different proposals for using capital to protect against the risk in banks' trading portfolios: the standardized approach, the internal models approach, and the pre-commitment approach. The result of the ensuing public discussion was the adoption of the internal models approach, scheduled to take effect January 1, 1998. However, this decision has not precluded continued consideration of future regulatory changes. In particular, the pre-commitment approach continues to be studied by the Federal Reserve Board. (See Greenspan [1996].)

Before analyzing the pre-commitment approach, it is helpful to summarize the other two approaches. The reader interested in more details should consult Kupiec and O'Brien (1995a) or Bliss (1995). The standardized approach, very roughly, requires regulators to handle market risk in the same way credit risk is handled: Assets are categorized, and capital charges corresponding to the riskiness of each category are imposed. A criticism of this approach is that trading accounts are complicated and regulators do not have the resources or knowledge to thoroughly evaluate these complications. Consequently, any uniform formula would probably do a poor job of evaluating banks' risks.

In contrast, both the internal models approach and the pre-commitment approach try to use banks' superior knowledge and expertise to deduce appropriate capital levels. The internal models approach works, as the name suggests, by using banks' own models. Each bank's model is used to estimate a statistic called value-at-risk (VAR). Value-at-risk is a measure of potential losses. It satisfies the following condition: losses will only exceed it a given function at a time. For example, a 1 percent VAR of 3 million dollars means that losses will only exceed 3 million dollars 1 percent of the time.

In theory, the approach requires capital to be set equal to the 1 percent VAR. In practice, the approach calculates a 1 percent VAR for a ten-day trading period and then multiplies the result by three. Assuming the models are accurate, the multiple means that the percent level is substantially less than one. There are some other features to the approach such as checks on the quality of banks' models. The interested reader may consult the previously mentioned citations for more information. One criticism of this approach is that it discour-

[1] The Federal Deposit Insurance Corporation Improvement Act of 1991 (FDICIA) also uses capital requirements. FDICIA restricts bank activities and even allows for regulatory intervention if bank capital levels get low enough. See Spong (1994) for a summary.

ages banks from developing accurate models and instead encourages them to develop models that produce low capital levels. See Bliss (1995) for a good discussion and some poignant criticisms.

1. MOTIVATION

What will happen if the pre-commitment approach is implemented? One way to find out is to try the proposal. While actual results provide the ultimate judgment on a policy, economy-wide regulatory experiments are expensive and risky. It is preferable to first use inexpensive and safer alternative sources of information. One such source, though indirect, is examples (or lack thereof) of contractual arrangements similar to the pre-commitment approach used in different settings. Still, in the absence of an actual experiment, the best source of information is to hypothesize the results, that is, to perform thought experiments. The most rigorous thought experiments use mathematical models. A good model increases knowledge of what might happen when an untried policy is implemented.[2]

This paper contains one such thought experiment conducted on the pre-commitment approach. A private information model of banking capital regulation is developed to argue the following points. First, the proposal may be interpreted as a *menu* of contracts, a well-established economic concept. Several examples of their use outside the banking industry are provided. Second, under the assumptions laid out in this paper, menus are beneficial. Third, there are principles underlying the optimal design of fine schedules. Proper use of these principles can minimize the distortions to capital holdings caused by private information. Fourth, schedules in which fines are assessed only when there are losses (as the proposal presently specifies) will potentially need to be large to be effective. Still, there are a few issues not directly addressed by the model. These issues and possible extensions of the model that can address them are discussed later.

One objective of the paper is to foster a better understanding of the basic concepts underlying the pre-commitment approach. Part of this basic understanding requires elucidating what the pre-commitment approach is *not*. In particular, the approach's use of incentives may give the mistaken impression that the approach is a plan for deregulation. For example, Allen (1996) describes the approach with the term "self-regulation." While it is true that banks choose their level of capital (so they "regulate" themselves in the same sense that one would set a thermostat), their choice is made under an explicit set of rules and penalties. In this sense the pre-commitment approach is just as much a regulatory scheme as the other approaches. It is most definitely not a

[2] See Lucas (1980) for a statement of this methodological view.

proposal for deregulation. Instead, it is a plan to alter and improve banking capital regulation.

2. MENUS OF CONTRACTS

Underlying the pre-commitment approach is the well-established economic principle that it may be desirable for economic agents (banks in this application) to choose from a *menu* of contracts. In this paper, banks each choose an item from a menu designed by the regulator. In the context of the pre-commitment approach, an item on the menu consists of a capital level and an associated fine schedule. For example, the menu could consist of two items, one with low capital requirements and high fines, and another with high capital requirements but low fines.

Menus of contracts are pervasive and have been studied extensively by economists. Examples of such menus include those presented by insurance companies to potential customers. Each company offers a number of combinations of a premium with contingent payments. The contingencies are usually identifiable events, such as a fire or an accident, and the payments are limited by deductibles and co-payments. Faced with the menu, the customer usually chooses a single combination from the menu of choices. A fundamental issue for the design of insurance contracts, and for this paper, is that customers know their risks better than insurers, so the menus must be carefully designed with this point in mind. Otherwise, as in the case of life insurance, a life insurer that does not price its policies properly may end up only selling policies to high-risk individuals. This problem is called adverse selection in the insurance literature.

Some public utilities also use menus. Wilson (1993) contains a striking description of the rate schedules of the French electrical utility *Electricité France*.[3] In addition to differentiating its rates among observable features of its customers, such as residential versus commercial use, this utility also allows customers to choose from several options. For example, after paying a fixed charge based on the power rating of their appliances, professional offices face a menu that includes a further fixed charge and a per-unit-of-usage charge that depends on the time of day.

The menu contains three items: basic, empty hours, and critical times options. The basic option charges a fixed monthly fee and a fee per kilowatt hour consumed that is independent of the time of day. The empty hours option differs from the basic option by charging a 25 percent higher fixed fee but offers a 50 percent discount on usage during off hours. The remaining item on the menu is the critical times option. Compared with the basic option, its fixed fee is 50 percent lower and its per-usage fee is 36 percent lower. Unlike the basic

[3] His description is based on its rate schedules as of February 1987.

option, however, the critical times option contains one important contingency; if the utility announces that power is in short supply, there is an 800 percent surcharge on energy usage!

The schedules described in Wilson (1993) are clearly designed to separate customers by their power needs. The critical times option seems designed to select consumers who can afford to shut down their office on short notice, while the empty hours option is designed for customers with off-peak demand. No doubt, the design of the menu has something to do with the short-run capacity of electrical production and associated problems of peak usage.

Features of the Model

The following four features of the model underlie the use of menus in this paper:

- Several types of banks, which differ in the probability distribution of the returns on their assets;
- Private information on bank types, i.e., the assumption that a bank knows its probability distribution of returns but the regulator does not;
- A regulator who desires banks' capital levels to depend on bank type;
- A regulator who has the ability to levy fines on banks.

The first feature, a heterogeneous group of banks, is necessary because otherwise all banks would hold the same amount of capital, eliminating the need for a menu. The second feature, private information, is necessary because without it, that is, if both the bank and the regulator know the quality of the bank's portfolio, the regulator could simply figure out each bank's capital level himself. In contrast, with private information the regulator cannot arbitrarily control the actions of banks but instead may only indirectly influence them by setting penalties based on variables, such as bank returns, that the regulator can observe. The assumption of private information seems realistic because it corresponds to the idea that banks are better at assessing their portfolio than regulators.

The third feature, the desirability of heterogeneous capital levels, creates a potential conflict with banks' behavior under the assumption of private information. Private information hides the fundamental characteristics of the bank from the regulator. In this way, private information can be a problem if a regulator tries to set capital levels that depend on banks' types. For example, suppose the regulator wanted one type of bank to hold more capital than another type and suppose that banks prefer to hold less capital to more. Because the regulator is ignorant of bank types, a bank that is supposed to hold the higher amount of capital could post the lower amount instead. The regulator would be powerless to do anything about it since as mentioned, he could not distinguish one type of bank from another.

Differential capital levels may be feasible, however, when combined with fines, the last essential feature of the environment. The implementation works by letting the regulator provide banks with a menu of contracts. Each item on the menu consists of a capital level and an associated fine schedule. The idea is that it may be possible (and desirable) to design the fine schedules to affect each type of bank differently. The differential effect may be enough to get each type of bank to hold the amount of capital the regulator desires for it. Exactly how the menu needs to be designed will be elaborated later.

3. THE MODEL

In order to illustrate menus of contracts as clearly as possible, a simple model is described. The model leaves out several realistic features of the banking system. In particular, the important moral-hazard problem of bankers taking on too much risk because of deposit insurance is left out. The reason for this omission is that the goal of this paper is to describe as clearly as possible how menus of contracts work, and it is private information on bank types, not moral hazard in bankers' actions, that underlies the use of menus of contracts. Moral hazard could be included, and more will be said later on how to do this, but only at the cost of considerable complication.

Environment

Imagine the following banking system, which possesses the previously described four features. A bank's type is θ; there are two types of banks, called θ_1 and θ_2 (or type-1 and type-2), that differ in the riskiness of their portfolios. Assume type-1 banks are riskier than type-2. There is a continuum of banks, and each type of bank comprises a positive fraction of the banking sector. Let $h(\theta)$ denote the positive fraction of the banking sector consisting of type-θ banks, where, of course, $h(\theta_1) + h(\theta_2) = 1$. Both the regulator and the banks know the distribution of bank types, $h(\theta)$. Each bank's type is private information: it knows the riskiness of its portfolio but the regulator does not.

For simplicity, assume that each bank has an equally sized fixed base of deposits. Bank assets produce returns, q, which are *net* of payments to its depositors. Returns, q, may be positive or negative and are a function of a bank's type and an idiosyncratic, that is bank-specific, shock.[4] Each bank's return is distributed according to the probability function $p(q|\theta)$. Type-1 and type-2 banks differ in the distribution of their returns. Unlike its type, a bank's return is public information, that is, observed by both itself and the regulator.

Before realizing returns, banks hold capital, k, which costs them $r > 0$ per unit. Capital is not invested but simply sits in the bank and is repaid at the end

[4] Because there is a continuum of banks there is no aggregate uncertainty in the economy. Later there will be a short discussion of what might happen if aggregate uncertainty is included.

of the period.[5] Regulators have the power to fine banks, $f \geq 0$, after returns are realized. Restricting fines to be nonnegative precludes regulators from making transfers to banks.

Preferences

Assume that banks are risk neutral and that their sole objective is to maximize profits (returns net of fines and the cost of capital). Their utility function is

$$U(f, q, k) = q - f - rk,$$

so expected utility for a type-θ bank is

$$\int_q p(q|\theta)U(f, q, k)\, dq.$$

Since utility equals returns minus fines and capital costs, it is possible in this model for utility to be negative. The pre-commitment approach focuses solely on the trading portfolio, so conceivably losses resulting from bad performance and from fines would be paid from the rest of the bank portfolio. In this case, regulators should consider the effect fines would have on the rest of the bank's portfolio. The model, like the proposal, does not explicitly confront this problem. The model, however, does indirectly address the problem through its treatment of negative profits. More will be said about this point later, when discussing bankruptcy.

Negative utility becomes more problematic if, as recent discussions of the proposal have suggested, the approach is expanded to other sources of risk like credit risk. (See Seiberg [1996].) For example, if the proposal is extended to the bank's entire asset portfolio, then there is no longer a "rest of the bank" to obtain funds from. Limited liability constraints will bind, limiting the regulator's ability to impose fines. These concerns can be incorporated into the model, though it does complicate some of the analysis. More will be said about limited liability later.

Allocations

The model can be solved to determine the optimal *allocation*. An allocation is a statement of two things: how much capital each type of bank posts and how much a type-θ bank is fined if it produces return q.

Definition 1 *An allocation in this model is a function $k(\theta)$ describing capital holdings and a function $f(q, \theta)$ describing fine schedules.*

[5] Needless to say, this model is not based on a sophisticated theory of bank capital structure. The model does, however, provide a simple non-Modigliani-Miller economy, where banks want to hold less capital than regulators want them to. This latter feature is consistent with the prevailing view that deposit insurance leads to excessive leveraging of banks.

In this model, an allocation is equivalent to the menu of contracts. The two items on the menu are the two pairs of functions, $(k(\theta_1), f(q, \theta_1))$, and $(k(\theta_2), f(q, \theta_2))$. Each item consists of a capital level and a fine schedule.

The Approach

This private-information problem is analyzed by solving a constrained-minimization program. Economists solve these programs to find an allocation that minimizes an objective function while satisfying a set of constraints.[6] An objective function is a way of ranking alternative allocations according to some criterion. Constraints are conditions that allocations must satisfy in order to be feasible. For example, if the economy contained a limited supply of a raw material, there needs to be a constraint that in the aggregate, firms do not use more than the total supply of the raw material. In this paper the constrained-minimization program represents the problem facing a regulator who is designing capital regulations to further society's objectives given the limitations imposed by constraints on the regulator's and the banks' behavior.

Objective Function

The objective function is the total cost of capital used by the banking system. The goal is to find a feasible allocation that minimizes the objective function's value. Admittedly, the total cost of capital is a simple measure of social welfare, but it makes sense in this context for the following reason. Since the distribution of bank types is fixed and banks do not undertake any investment, allocating the returns is simply a transfer among the participants. Rather than specifying what happens to fines or how bank profits are distributed to consumers, it is simplest to ignore these distributional issues. Consequently, attention is focused on the resource cost in the economy, the cost of capital. The idea is that the cost of capital represents the opportunity cost of alternative uses of capital outside the banking system. Accordingly, the objective function is

$$\sum_{\theta} h(\theta) r k(\theta).$$

Regulator's Constraint

The perspective underlying this model is that the regulator designs the capital regulations to minimize banking capital. The earlier discussion, however, argued that the regulator wants to protect the insurance fund. This desire is modeled by requiring that the regulator limit the number of bankruptcies in the economy. Besides protecting the insurance fund, other reasons for this behavior might include preventing potentially harmful systemic events like a

[6] Actually, economists usually maximize an objective function, but in this model minimization is appropriate.

banking panic or even avoiding the political repercussions from too many bank failures. These concerns are modeled by simply requiring that the regulator set capital levels so that no more than a fraction, α, of banks fail. Bankruptcy is defined as an event in which losses exceed capital.

Definition 2 *A bank is bankrupt if losses exceed capital, that is, if* $q + k < 0$.

The regulator's constraint is written

$$\sum_{\theta} h(\theta) \int_{q+k(\theta)<0} p(q|\theta)\, dq \leq \alpha. \tag{1}$$

The term $\int_{q+k(\theta)<0} p(q|\theta)\, dq$ is the fraction of type-θ banks that fail.

Constraints on Fines

In this model, fines only transfer resources among members of the economy and do not enter the objective function. As a consequence, large fines could be imposed to enforce capital allocations. To avoid this possibility, and to capture the idea that there are limitations or costs to imposing fines, explicit restrictions on fines are imposed. Individual fines are limited to be no more than a fixed amount, $\bar{f}$. Since fines must also be nonnegative, each fine $f(q, \theta)$ is then required to be in the range

$$0 \leq f(q, \theta) \leq \bar{f}. \tag{2}$$

For similar reasons, the total amount of fines that can be assessed on the banking sector are not allowed to exceed $\bar{F}$, that is,

$$\sum_{\theta} h(\theta) \int_{q} p(q|\theta) f(q, \theta)\, dq \leq \bar{F}. \tag{3}$$

Again, the goal of these constraints is to limit the imposition of fines. If the level of any particular fine is too high, then limited liability concerns, as discussed earlier, need to be considered explicitly. Furthermore, if total fines are too large, no one would run a bank. These constraints are a crude but convenient way of limiting fines. A more realistic alternative would be to assume that total fines may be assessed only in an amount equal to insurance fund payments to depositors of failed banks.[7]

[7] This latter specification was studied under the assumption of limited liability. Unfortunately, it complicated the analysis and hid some of the basic insights of menus of contracts. In particular, the parameters $\bar{F}$ and $\bar{f}$ become functions of the capital levels. Yet another specification is to make banks risk averse and put bank utility in the objective function. This specification avoids the extremely high fines that are characteristic of models with risk neutrality; but it produces the unappealing result that the regulator is insuring banks (not just depositors) and doing so by often making transfers *to* them.

Incentive Constraints

Incentive constraints take into account the effect of private information. To see how private information restricts the set of feasible allocations, consider the following allocation, which is assumed to satisfy constraints (1), (2), and (3). Set $k(\theta_1) > k(\theta_2)$ and set $f(q, \theta) = 0$ for all returns q and types θ. This allocation makes capital depend on a bank's type but never fines banks. If bank types are known, that is, they are not private information, then this menu of contracts could be implemented by fiat. The regulator simply orders each type-θ bank to hold capital level $k(\theta)$.

Now consider the same allocation, but under the assumption that bank types are private information. Since the regulator does not know a bank's type, it cannot order a bank to hold $k(\theta)$. After all, a bank of one type could simply claim to be a different type. Instead, the regulator must induce banks to hold $k(\theta)$ by letting them choose from a menu of contracts.

Under private information, a type-1 bank is faced with the following decision: Does it choose a type-1 or a type-2 contract? The answer in this case is that it chooses the type-2 contract, as the following equation demonstrates:

$$\int_q p(q|\theta_1)q\,dq - rk(\theta_1) < \int_q p(q|\theta_1)q\,dq - rk(\theta_2). \tag{4}$$

The left-hand side of equation (4) is the utility of a type-1 bank that posts $k(\theta_1)$ units of capital. This level is less than the right-hand side of the inequality; that is, the utility of the same bank if it pretends to be a type-2 bank and posts $k(\theta_2)$ units of capital. Thus, this allocation is not feasible if there is private information because no type-1 bank acting in its self-interest would ever hold the higher level of capital.

Economists ascertain which allocations are feasible under private information by using the *revelation principle*. The revelation principle says that it is sufficient to consider a menu of contracts with one item for each type as long as the menu, or equivalently the allocation, is *incentive compatible*. As a matter of convenience, economists index each item on the menu by the θ of the type-θ bank choosing that item. Because economists index each item by the type choosing it, incentive constraints are sometimes called truth-telling constraints.

Definition 3 *In this model an allocation is incentive compatible if*

$$\int_q p(q|\theta_1)(q - f(q, \theta_1))\,dq - rk(\theta_1) \geq \int_q p(q|\theta_1)(q - f(q, \theta_2))\,dq - rk(\theta_2), \tag{5}$$

and

$$\int_q p(q|\theta_2)(q - f(q, \theta_2))\,dq - rk(\theta_2) \geq \int_q p(q|\theta_2)(q - f(q, \theta_1))\,dq - rk(\theta_1). \tag{6}$$

As the previous example suggested, incentive compatibility embodies the ability of banks to act in their own interest. Each incentive constraint is a way

of writing the maximization problem facing a bank. For example, a type-1 bank has two choices. It can claim to be a type-1 or a type-2 bank. Constraint (5) states that a type-1 bank prefers to claim it is a type-1 bank rather than a type-2 bank. If there was also a third type of bank, there would need to be four additional incentive constraints. One constraint would state that a type-1 bank prefers a type-1 allocation to a type-3 allocation. Another constraint would ensure that a type-2 bank prefers a type-2 allocation to a type-3 allocation. And two more constraints would be necessary to ensure that it is incentive compatible for the type-3 bank to claim to be a type-3.

Now that the description of the constraints is complete, all the pieces are in place to formally state the problem of finding a feasible allocation that minimizes the cost of capital used by the banking sector. For an allocation to be feasible, it must satisfy the following constraints: prevention of too many bankruptcies, limitations on the regulator's power to levy fines, and compatibility with banks' incentives. The optimal allocation in this economy will be the solution to the following constrained-minimization program.

The Constrained-Minimization Program

$$\min_{k(\theta)\geq 0,\, f(q,\theta)} \sum_{\theta} h(\theta)\, rk(\theta)$$

s.t. (1), (2), (3), (5), and (6).

The Solution

The analysis makes the following two assumptions about the distribution of bank returns, $p(q|\theta)$.

Assumption 1 *For all* $q < 0$, $p(q|\theta_1) > p(q|\theta_2)$.

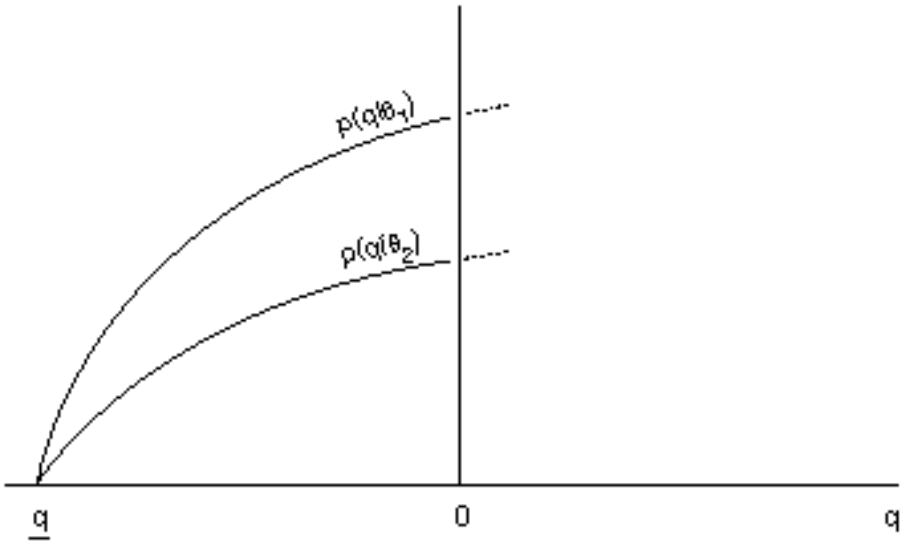
Assumption 1 says that there is a higher probability of each loss level for type-1 (risky) banks than for type-2 (safe) banks.

Assumption 2 *For all* θ, $p(q|\theta)$ *is increasing and weakly concave over the range* $q < 0$.

An example of a pair of probability functions that satisfy Assumptions 1 and 2 is illustrated in Figure 1. The figure shows only the probabilities for the portion of returns, q, that is negative.[8] These assumptions are one way of expressing the idea that type-1 banks are riskier than type-2 banks.

[8] The negative portion of the returns is important because it is all that matters for determining whether the set of allocations satisfying constraint (1) is convex.

Figure 1 A Pair of Probability Functions (Shown only over the Range $\underline{q} < q < 0$) that Satisfy Assumptions 1 and 2



Capital Allocation

Under Assumptions 1 and 2, there is a decreasing marginal decline in bankruptcies for both types of banks as their capital level increases. Furthermore, for a fixed level of capital, the marginal decline is higher for type-1 banks then type-2 banks. The relative sizes of the marginal declines in bankruptcies leads to the following proposition.

Proposition 1 *The optimal allocation must satisfy* $k(\theta_1) \geq k(\theta_2)$.

Proof: If $k(\theta_1) < k(\theta_2)$, then raise $k(\theta_1)$ and lower $k(\theta_2)$ such that they are equal and do not violate the regulator's constraint. Also, set all fines to zero. This allocation is trivially incentive compatible. But because of Assumptions 1 and 2, $h(\theta_1)k(\theta_1)$ is raised by less than $h(\theta_2)k(\theta_2)$ is lowered, thus lowering the total amount of capital in the system.

The proposition should be evident from inspecting Figure 1. A $k(\theta_1) = k(\theta_2)$ allocation is trivially incentive compatible and uses less total capital than a $k(\theta_1) < k(\theta_2)$ allocation. One useful implication of Proposition 1 is that incentive constraint (6) does not bind at an optimal allocation. In other words, a type-2 bank has no incentive to pretend that it is a type-1 bank, so this constraint can be ignored in the analysis.

For further analysis of capital levels it is necessary to study the first-order conditions to the program.[9] Let ν denote the Lagrangian multiplier on the regulator's constraint (1), and μ_1 the multiplier on the type-1 bank's incentive constraint (5). The first-order condition on $k(\theta_1)$ is

$$r + \nu p(-k(\theta_1)|\theta_1) - \frac{r\mu_1}{h(\theta_1)} = 0, \tag{7}$$

and on $k(\theta_2)$ it is

$$r + \nu p(-k(\theta_2)|\theta_2) + \frac{r\mu_1}{h(\theta_2)} = 0. \tag{8}$$

Equating the two first-order conditions and rearranging terms produces

$$-\mu_1 \frac{r}{\nu}\left(\frac{1}{h(\theta_1)} + \frac{1}{h(\theta_2)}\right) + p(-k(\theta_1)|\theta_1) = p(-k(\theta_2)|\theta_2). \tag{9}$$

Lagrangian multipliers on binding inequality constraints are positive so the first term in equation (9) is negative, which implies that $p(-k(\theta_1)|\theta_1) > p(-k(\theta_2)|\theta_2)$. The inequality means that at the solution the marginal decrease in bankruptcies from an increase in capital of type-1 banks is greater than that of type-2 banks.

It is easy to see the role of private information in determining fines and capital levels if the private-information solution is compared with the *full-information* solution. By full information it is meant that a bank's type is not only known by the bank but also by the regulator. In terms of the program, solving for the full-information optimum requires first removing the incentive constraints, equations (5) and (6). The first-order conditions for the full-information program are identical to (7) and (8) except now $\mu_1 = 0$. This drops the first term from equation (9). Letting $k_f(\theta)$ denote the full-information solution, the first-order conditions imply that $p(-k_f(\theta_1)|\theta_1) = p(-k_f(\theta_2)|\theta_2)$.

As the following proposition proves, the optimal full-information allocation is for type-1 (risky) banks to hold strictly more capital than type-2 (safe) banks.

Proposition 2 $k_f(\theta_1) > k_f(\theta_2)$.

Proof: Assumptions 1 and 2 and the first-order condition $p(-k_f(\theta_1)|\theta_1) = p(-k_f(\theta_2)|\theta_2)$ imply that $-k_f(\theta_1) < -k_f(\theta_2)$. Therefore, $k_f(\theta_1) > k_f(\theta_2)$.

While both the private-information and full-information models are characterized, in general, by risky banks holding more capital, the amounts differ. As the next proposition shows, type-1 (risky) banks hold less capital under private information than they do under full information, while the order is reversed for type-2 (safe) banks.

[9] First-order conditions are sufficient for finding the solution to a constrained-minimization problem when the objective function is weakly convex and the set of feasible allocations is convex. Both conditions are satisfied by this model. Satisfying the latter condition was the reason for making Assumptions 1 and 2.

Proposition 3 $k_f(\theta_2) < k(\theta_2) \leq k(\theta_1) < k_f(\theta_1)$.

Proof: First-order conditions imply

$$p(-k_f(\theta_1)|\theta_1) = p(-k_f(\theta_2)|\theta_2), \text{ and}$$
$$p(-k(\theta_1)|\theta_1) > p(-k(\theta_2)|\theta_2). \tag{10}$$

Since $p(q|\theta)$ is increasing over the range $q < 0$, for the regulator's constraint (1) to be satisfied by a private-information allocation there are only two possible capital allocations: Either, $k(\theta_1) < k_f(\theta_1)$ and $k(\theta_2) > k_f(\theta_2)$, which implies fewer type-1 banks and more type-2 banks fail. Or, $k(\theta_1) > k_f(\theta_1)$ and $k(\theta_2) < k_f(\theta_2)$, which implies more type-1 banks and fewer type-2 banks fail. Only the first option, however, satisfies equation (10), because $p(q|\theta)$ is increasing and weakly concave. The remaining claim of the proposition, $k(\theta_2) \leq k(\theta_1)$, was proven in Proposition 1.

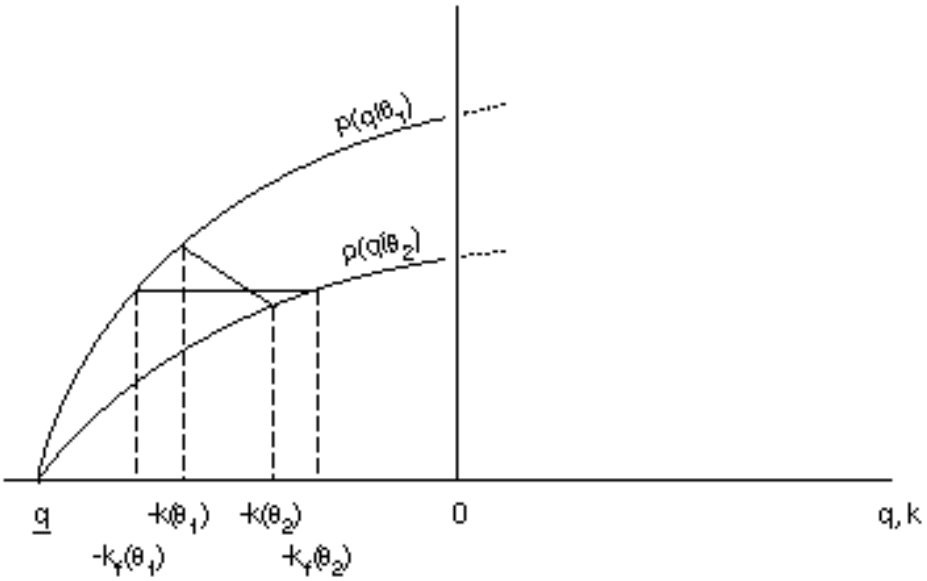
As Proposition 3 implies and Figure 2 summarizes, private information reduces the spread between $k(\theta_1)$ and $k(\theta_2)$. The private-information solution would like to replicate the full-information solution but it cannot because the full-information solution does not satisfy the incentive constraints. In fact, the full-information solution uses less capital than the private-information solution. (This result can be formally shown by using Proposition 3 and noting that the full-information problem is identical to the private-information problem except with fewer constraints.)

As a final point of reference, consider an allocation where $k(\theta_1) = k(\theta_2)$. This is an allocation where all types of banks are treated identically as in the standardized approach discussed in the introduction. Proposition 1 does not prove that the private-information solution is always better than treating all banks identically, but it suggests that it usually is. Later, a numerical example will be provided where, indeed, it is better. The value of using the pre-commitment approach relative to the standardized approach will be calculated from the objective function. It will be the difference in total capital used by the private-information solution and the total capital used by the best allocation where all banks hold the same level of capital. The measure will depend on the ability of the regulator to spread out capital allocations by using fines.

Fine Schedules

The previous analysis showed that, despite private information, capital levels may still differ across bank types. Any difference has to be supported by a fine schedule which discourages type-1 banks from pretending to be type-2 banks. Equation (3) limits the total amount of fines which may be assessed on the banking sector. If this constraint binds, then precisely how fines are assessed is important.

Figure 2 Capital Allocations for the Private-Information and Full-Information Models



The first observation regarding fines is that since only incentive constraint (5) binds, fines need to be assessed only on banks which claim they are type-2 (safe) banks. At first, it might not seem intuitive that only the safe banks are fined. But it makes sense when it is realized that safe banks get the benefit of lower capital levels. This consideration necessitates penalties to dissuade type-1 banks from pretending they are type-2 banks. True, it is possible to fine both types of banks, but any fine on banks declaring themselves to be type-1 would be wasted since type-2 banks have no incentive to pretend they are type-1 banks. Fines on type-1 banks would only give them an additional incentive to pretend they are type-2 banks. This discussion is summarized in the following point about optimal fine schedules.

Lesson 1 *Not all bank types need to be subject to fines. In particular, banks which post the highest level of capital do not need to be fined.*[10]

[10] With more than two types there can be varying degrees to which banks are fined, as the numerical example in the following section illustrates. Furthermore, if moral hazard is added, then it might be necessary to fine all banks for one return or another. Still, even with the addition of moral hazard, different types of banks would be fined to varying degrees.

The next issue is how the regulator should assess fines on type-2 banks. The answer can be deduced from the constraints on fines, (2) and (3), along with the binding incentive constraint (5). The right-hand side of the incentive constraint (5) shows that a type-1 bank receives expected *disutility* from declaring it is a type-2 bank in the amount

$$\int_q p(q|\theta_1)f(q,\theta_2)\,dq. \tag{11}$$

The size of this term is important because it has to be large enough to convince type-1 banks to decline the benefit from choosing the lower capital level.

For each realization of the return, q, the penalty has two components, the size of the fine and the probability the fine will be imposed. Since the bank is risk neutral, it does not care whether the size of the fine or the probability of the return is high. It cares only that the sum of their products is high. Banks heed only the expected value of fines, not their distribution across returns.

The distribution of fines across returns does matter, however, to the regulator. The first, and most obvious, way it matters is that fines can be no more than $\bar{f}$. There is also a second, less obvious, way in which the distribution of fines matters to the regulator. Constraint (3) limits the total amount of fines the regulator may impose in equilibrium. Because of this constraint, a fine of $f(q,\theta_2)$ lowers the amount of fines the regulator may assess if other returns are realized. This quantity is lowered by

$$p(q|\theta_2)f(q,\theta_2). \tag{12}$$

This product depends on $p(q|\theta_2)$ and not $p(q|\theta_1)$ because in *equilibrium* only type-2 banks receive fine $f(q,\theta_2)$. Remember, incentive compatibility requires that fines are set so that type-1 banks never pretend to be type-2 banks (or vice versa). The distribution of fines matters because the fine schedule $f(q,\theta_2)$ is multiplied by $p(q|\theta_1)$ in incentive constraint (5), but it is multiplied by $p(q|\theta_2)$ in fine constraint (3).

The trade-off between the fine's effect on the incentive constraint and its effect on the fine constraint can be measured by the deterrent effect per unit of assessed fine.

$$\frac{p(q|\theta_1)f(q,\theta_2)}{p(q|\theta_2)f(q,\theta_2)} = \frac{p(q|\theta_1)}{p(q|\theta_2)}.$$

The quotient on the right-hand side of the equation is often called a *likelihood ratio*. It is very important in private-information models, and it is an important point of this analysis.

Lesson 2 *Fines are best assessed on returns, q, with the highest likelihood ratio $\frac{p(q|\theta_1)}{p(q|\theta_2)}$.*

Lessons 1 and 2 suggest the best way for the regulator to assess fines. First, only fine banks declaring themselves to be type-2 banks. Next, set the fine

$f(q, \theta_2)$ as high as possible on the return, q, with the highest likelihood ratio $p(q|\theta_1)/p(q|\theta_2)$. Once the maximum fine, $\bar{f}$, is reached, then the regulator should set fines as high as possible on the return with the next highest likelihood ratio. This procedure should be continued until no more fines can be levied.

4. A NUMERICAL EXAMPLE

This section reiterates the lessons of the previous analysis by presenting a numerical example. The example shows how private information distorts allocations and how likelihood ratios influence fines. It also shows how it may be beneficial, despite private information, to differentiate banks by type. This is done by comparing the private-information solution with an allocation in which all banks hold the same level of capital. As discussed earlier, this latter allocation can be viewed as the standardized approach, though for reasons to be discussed later, the analogy leaves out at least one important feature of that approach.

It should be noted that the numbers used in this example are *not* drawn from any data but instead are purely hypothetical. Thus, the quantitative implications of the example, that is, the size of fines and the size of welfare costs, do not describe the actual economy. Instead, the results should be viewed as emphasizing the qualitative properties of the models.

This example adds a third type of bank, θ_3, to the previous analysis. As noted earlier, the addition of a third bank type only requires that more incentive constraints are added to the constrained-minimization program. As before, type-2 banks are safer than type-1 banks, but now, type-3 banks are the safest of all. The three types comprise the following fraction of banks in the banking system:

$$h(\theta_1) = 0.3, \qquad h(\theta_2) = 0.3, \qquad h(\theta_3) = 0.4.$$

Only a fraction $\alpha = 0.06$ of the banks may fail, and the cost of capital for banks is $r = 0.12$. Total fines $\bar{F}$ are restricted to be less than or equal to 0.04. Fines imposed on returns, $\bar{f}$, are limited to be less than 0.1, though this latter constraint will not bind in equilibrium.

The assumed probability functions, $p(q|\theta)$, for the three types of banks are as follows:

	$q \in [-1, 0]$	$q \in (0, 1]$	$q \in (1, 2]$	$q \in (2, 3]$
$p(q\|\theta_1)$	$0.6q + 0.6$	0.35	0.20	0.15
$p(q\|\theta_2)$	$0.3q + 0.3$	0.10	0.45	0.30
$p(q\|\theta_3)$	$0.2q + 0.2$	0.10	0.20	0.60

The functions can be broken into two parts: probabilities on negative returns and probabilities on positive returns. The probability on negative returns,

$q \in [-1, 0]$, that is, $-1 \leq q \leq 0$, is the only portion of the distribution which matters for the bankruptcy constraint. For each type of bank, the probability function increases linearly over this range. As indicated in Figure 3, these functions satisfy Assumptions 1 and 2. For the positive returns, the probability functions are linear but discontinuous. For example, $p(q \in (0, 1]|\theta_1) = 0.35$ means that there is a 35 percent chance that a type-1 bank's return will fall in this range and that each return within this range is equally probable.

The following table lists computed optimal capital levels for all three models. The first three rows list the capital holdings for each type of bank. The fourth row lists the total capital held by the banking system. If scaled by the cost of capital, the fourth row is also the value of the objective function at the optimal allocation. The first column of numbers, denoted by "Stand." represents the standardized approach; that is, all banks are treated identically by requiring them to hold the same capital level. The second column denotes the private-information model, when the regulator can offer a menu of contracts, while the third column lists capital allocations under the full-information model.

	Stand.	**Priv. Info.**	**Full Info.**
θ_1	0.415	0.484	0.690
θ_2	0.415	0.420	0.380
θ_3	0.415	0.278	0.077
Total	0.415	0.382	0.352

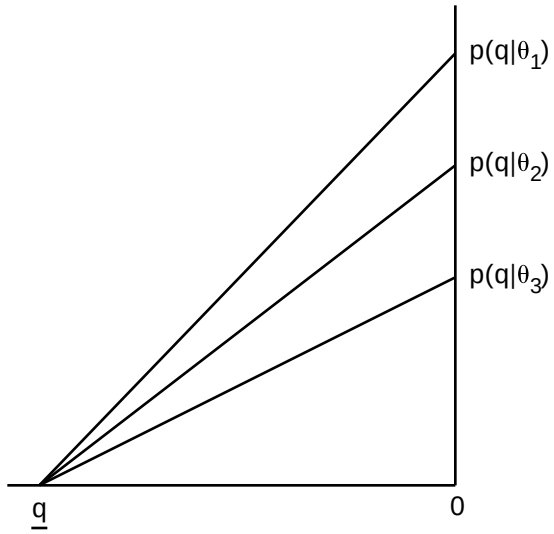
The table contains two implications. First, moving from left to right, total capital decreases over successive models, as it should, since allocations are less constrained with each model. Second, the profile of capital levels changes. The private-information allocation spreads out capital levels, compared to the standardized allocation, but not as much as the full-information solution, which is consistent with Proposition 2.

Private information also affects the distribution of bank failures. The next table lists failure rates for each type of bank.

	Stand.	**Priv. Info.**	**Full Info.**
θ_1	0.103	0.080	0.029
θ_2	0.051	0.051	0.058
θ_3	0.034	0.052	0.085
Total	0.060	0.060	0.060

Again, the choice of model has a sizable effect on the distribution of bank failure rates. Under the standardized approach, type-1 banks, the riskiest ones,

Figure 3 Probability Functions (Shown over the Range $\underline{q} < q < 0$) Used in the Numerical Example



have the highest failure rate. For each increasingly safer type, the fraction of failures decreases. Under private information, this ranking slightly changes, and failure rates for all three types are bunched closely together. The full-information solution spreads out failure rates across types but in a different direction than the standardized solution. Under full information, type-3 (safe) banks fail the most.

As this paper has consistently emphasized, a schedule of expected fines is necessary to implement menus of contracts under private information. (Remember, fines are not required for the other two models.) The following table shows the calculated optimal fine schedules as a function of the report, θ, and the return, q, for the private-information model.

θ	$q \in [-1, 0]$	$q \in (0, 1]$	$q \in (1, 2]$	$q \in (2, 3]$
θ_1	0.0	0.000	0.000	0.000
θ_2	0.0	0.022	0.000	0.000
θ_3	0.0	0.053	0.031	0.000

In reporting the fines, the convention was adopted to assess fines in equal amounts within each interval of positive returns. The reason for making this assumption is that within each range the likelihood ratio is a constant, so

there is an indeterminacy in how fines are allocated within each of these ranges. Not varying the fine within each range seems to be the simplest way of presenting the results.

As the earlier analysis showed, fines are never assessed on a bank which declares itself to be type-1, the riskiest type. The reason is that type-1 banks post the highest capital level, so the other banks have no incentive to claim to be type-1. Type-1 banks, however, have incentives to declare themselves type-2 or even type-3 banks.

As the earlier analysis also showed, fines are effective when levied on the return with the highest likelihood ratio, $\frac{p(q|\theta_1)}{p(q|\theta_2)}$. Consequently, fines are assessed if a bank claims to be type-2 and the return is $q \in (0, 1]$. For this same return, banks declaring themselves to be a type-3 bank are also fined. The reason is again to preclude type-1 banks from declaring themselves to be a type-3 bank.

Along with type-1 banks, type-2 banks have an incentive to post a lower capital level, though in their case, they must only be prevented from declaring themselves to be a type-3 bank. Again, likelihood ratios provide a guide for the best way to arrange fines. Fines are most effective against type-2 banks when imposed on returns with $q \in (1, 2]$.

It should be noted that, in general, the addition of a third type of bank complicates the analysis. For example, fines which help one incentive constraint bind might weaken another. In general, optimal fines need not take the exact form of the schedules listed in the table, although any optimal fine schedule will make extensive use of the likelihood ratios.

5. THE PRE-COMMITMENT APPROACH

What does the previous analysis say about the conceptual basis of the pre-commitment approach? It makes a clear statement in support of menus of contracts. Menus of contracts reduce the amount of capital used in the system by allocating it more efficiently across types of banks. Insofar as the pre-commitment approach is a menu of contracts, it is based on conceptually firm economic grounds.

What does the previous analysis say about the specifics of the pre-commitment approach? Remember, the proposal advocates fining banks only when losses exceed capital. One of the lessons of the previous analysis was that fines should be imposed when a high likelihood ratio exists, regardless of whether or not losses exceed capital. Another lesson was that the size of the fine should depend on the amount of capital posted. Lower capital levels require higher fines to preserve incentive compatibility, while higher capital levels require fewer fines. The pre-commitment approach is silent on this issue. In the context of the model, the approach's proposed fine schedule is not optimal.

Now, it might very well be that the proposal's schedule of fines, while not optimal, works reasonably well. After all, optimality is a statement about

ranking alternatives, not about the absolute size of any differences in the value of the objective function. To make this latter assessment requires numbers based on data, particularly data on the distribution of returns. That exercise is outside the scope of this paper. Nevertheless, there is enough information to obtain an idea about the *quantitative* size of fines required to implement the proposal's fine schedule. This calculation at least provides some sense of the quantitative implications of the proposal.

To pursue this aim, consider an economy, much like the one in the numerical example, with only two types of banks, one riskier than the other. As before, the risky bank is indexed by θ_1 and the safer bank is indexed by θ_2. Also, assume that $k(\theta_1) > k(\theta_2)$ is desired by the regulator.

It is convenient to divide the range of losses into two portions. Let q_1 denote losses exceeding $-k(\theta_1)$ and let q_2 denote losses exceeding $-k(\theta_2)$ but not $-k(\theta_1)$, that is, $q_1 < -k(\theta_1) \leq q_2 < -k(\theta_2)$. The purpose of making this division is that, in the proposal, banks are fined only if losses exceed capital, which means type-1 banks are fined only when q_1 is realized, while type-2 banks are fined if either q_1 or q_2 is realized. Also, let $q_{1,2}$ denote the range consisting of both q_1 and q_2.

Recall that type-2 banks hold less capital than type-1 banks, so assume that the only binding incentive constraint is on type-1 banks, equation (5). Furthermore, to be consistent with the proposal, fines are set to zero, $f(q,\theta) = 0$, if losses do not exceed capital, $k(\theta)$. Any return for which this is the case has no effect on the incentive constraint and does not need to be written down explicitly. The incentive constraint, after removing the terms equaling zero and subtracting out expected returns, is

$$\int_{q_1} p(q|\theta_1)(-f(q,\theta_1))\,dq - rk(\theta_1) \geq \int_{q_{1,2}} p(q|\theta_1)(-f(q,\theta_2))\,dq - rk(\theta_2). \quad (13)$$

Again, the left-hand side is the utility a type-1 bank gets from telling the truth, while the right-hand side is its utility from lying (after subtracting out expected returns).

At this point, it is helpful to introduce some new notation. First, define $\Delta k = k(\theta_1) - k(\theta_2)$ as the difference in capital levels. Next, let $f_{av}(q_{1,2},\theta_2)$ be the average deterrent effect of fines over the range $q_{1,2}$. More precisely,

$$f_{av}(q_{1,2},\theta_2) = \frac{\int_{q_{1,2}} p(q|\theta_1) f(q,\theta_2)\,dq}{\int_{q_{1,2}} p(q|\theta_1)\,dq}.$$

A constant level of fines set at this value over the range $q_{1,2}$ would be enough to preserve incentive compatibility. This calculation is useful for obtaining some idea about necessary magnitudes of fines.

Now, the incentive constraint (13) can be rearranged to obtain

$$f_{av}(q_{1,2}, \theta_2) \geq \frac{\int_{q_1} p(q|\theta_1) f(q, \theta_1)\, dq}{\int_{q_{1,2}} p(q|\theta_1)\, dq} + \frac{r\Delta k}{\int_{q_{1,2}} p(q|\theta_1)\, dq}. \tag{14}$$

The average deterrent effect of the fines must be no less than the sum of the two benefits a type-1 bank obtains from claiming to be a type-2 bank: the gain from no longer receiving fines, $f(q, \theta_1)$, plus the lower capital costs, $r\Delta k$. Both terms are divided by the probability that the average fine is imposed, $\int_{q_{1,2}} p(q|\theta_1)\, dq$.

The size of the first term depends on the distribution of losses, something on which this paper has little to say. Still, its sign is positive, so the second term provides at least a lower bound on the size of the average fine. This lower bound is

$$f_{av}(q_{1,2}, \theta_2) > \frac{r\Delta k}{\int_{q_{1,2}} p(q|\theta_1)\, dq}. \tag{15}$$

At a minimum the fine has to be large enough to offset any cost savings from a type-1 bank posting a lower capital level. Kupiec and O'Brien (1995b) contains a similar equation.

To assess the size of fines requires several parameter values. Two, which the regulator will set, are the probability that losses will exceed the chosen capital level and the time frame for evaluating the portfolio. Because of the preliminary state of the proposal, it is not clear what parameter values shall be used. Still, the discussion in the literature seems to focus on setting a capital level such that losses occur no more than 1 to 5 percent of the time. For example, see Kupiec and O'Brien (1995c), Bliss (1995), or Marshall and Venkataraman (1996).

Probably a better source of information is the criteria regulators will actually use in the internal models approach. This approach uses a 1 percent criterion over a ten-day trading period. However, it then multiplies by three the number produced by the bank's VAR model. This multiplication means that in practice the percent criterion is substantially less than one. The exact number depends on the tail of the distribution and would seem difficult to ascertain.

Despite the inability to obtain specific numbers, the internal models approach indicates two features the chosen numbers need. The time period should be relatively short and capital should be set so capital rarely exceeds losses. Accordingly, for the following calculations assume that the time frame is a quarter and set the probability of losses $\int_{q_{1,2}} p(q|\theta_1)\, dq$ over a range of 0.005 to 0.03. The time frame is longer than that which the internal models approach uses but the probability of a loss is higher. As a starting point, these numbers seem as good as any others.

The remaining number, the cost of capital, is more difficult to estimate. The model, while effective for illustrating menus of contracts, is not a good theory of capital structure. In the model, capital is only invested in a riskless storage technology and is used to satisfy claims in the event of a loss. In reality, the

cost to a bank of a different equity structure is the change in the bank's value. Its value may depend on the equity structure because of deposit insurance or it may depend on other factors often cited by the corporate finance literature such as taxes, bankruptcy, or managerial incentives. Consequently, rather than taking a stand on a particular number, calculations are made for a range of possible numbers.

Say the lower bound on the quarterly cost of capital is 0.5 percent. For an upper bound, the real cost of equity capital for banks is used. Kuprianov (1997) calculates a nominal cost of equity capital to be 14.5 percent in 1995. In real terms, this is close to 12 percent, or 3 percent quarterly.

The following table reports the average fine as a percentage of assets satisfying equation (15) for the ranges described above. The numbers are calculated for a 1 percent difference in capital level, Δk. The rows list the cost of capital, while the columns list the probability of losses exceeding capital.

	$\int_{q_{1,2}} p(q\|\theta_1)\,dq$			
r in %	**0.005**	**0.010**	**0.020**	**0.030**
0.5	1.000	0.500	0.250	0.167
1.5	3.000	1.500	0.750	0.500
3.0	6.000	3.000	1.500	1.000

Remember, the numbers in the table ignore the first term in equation (14), so they still might not be sufficient to implement the proposal.

Many of these numbers seem high. For example, with a quarterly cost of 1.5 percent and a 0.01 chance of a loss, the fine must be set at 1.5 percent of assets per unit of capital. If there is a 5 percent difference in capital levels between the two items on the menu, the average fine per *quarter* on safe banks must be 7.5 percent of assets. And this number ignores the first term in equation (14). Still, other numbers, particularly for low costs of capital, do not seem so unreasonable.

Equation (15) and the calculations presented in the table should be viewed as providing the following cautionary note to the proposal's fine schedule.

Lesson 3 *If fines are assessed only in low probability states, then they need to be set at high levels to offset the certain benefit of choosing a lower capital level.*

The potentially large size of required penalties, which is also noted in Kupiec and O'Brien (1995b), is a concern for the proposal. In its defense, Kupiec and O'Brien (1995c, 1995b) argue that other penalties, such as higher future capital requirements or increased supervision, may also be imposed. Since the model is static and penalties are pecuniary, the model says nothing

about these alternatives, though conceivably it could be modified to handle them. A dynamic variant might involve the repeated version of Program 2, randomly redrawing each bank's riskiness every period. The tools exist to handle this problem. For example, see Green (1987), Phelan and Townsend (1991), and Atkeson and Lucas (1992) for analysis of dynamic versions of other private-information problems. This modification of the model, however, is outside the scope of the paper.

Other Issues

The model in this paper abstracted from numerous issues not necessary to illustrate menus of contracts. Still, it is worthwhile to discuss what was left out and whether the issues not addressed by the model are important. One such issue is the previously discussed dynamic penalty schemes. This section lists several additional issues not addressed by the model, and for some of the issues it discusses how the model may be extended to analyze them.

Moral Hazard

Banks do not control their portfolios in this model, although of course they do so in reality. By changing their asset holdings, banks alter the distribution of their returns. As with bank types, it is reasonable to assume that many of the adjustments a bank makes to its portfolio are private information. It is these unobserved adjustments, plus their possible harmful effects, which have led to the use of the term moral hazard to describe these problems. Penalty schemes should be designed to handle moral hazard as well as bank heterogeneity.

One way to model banks' ability to alter their portfolio would be to modify the technology to $p(q|a, \theta)$, where a is a costly action taken by the bank, and along with θ, is not observed by regulators. This specification would incorporate moral hazard, which is usually associated with deposit insurance. Now, θ might be interpreted as the quality of the management.[11] There is a nonbanking literature on variants of this problem (see, for example, Christensen [1981], Laffont and Tirole [1986], or Prescott [1995]), but none of these models include capital or anything resembling it.[12] The addition of moral hazard will modify but not negate the messages of the three lessons. For example, Lesson 1 said that risky banks should not be fined. With moral hazard, however, it might be necessary

[11] Another option is to put θ into a bank's preferences and let it represent a bank's taste for risk.

[12] To be sure, there are a few banking papers that include capital. Giammarino, Lewis, and Sappington (1993) and Besanko and Kanatas (1996) both assume that returns are determined by $p(q|a + \theta)$. However, they assume that not only returns, q, are observable but also the sum $a + \theta$. Consequently, there is no need for return-dependent fines, which is a fundamental issue for the pre-commitment approach. Chan, Greenbaum, and Thakor (1992) separately analyze moral hazard and hidden information in a banking model with capital.

to fine risky (high capital) as well as safe (low capital) banks, but the relative size of the fines will differ across types.

Limited Liability

As discussed earlier, the model allowed for negative utility. If the bank experienced a loss, the regulator could still impose a fine. In practice, because of limited liability, if a bank experiences a loss, there are no assets to fine.

It is straightforward to add limited liability to the model, though it does complicate the analysis. Adding limited liability does not change this paper's message that menus of contracts may be valuable. However, adding limited liability does limit the ability of the regulator to assess fines. The limitation is particularly restrictive if a bank experiences a loss or a low return. If the proposal is to be extended to a bank's entire portfolio, then it will be necessary to fine banks when they do not experience a loss. The pre-commitment approach, as presently proposed, only assesses fines when there is a loss.

Aggregate Shocks

What would happen if there was a large shock to the market? Would regulators want to enforce fines? For example, consider a large number of banks trading in derivatives markets. If there was a large drop in market price, analogous to a stock market crash, banks might try to liquidate their holdings to avoid future fines. Such liquidation could cause further price declines if the banks comprised a large enough portion of the market.

In the context of the model, an aggregate shock could be included by indexing fines by the aggregate shock, ε, in addition to bank-specific shocks. The fine schedule would be written $f(q, \theta, \varepsilon)$ and might contain contingencies reducing fines on banks if the loss on the portfolio was due to an aggregate shock as opposed to a bank-specific shock. This sort of schedule contains relative performance features, where banks are compared with a market aggregate.

Time Inconsistency

Can regulators commit to imposing fines? Committing to future actions may be difficult. For example, consider the taxation problem facing a government at any particular point in time. At that moment, it seems optimal to tax all existing capital and to promise never to tax capital in the future. That way, there are no economic distortions since the initial capital stock is inelastically supplied and the promise to not tax future capital gives people an incentive to invest. However, the next period the government will face the same problem and tax all the capital in that period. If the government taxes this period, then people will realize that the government could make the same promise the next period, reneging on this period's promise. Consequently, they do not invest this period. This problem is called time inconsistency.

For banking regulation, the same logic applies to fines. Once returns are realized, it may be "optimal" from the perspective of that period to assess no fines (particularly if the fine would cause bankruptcy) and to promise never to forgive fines again. However, if the regulator can forgive now, he can forgive in the future. For fines to be effective, their imposition must be credible. Large fines which cause a bank to fail, or fines during adverse macroeconomic conditions, may not be credible.

Standardized Approach

The allocation in which all banks held the same amount of capital was described earlier as the standardized approach. In the model, it performed poorly as a regulatory scheme. The allocation was included to show the potential benefits of differentiating bank capital levels by their type. In the context of the model, that allocation is the best approximation of the standardized approach. However, there is an aspect to the approach, not incorporated by the model, which may be beneficial.

Consider a modification to the paper's model where now, before a bank reports its type, the regulator gains access to the bank's portfolio and evaluates its riskiness. Now, the regulator's evaluation need not be as sophisticated as the bank's. It just needs to have some degree of accuracy. The standardized approach, as described in the introduction, could be considered one such evaluation, albeit a crude one.

This approach is different from the paper's model because by evaluating the bank's portfolio the regulator has obtained a *signal*. If this signal is at all correlated with the true riskiness of the portfolio, then it is valuable to include it in the contract. The reason for including it is that the signal, if correlated, affects the regulator's posterior distribution about a bank's type. In other words, it provides information to the regulator about the bank's type. When viewed in this context, the standardized approach may be viewed as a form of monitoring. The value of the signal, of course, would depend on both the quality of the signal and the cost of obtaining it.

It should be emphasized that the pre-commitment approach and the standardized approach are not incompatible. For example, if the signal is partially correlated with a bank's type, then it still might be valuable for the bank to choose from a menu of contracts. The difference from the contract in the paper's model would be that the menu faced by the bank would depend on the signal observed by the regulator.[13]

[13] If the signal was perfectly correlated with bank type and did not cost anything to obtain, the model would be equivalent to the full-information model discussed earlier.

6. CONCLUSION

To conclude, this paper makes several statements about the pre-commitment approach and menus of contracts. First, the approach is a proposal to use menus of contracts, a widely used contracting device. Second, in the model presented, properly designed menus are beneficial. Third, the proper design of fine schedules entails fining safe (low capital) banks but not risky (high capital) banks and basing the size of fines on likelihood ratios. Fourth, the fine schedules associated with the proposal should be viewed with caution. Fines which only occur in low probability states, as suggested by the proposal, potentially need to be large to offset the certain benefits of lowering capital. Last, the pre-commitment approach is not a market-based system. Instead, it is a regulatory scheme, just like the other proposals, but one which employs incentives.

REFERENCES

Allen, James C. "Fed to Test Self-Regulation Idea for Setting Derivatives Capital," *American Banker*, April 18, 1996.

Atkeson, Andrew G., and Robert E. Lucas, Jr. "On Efficient Distribution with Private Information," *Review of Economic Studies*, vol. 59 (July 1992), pp. 427–53.

Besanko, David, and George Kanatas. "The Regulation of Bank Capital: Do Capital Standards Promote Bank Safety?" *Journal of Financial Intermediation*, vol. 5 (April 1996), pp. 160–83.

Bliss, Robert R. "Risk-Based Bank Capital: Issues and Solutions," Federal Reserve Bank of Atlanta *Economic Review*, vol 80 (September/October 1995), pp. 32–40.

Chan, Yuk-Shee, Stuart I. Greenbaum, and Anjan V. Thakor. "Is Fairly Priced Deposit Insurance Possible?" *Journal of Finance*, vol. 47 (March 1992), pp. 227–45.

Christensen, John. "Communication in Agencies," *Bell Journal of Economics*, vol. 12 (Autumn 1981), pp. 661–74.

Dewatripont, Mathias, and Jean Tirole. *The Prudential Regulation of Banks*. Cambridge, Mass.: MIT Press, 1993.

Giammarino, Ronald M., Tracy R. Lewis, and David E. M. Sappington. "An Incentive Approach to Banking Regulation," *Journal of Finance*, vol. 48 (September 1993), pp. 1523–42.

Green, Edward J. "Lending and the Smoothing of Uninsurable Income," in Edward C. Prescott and Neil Wallace, eds., *Contractual Arrangements for Intertemporal Trade*. Minneapolis: University of Minnesota Press, 1987.

Greenspan, Alan. Speech to the 32nd Annual Conference on Bank Structure and Competition, Federal Reserve Bank of Chicago, May 2, 1996.

Kupiec, Paul H., and James O'Brien. "Recent Developments in Bank Capital Regulation of Market Risks," Working Paper 95–51. Washington: Board of Governors of the Federal Reserve System, December 1995a.

_____. "Model Alternative," *Risk*, vol. 8 (July 1995b), pp. 37–40.

_____. "A Pre-Commitment Approach to Capital Requirements for Market Risk," Working Paper 95–36. Washington: Board of Governors of the Federal Reserve System, July 1995c.

Kuprianov, Anatoli. "Tax and Regulatory Disincentives to Commercial Banking," Federal Reserve Bank of Richmond *Economic Quarterly* (forthcoming 1997).

Laffont, Jean-Jacques, and Jean Tirole. "Using Cost Observation to Regulate Firms," *Journal of Political Economy*, vol. 94 (June 1986), pp. 614–41.

Lucas, Robert E., Jr. "Methods and Problems in Business Cycle Theory," *Journal of Money, Credit, and Banking*, vol. 12 (November 1980, Part 2), pp. 696–715.

Marshall, David, and Subu Venkataraman. "Bank Capital for Market Risk: A Study in Incentive-Compatible Regulation," Federal Reserve Bank of Chicago *Chicago Fed Letter*, Number 104, April 1996.

Phelan, Christopher, and Robert M. Townsend. "Computing Multiperiod, Information-Constrained Optima," *Review of Economic Studies*, vol. 59 (October 1991), pp. 853–82.

Prescott, Edward Simpson. "Communication in Models with Private Information: Theory, Applications, and Computation." Ph.D. Dissertation. University of Chicago, May 1995.

Seiberg, Jaret. "The Fed Considers Sweeping Changes in Risk-Based Capital Requirements," *American Banker*, December 13, 1996.

Spong, Kenneth. *Banking Regulation: Its Purposes, Implementation, and Effects*. Kansas City: Federal Reserve Bank of Kansas City, 1994.

Wilson, Robert B. *Nonlinear Pricing*. New York: Oxford University Press, 1993.

Algebraic Production Functions and Their Uses Before Cobb-Douglas

Thomas M. Humphrey

Fundamental to economic analysis is the idea of a production function. It and its allied concept, the utility function, form the twin pillars of neoclassical economics. Written

$$P = f(L, C, T \ldots),$$

the production function relates total product P to the labor L, capital C, land T (terrain), and other inputs that combine to produce it. The function expresses a technological relationship. It describes the maximum output obtainable, at the existing state of technological knowledge, from given amounts of factor inputs. Put differently, a production function is simply a set of recipes or techniques for combining inputs to produce output. Only efficient techniques qualify for inclusion in the function, however, namely those yielding maximum output from any given combination of inputs.

Production functions apply at the level of the individual firm and the macro economy at large. At the micro level, economists use production functions to generate cost functions and input demand schedules for the firm. The famous profit-maximizing conditions of optimal factor hire derive from such micro-economic functions. At the level of the macro economy, analysts use aggregate production functions to explain the determination of factor income shares and to specify the relative contributions of technological progress and expansion of factor supplies to economic growth.

For valuable comments on earlier drafts of this article, the author is indebted to his Richmond Fed colleagues Bob Hetzel, Ned Prescott, Pierre-Daniel Sarte, and Alex Wolman. The views expressed herein are the author's and do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

The foregoing applications are well known. Not so well known, however, is the early history of the concept. Textbooks and survey articles largely ignore an extensive body of eighteenth and nineteenth century work on production functions. Instead, they typically start with the famous two-factor Cobb-Douglas version

$$P = bL^k C^{1-k}.$$

That version dates from 1927 when University of Chicago economist Paul Douglas, on a sabbatical at Amherst, asked mathematics professor Charles W. Cobb to suggest an equation describing the relationship among the time series on manufacturing output, labor input, and capital input that Douglas had assembled for the period 1889–1922.[1]

The resulting equation

$$P = bL^k C^{1-k}$$

exhibited constant returns to scale, assumed unchanged technology, and omitted land and raw material inputs. With its exponents k and $1 - k$ summing to one, the function seemed to embody the entire marginal productivity theory of distribution. The exponents constitute the output elasticities with respect to labor and capital. These elasticities, in competitive equilibrium where inputs are paid their marginal products, represent factor income shares that just add up to unity and so exhaust the national product as the theory contends.

The function also seemed to resolve the puzzling empirical constancy of the relative shares. How could those shares remain unchanged in the face of secular changes in the labor force and the capital stock? The function supplied an answer. Increases in the quantity of one factor drive down its marginal productivity and hence its real price. That price falls in the same proportion as the increase in quantity so that the factor's income share stays constant. The resulting share terms k and $1 - k$ are fixed and independent of the variables P, L, and C. It follows that even massive changes in those variables and their ratios would leave the shares unchanged.

From Cobb-Douglas, textbooks and surveys then proceed to the more exotic CES, or constant elasticity of substitution, function

$$P = [kL^{-m} + (1 - k)C^{-m}]^{-1/m}.$$

They observe that the CES function includes Cobb-Douglas as a special case when the elasticity, or flexibility, with which capital can be substituted for labor or vice versa approaches unity.

[1] Even before Douglas's collaboration with Cobb, his research assistant at Chicago, Sidney Wilcox, had devised in 1926 the formula $P = [L^2 + C^2]^{1/2\epsilon} L^k C^h$, where the exponents ϵ, k, and h sum to unity. Wilcox's function reduces to the Cobb-Douglas function in the special case when ϵ is zero, but not otherwise (see Samuelson 1979, p. 927).

Finally, the texts arrive at functions that allow for technological change. The simplest of these is the Tinbergen-Solow equation. It prefixes a residual term e^{rt} to the simple Cobb-Douglas function to obtain

$$P = e^{rt}L^kC^{1-k}.$$

This term captures the contribution of exogenous technological progress, occurring at trend rate r over time t, to economic growth. Should new inventions and innovations fail to materialize exogenously like manna from heaven, however, more complex functions are available to handle endogenous technical change. Of these and other post-Cobb-Douglas developments, texts and surveys have much to say. Of the history of production functions before Cobb-Douglas, however, they are largely silent.

The result is to foster the impression among the unwary that algebraic production functions are a twentieth century invention. Nothing, however, could be further from the truth. On the contrary, the idea, if not the actuality, of such functions dates back at least to 1767 when the French physiocrat A. R. J. Turgot implicitly described total product schedules possessing positive first partial derivatives, positive and then negative second partial derivatives, and positive cross-partial derivatives. Thirty years later, Parson Thomas Malthus presented his famous arithmetic and geometric ratios (1798), which imply a logarithmic production function. Likewise, a quadratic production function underlies the numerical examples that David Ricardo (1817) used to explain the trend of the relative shares as the economy approaches the classical stationary state. In roughly the same period, pioneer marginalist Johann Heinrich von Thünen hypothesized geometrical series of declining marginal products implying an exponential production function. Before he died in 1850, Thünen wrote an equation expressing output per worker as a function of capital per worker. When rearranged, his equation yields the Cobb-Douglas function.

Others besides Thünen presaged modern work. In 1877 a mathematician named Hermann Amstein derived from a production function the first-order conditions of optimal factor hire. Moreover, he employed the Lagrangian multiplier technique in his derivation. And in 1882 Alfred Marshall embedded an aggregate production function in a prototypal neoclassical growth model. From the mid-1890s to the early 1900s a host of economists including Philip Wicksteed, Léon Walras, Enrico Barone, and Knut Wicksell used production functions to demonstrate that the sum of factor payments distributed according to marginal productivity exactly exhausts the total product. One of these writers, A. W. Flux, introduced economists to Leonhard Euler's mathematical theorem on homogeneous functions. Finally, exemplifying the adage that no scientific innovation is christened for its true originator, Knut Wicksell presented the Cobb-Douglas function at least 27 years before Cobb and Douglas presented it.

The following paragraphs trace this evolution and identify specific contributions to it. Besides exhuming lost or forgotten ideas, such an exercise

may serve as a partial corrective to the tendency of textbooks and surveys to neglect the early history of the concept. One thing is certain. Algebraic production functions developed hand-in-hand with the theory of marginal productivity. That theory progressed from eighteenth century statements of the law of diminishing returns to late nineteenth and early twentieth century proofs of the product-exhaustion theorem.

Each stage saw production functions applied with increasing sophistication. First came the idea of marginal productivity schedules as derivatives of a production function. Next came numerical marginal schedules whose integrals constitute particular functional forms indispensable in determining factor prices and relative shares. Third appeared the pathbreaking initial statement of the function in symbolic form. The fourth stage saw a mathematical production function employed in an aggregate neoclassical growth model. The fifth stage witnessed the flourishing of microeconomic production functions in derivations of the marginal conditions of optimal factor hire. Sixth came the demonstration that product exhaustion under marginal productivity requires production functions to exhibit constant returns to scale at the point of competitive equilibrium. Last came the proof that functions of the type later made famous by Cobb-Douglas satisfy this very requirement. In short, macro and micro production functions and their appurtenant concepts—marginal productivity, relative shares, first-order conditions of factor hire, product exhaustion, homogeneity and the like—already were well advanced when Cobb and Douglas arrived.

1. PRODUCTION FUNCTIONS IMPLICIT IN VERBAL STATEMENTS OF THE LAW OF DIMINISHING RETURNS

The notion of an algebraic production function is implicit in the earliest verbal statements of the operation of the law of diminishing returns in agriculture. A. R. J. Turgot, the French physiocratic economist who served as Louis XVI's Minister of Finance, Trade, and Public Works for a year until dismissed for enacting free-market reforms against the wishes of the king, provided the best of these early statements. In his 1767 *Observations on a Paper by Saint-Péravy,* Turgot discusses how variations in factor proportions affect marginal productivities.[2]

Suppose, he writes, that equal increments of the variable factor capital are applied to a fixed amount of land. Each successive increment adds a positive increase to output such that capital's marginal productivity is positive. But that marginal productivity, which at first rises with increases in the capital-to-land

[2] On Turgot's discovery of the law of diminishing returns, see Lloyd (1969, p. 22), Niehans (1990, pp. 75–76), and Schumpeter (1954, pp. 259–61). On the production function implicit in Turgot's discovery, see Schumpeter (1954, p. 1036).

ratio, eventually attains a peak and then falls until it reaches zero. At that latter point, the total product of capital—the sum of the marginal products—is at a maximum.

Here is the first clear articulation of the law of variable proportions, or diminishing marginal productivity. Although Turgot applied the law strictly to capital, he realized that it holds for any variable factor including labor. He also recognized a corollary proposition, namely that increases in any factor raise the marginal productivities of the other cooperating factors, which now have more of the first factor to work with. Thus additions to capital, while eventually lowering capital's own marginal productivity, raise the marginal productivities of labor and land.

Turgot's Production Function

Marginal productivity, when expressed mathematically, is the first-order partial derivative of the production function with respect to the input in question, or

$$\partial P/\partial C.$$

And the rate of change of that marginal productivity, again with respect to the associated input, is the second-order partial derivative

$$\partial[\partial P/\partial C]/\partial C = \partial^2 P/\partial C^2.$$

Finally, the response of an input's marginal productivity to changes in complementary inputs is a cross-partial derivative

$$\partial[\partial P/\partial C]/\partial L = \partial^2 P/\partial C \partial L.$$

From what has been said above, it follows that Turgot implicitly described a production function possessing positive first partial derivatives, positive then negative second partial derivatives, and positive cross-partial derivatives. His function, with its initially rising marginal productivity of capital, differs from Cobb-Douglas. In Cobb-Douglas, of course, the marginal productivity of a variable factor declines monotonically from the outset so that the second partial derivative is always negative. Also, Turgot's function, because of the fixity of land, cannot exhibit constant returns to scale like Cobb-Douglas.

2. PRODUCTION FUNCTIONS IMPLICIT IN NUMERICAL TABLES AND SERIES

More than 30 years after Turgot, English classical economists independently rediscovered his notion of production functions obeying the law of variable proportions. Unlike him, however, they expressed the concept numerically. Thus several British classicals, though presenting no explicit mathematical production functions, nevertheless used hypothetical numerical examples and series

that imply specific functional forms. A logarithmic function underlies Thomas Malthus's famous arithmetical and geometrical series, which he used to illustrate the law of diminishing returns. In his 1798 *An Essay on the Principle of Population,* Malthus wrote that population, if unchecked, tends to increase indefinitely over time at the geometric ratio 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 . . . . Food output, on the other hand, increases at the arithmetic ratio 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 . . . .

Thomas Malthus's Logarithmic Production Function

Let L denote the labor force or its proxy, the population. Similarly, let P denote food output and t denote time, normalized so that one unit is the interval required for population to double. (Malthus estimated this doubling time to be 25 years.) Then the equation

$$L = 2^t$$

generates Malthus's geometric series for population as time t assumes successive values of 0, 1, 2, 3, etc. Similarly, his arithmetic series for food evolves from the equation

$$P = t + 1.$$

Treating the labor force L and total product P in the spirit of Malthus as interdependent, interacting variables, one can reduce the two equations to a single logarithmic expression.[3] Solve the second equation for time t, substitute the result into the first equation, take logarithms, and then solve for P to obtain the production function

$$P = f(L) = 1 + (1/\log 2)\log L = 1 + (\text{constant})\log L.$$

This production function establishes no absolute upper limit to output. But it does display continuously falling marginal and average productivities of labor. These productivities,

$$dP/dL = f'(L) = (1/\log 2)(1/L)$$

and

$$P/L = f(L)/L = (1/L) + (1/\log 2)(\log L/L),$$

respectively, approach zero asymptotically as the labor force becomes very large. Here are Malthusian diminishing returns with a vengeance.

Malthus put his diminishing-returns production function to immediate use. He employed it to rationalize his minimum-subsistence theory of wages. He

[3] George Stigler ([1952] 1965, p. 193) was the first to call attention to Malthus's production function. Peter Lloyd (1969, pp. 22–26) presents an expanded treatment.

argued that labor-force size responds to gaps between actual and subsistence wages. Its response keeps wages at subsistence. Thus above-subsistence wage rates act to raise birth rates, lower death rates, and spur labor-force growth. Because of diminishing returns, however, the extra workers reduce labor's marginal productivity and hence the real wage rate to subsistence. Conversely, below-subsistence wages lead to starvation, low birth rates, and labor-force decline. Fewer workers mean higher marginal productivity of labor, thereby restoring wages to subsistence.

Other classical economists seized on Malthus's population mechanism. Thus was born the classical notion of an unlimited, or infinitely elastic, long-run supply of labor at the subsistence wage rate.

David Ricardo

Malthus was hardly the only classical economist to work with production functions exhibiting diminishing returns. David Ricardo was the most prominent of the many others who did so. His famous theory of growth and distribution in the economy's progress toward the stationary state rests on a quadratic production function yielding linearly declining marginal and average product schedules. Thus, in his 1817 *Principles of Political Economy and Taxation,* he combines his particular production function with Malthus's minimum-subsistence wage theory to predict that scarcity of land ultimately will bring growth to a halt.

According to Ricardo, growth ceases when diminishing returns to capital applied to scarce land lower capital's real reward to a minimum consistent with zero net investment. At this point, the incentive to invest as well as the means to finance investment vanish and the economy approaches the classical stationary state.

In constructing his production function $P = f(L)$, Ricardo assumed that labor and capital combine in rigidly fixed proportions. Each worker, for example, comes equipped with a shovel. The resulting composite input labor-and-capital L then combines with uniformly fertile land in variable proportions to generate diminishing returns. Ricardo believed that diminishing returns in agriculture were powerful. Indeed, he thought they were so powerful as to overwhelm increasing returns in manufacturing stemming from technological progress and the division and specialization of labor. For that reason, he concentrated on the agricultural sector and omitted variables representing technological progress from his production function. Fixed land, too, was omitted on the grounds that it was a constant rather than a variable. Finally, Ricardo drew no distinction between the aggregate production function for the whole economy and the corresponding micro function for the representative farm. He simply viewed the aggregate function as a scaled-up version of the micro function and treated the economy as one giant farm.

Ricardo's Quadratic Production Function

Like Malthus, Ricardo presents his function in the form of a numerical example rather than an algebraic equation.[4] His *Principles* displays a table showing hypothetical marginal products of successive homogeneous doses of labor-and-capital L applied to land of uniform fertility. The first dose produces 180 units of output. Each succeeding dose contributes 10 fewer units than its immediate predecessor—the second dose contributing 170 units, the third 160 units, and so on. These numbers imply the linearly declining marginal productivity schedule

$$dP/dL = f'(L) = 190 - 10L,$$

which, upon integration, yields the quadratic production function

$$P = f(L) = 190L - 5L^2.$$

One property is absolutely crucial to Ricardo's theory of the trend of relative shares as the economy approaches the stationary state. The function's associated average product schedule

$$P/L = f(L)/L = 190 - 5L$$

declines at half the rate of the marginal product schedule so that the ratio of marginal to average product falls as L increases.

Ratio of Marginal to Average Products and the Trend of Relative Shares

This property, together with Ricardo's assumption that Malthusian population growth keeps the wage rate at subsistence, determines the trend of relative shares in his model. For it is easy to show that the shares going to rent on the one hand and wages plus profit on the other vary inversely and directly, respectively, with the ratio of the marginal to the average product of labor-and-capital. After all, land's absolute real rental income R is simply what remains of total product $P = f(L)$ after the variable composite factor L receives its marginal product $f'(L)$. That is,

$$R = f(L) - Lf'(L).$$

Dividing through by total product gives rent's relative share

$$R/P = [f(L) - Lf'(L)]/f(L) = 1 - \{f'(L)/[f(L)/L]\}$$

as one minus the term in braces. This latter term represents the combined share of total product going to labor and capital together. It is nothing other than the crucial ratio of the marginal to the average product of the composite variable

[4] Haim Barkai deduced this function from Ricardo's tables in 1959. Both he and Blaug (1985, pp. 88–92, 103–05, 118–21) discuss how Ricardo used it to predict the trend of rent's distributive share as the economy approaches the stationary state.

L. Since the ratio falls with increasing applications of *L*, it follows that rent's share rises while the combined share of wages and profit falls.

Of the combined share, the wage component must rise and the profit component fall. The reason is simple. Since the Malthusian mechanism holds the wage rate at subsistence, the total wage bill consisting of the wage rate times the work force grows proportionally with the number of workers. Output, however, grows less than proportionally to labor because of diminishing returns. As a result, the ratio of the wage bill to total product, namely labor's share, increases with *L*. And with the relative shares of rent and wages both rising, it follows that the remaining relative share of profit necessarily falls.

Approach to the Classical Stationary State

Eventually, profit and its relative share fall to a minimum, perhaps zero. There both the means and the incentive to finance net investment vanish. At that point, profit is just sufficient to maintain rather than to increase the capital stock.

With capital formation stymied, growth halts. Here is Ricardo's prediction that diminishing returns stemming from scarcity of land overwhelm increasing returns due to technological progress and so lead inevitably to the classical stationary state. His pessimistic, dismal prophecy derives directly from a production function exhibiting linearly declining marginal and average returns (see Figure 1).

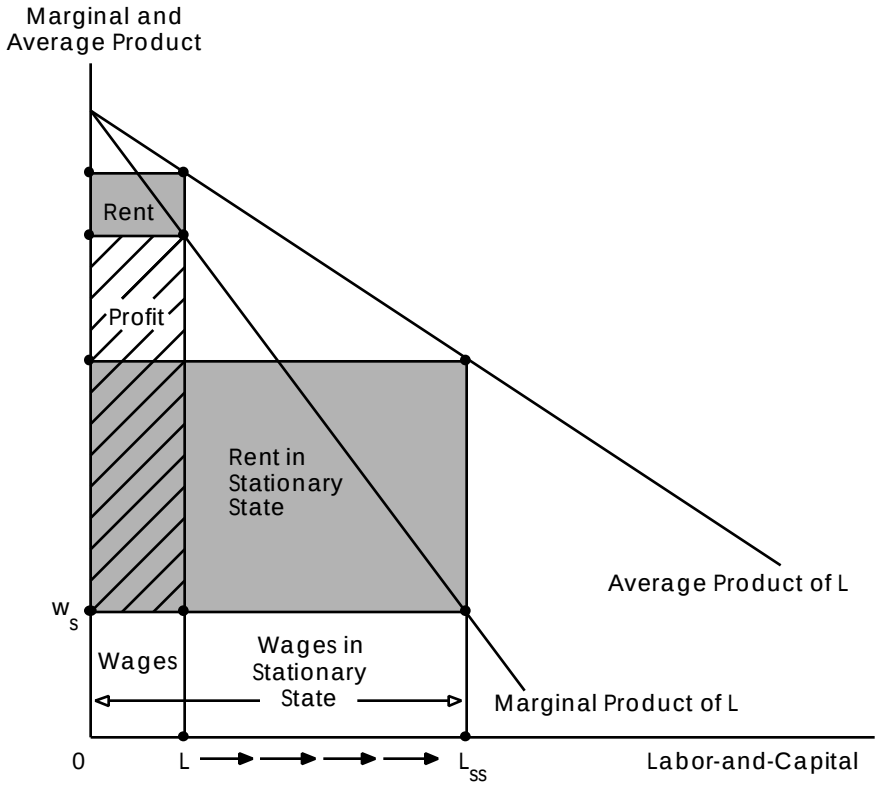
Ricardian Rent Theory as Incomplete Marginalism

As a marginal productivity theory of factor pricing and distribution, Ricardo's rent analysis left much to be desired. True, it did establish the marginal principle. It stressed that capitalist farmers should cultivate land to the point where the incremental return to the last dose of labor-and-capital applied just equals the cost of the dose. But it employed marginal analysis only to determine the joint payment going to labor and capital combined.

To account for the rewards going to each factor separately, Ricardo had to resort to other explanations. He relied on the Malthusian minimum-subsistence theory to determine labor's wage rate and income share. Similarly, he explained capital's profit rate as a pure residual, namely what remained of the marginal product of labor-and-capital after deduction of subsistence wages. Likewise, he viewed land's rental rate as a surplus determined by the gap between the average and marginal products of the variable factor, or alternatively, by the superior productivity of the intramarginal doses of the factor.

In short, Ricardo resorted to subsistence and residual theories to determine factor prices. Marginal productivity served only to split the total product into its rent and non-rent components. What was needed was someone to transform the primitive, incomplete marginalism of classical Ricardian theory into comprehensive neoclassical marginal productivity.

Figure 1 Ricardo's Theory of the Trend of the Relative Shares



Ricardo's quadratic production function generates linearly declining marginal and average product schedules. Total product consists of the rectangular area inscribed under the average product curve at any given level of labor-and-capital. The composite labor-and-capital input receives its marginal product, of which labor gets subsistence wages $w_S L$ and capital gets the cross-hatched area as profit. Land rent accounts for the remaining rectangle bounded by the gap between the marginal and average product curves. Such rent consists of the surplus product of the inframarginal doses of the composite factor. Bidding for scarce land transfers the surplus to landowners.

Initially, with labor-and-capital at L, profit exists to fuel capital formation and growth. Growth stops when the composite factor expands to its stationary-state level L_{SS}. There diminishing returns reduce profit to zero, thus extinguishing the source of growth. In the classical stationary state, all product accrues to labor and land. The widening gap between the marginal and average product curves ensures that the profit share falls while the rent and wage shares rise as the economy expands along the horizontal axis.

Johann Heinrich von Thünen's Contributions to Marginal Productivity Theory

Credit for doing so goes to the German mathematical economist, location theorist, and agronomist Johann Heinrich von Thünen (1783–1850), whose work on production functions was far ahead of its time. A true original, he owed nothing to the productivity doctrines of Malthus and Ricardo. Having read neither writer, Thünen constructed his theories fresh from the detailed records that he kept for his own agricultural estate. In a lifelong effort to identify empirically the exact relations of production on his farm, he applied marginal analysis to all factor inputs and prices. The entire neoclassical theory of production and distribution traces its origin to Volume II of his great book *The Isolated State*.

As the earliest neoclassical marginalist, Thünen boasts several distinctions. He was the first to apply the differential calculus to productivity theory and perhaps the first to use it to solve economic optimization problems.[5] He was likewise the first to interpret marginal productivities essentially as partial derivatives of the production function. In so doing, he made explicit what was merely implicit in Turgot's analysis.

Decomposing Ricardo's composite input into its separate components, Thünen was the first to treat labor and capital symmetrically, to show that each is subject to diminishing incremental returns, and to state that labor's marginal productivity is an increasing function of the quantity of capital per worker. Moreover, he was the first to state precisely that capital's real reward and labor's too are determined by the additional product resulting from the last increment of each factor hired, all other factors held constant. Likewise, he was the first to show that when capital's interest rate is determined marginally, wages may appear to be a residual. Conversely, when the wage rate is determined marginally, interest appears as a residual. Unlike Ricardo, who assumed fixed factor proportions, Thünen stressed that labor and capital are substitutes for each other and can be combined in variable proportions.

Always Thünen insisted that economic efficiency requires that factors be hired until the ratio of their respective marginal products equals the ratio of their unit prices. Always he held that net revenue peaks when each input's marginal value product matches its marginal factor cost. In short, he relentlessly applied the principle that equimarginal resource allocation maximizes the total product.

[5] The vexed question of priority of discovery raises its head here. Thünen evidently used the differential calculus to solve an optimization problem as early as 1824. But in that same year Thomas Perronet Thompson employed rudimentary elements of the calculus to compute the optimal inflation tax. Even earlier, in a book published in 1815, Georg von Buquoy used the calculus to determine the optimal plowing depth of the soil. In any case, Thünen's optimization calculus remained unpublished until 1850, by which time Augustin Cournot's 1838 *Researches into the Mathematical Principles of the Theory of Wealth* had eclipsed it. On these contributions see Niehans (1990, pp. 173–74) and the sources cited there.

These contributions identify him as the true founder of neoclassical marginal productivity theory.[6]

Thünen's Exponential Production Functions

Bolstering his theory with numbers derived from agricultural experiments on Tellow, his estate in Mecklenburg, Thünen presented tables depicting the marginal productivities of labor L, capital C, and fertilizer F applied to fixed land. His tables show the marginal productivities declining at constant geometrical rates. In other words, his experiments suggested to him that successive unit increases of any variable input, the others held constant, add to output a constant fraction of the amount added by the preceding unit. For labor the fixed fraction was two-thirds, for capital nine-tenths, and for fertilizer one-half.

Let r denote the fractional ratio between successive marginal products of any variable factor. Then Thünen's schedule of the factor's incremental returns constitutes the terms of the decreasing geometric series $a, ar, ar^2, ar^3, \ldots ar^{n-1}$. Here a is the marginal product of the first unit of the factor, ar the marginal product of the second, ar^2 that of the third and so on until we reach the last, or nth, unit whose marginal product is ar^{n-1}.

Using the sum-of-the-series formula

$$S = [a(1 - r^n)/(1 - r)]$$

to sum the n marginal products gives the factor's total product schedule as

$$P = A(1 - r^n),$$

where the letter A denotes the constant term $a/(1 - r)$ and the exponent n is the number of units of the variable factor hired. Since r is a fraction such that r^n becomes zero when n becomes infinite, it follows that A is the limit that the sum $A(1 - r^n)$ approaches as the number of factor units n becomes indefinitely large. The upshot is that the factor's total product asymptotically converges to the finite maximum A.

Exactly the same analysis holds for each and every variable input. Consequently, when all factors—labor, capital, and fertilizer—are allowed to vary simultaneously, the production function underlying Thünen's numerical examples can be expressed as

$$P = f(L, C, F) = A(1 - 2/3^L)(1 - 9/10^C)(1 - 1/2^F),$$

where the exponents denote the amounts of each factor employed.

[6] On Thünen and his contributions to marginal productivity theory, see Blaug (1985, pp. 322–24, 449, 616–17), Dickinson (1969), Leigh (1968), Niehans (1990, pp. 164–75), and Schumpeter (1954, pp. 465–68).

The foregoing applies when inputs vary in units of discrete size. Should those units be infinitely divisible and so continuously variable, then the term e^{-k} replaces the fraction r in the production function. Here k denotes the instantaneous rate of decline of marginal productivity and e^{-k} is the factor of proportionality over the unit interval. The result is that each input's total product schedule becomes

$$P = A(1 - e^{-kn}),$$

and the corresponding Thünen production function is

$$P = f(L, C, F) = A(1 - e^{-.405L})(1 - e^{-.105C})(1 - e^{-.693F}).$$

This function, like its discrete counterpart, possesses two properties. Output is zero when any factor is zero. Output approaches its maximum level A as all factors are increased indefinitely.[7]

Rediscovery of Thünen's Exponential Functions

Thünen's exponential production functions together with their associated marginal schedules passed largely unnoticed for more than 40 years after their publication in 1863. They (but not their authorship) were rediscovered by two agricultural scientists, W. J. Spillman and E. A. Mitscherlich, working independently in the early twentieth century. Spillman labeled Thünen's decreasing geometric series the "law of the diminishing increment" and wrote out the corresponding marginal and total product equations for fertilizer and irrigation water. Mitscherlich christened the same phenomenon the "law of the soil." This law he represented by the equation

$$dP/dF = k(A - P),$$

expressing fertilizer's marginal productivity dP/dF as a constant fraction of the (diminishing) gap between maximum A and actual current levels P of output. Upon integration, his equation yields Thünen's total product schedule

$$P = f(F) = A(1 - e^{-kF})$$

for the continuous case. Neither author, however, was aware of Thünen's earlier formulation of these concepts.

3. THE FIRST ALGEBRAIC PRODUCTION FUNCTION

In addition to the functions implicit in his numerical examples, Thünen wrote down the first explicit algebraic production function to appear in print (see Lloyd [1969], pp. 31–33). As presented in Chapter 2 of the second part of

[7] Lloyd (1969, pp. 26–31) provides a complete account of Thünen's exponential production functions and their derivation. See also Stigler (1946, p. 125) for a textbook treatment.

Volume II of *The Isolated State*, the function evidently replaces the geometric series of marginal products he presented in the first part of that same volume. Expressed in per-worker form, his function is

$$p = hq^n,$$

where p is output per worker, h is a constant parameter determined by such considerations as the fertility of the soil and the strength and diligence of the workers who till it, q is capital per worker (the capital-to-labor ratio), and the exponent n is a fraction between zero and one.

It turns out that Thünen's production function is none other than the Cobb-Douglas function $P = bL^kC^{1-k}$ in disguise. For, when one multiplies both sides of Thünen's equation by labor L, one obtains

$$P = pL = hLq^n = hL(C/L)^n = hL^{1-n}C^n.$$

The resulting function

$$P = hL^{1-n}C^n$$

is virtually the same as the Cobb-Douglas function. The conclusion is inescapable. Credit for presenting the first Cobb-Douglas function, albeit in disguised or indirect form, must go to Thünen in the late 1840s rather than to Douglas and Cobb in 1928. All of which goes to show that there is nothing new under the sun. Or as statistician Stephen M. Stigler expressed it in his famous Law of Eponymy, "No scientific discovery is named for its original discoverer."

Thünen's equation states that production requires inputs of both labor and capital such that labor working alone produces nothing. Thünen was uncomfortable with this result. Surely labor unaided by capital has some productivity, however low. To ensure that labor's output is positive even if capital is zero, Thünen modified his production function to read

$$p = h(1 + q)^n,$$

or, when multiplied through by labor L,

$$P = hL^{1-n}(L + C)^n.$$

This equation, which Thünen estimated empirically for his own agricultural estate and which he declares he discovered only after more than 20 years of fruitless search, states that labor produces something even when unequipped with capital.[8]

[8] The variable q in this equation expresses the capital-labor ratio when capital is measured in units of work effort. Alternatively, Thünen uses the letter k to denote the ratio when capital is expressed in the workers' means of subsistence. In this latter case, his formula becomes

$$p = h(g + k)^n,$$

where g is a positive constant not necessarily equal to unity.

4. FIRST USE OF AN AGGREGATE PRODUCTION FUNCTION IN A NEOCLASSICAL GROWTH MODEL

For all its brilliance and originality, Thünen's productivity analysis had little impact on his contemporaries and immediate successors. Some were intimidated by its formidable mathematics and shunned it for that reason. Thünen's own countrymen largely ignored it because it was theoretical and thus ran counter to the anti-theoretical bias of the dominant German Historical School. Still others overlooked it because it was hidden amidst the profusion of cryptically written notes, comments, digressions, repetitions, numerical examples, and mathematical formulas that constituted the disjointed and cumbersome narrative of *The Isolated State*. Another reason for neglect was that Thünen's readers concentrated on his celebrated but misguided formula $w = \sqrt{ap}$ to the exclusion of his other work. That formula, which Thünen thought sufficiently important to have engraved on his tombstone, specified the natural wage w as the geometric mean of the worker's minimum subsistence a and his average product p. Preoccupied with the formula, readers tended to overlook Thünen's genuine contributions to production theory.

One economist who was influenced, however, was Alfred Marshall. He acknowledges his debt to Thünen in a fragment preserved in the 1925 *Memorials of Alfred Marshall*. There he credits Augustin Cournot with teaching him pure analytic technique and Thünen with teaching him economics. Confessing that he derived more of his ideas from Thünen than from Cournot, he declares that he reveres Thünen above all his masters.

Marshall's Growth Model

Given his indebtedness to Thünen's productivity ideas, it is hardly surprising to find Marshall, in a note written in 1881 or 1882, employing an aggregate production function. His function appears in what is best described as a prototypical neoclassical growth model. In that model, Marshall expresses aggregate annual output or real national product P as a function of four determinants. These are, respectively, the number L times the average efficiency E of the working population (the work force measured in efficiency units), the accumulated stock of capital C, the level of technology or state of the arts of production A, and the fertility of the soil F, which Marshall treats as a fixed constant. In symbols,

$$P = f(L \cdot E, C, A, F).$$

Taken together, the growth rates of the arguments of Marshall's function determine the growth rate of aggregate output. Marshall expresses these input growth rates as time derivatives—dL/dt, dE/dt, dC/dt, dA/dt—each treated as a function of several relevant variables including wage and interest rates, the standard of comfort, time, and the arguments in the production function itself. In principle, the resulting dynamic system could be solved to yield secular growth paths for population, capital, technology, and output.

Marshall of course did not solve the system or investigate the qualitative properties of its growth paths. His formulation was too sketchy for such exercises. Nevertheless, he did incorporate an aggregate production function in what may be regarded as the first neoclassical growth model. And he did so at least 60 years before Tinbergen and Solow, generally regarded as the fathers of the neoclassical growth model. Unfortunately, however, Marshall never published his model and thus denied himself the credit he deserved. It remained for J. K. Whitaker to discover the model among the unpublished manuscript notes deposited in the Marshall Library at Cambridge University and to publish it in 1975.

5. PRODUCTION FUNCTIONS USED TO DERIVE THE CONDITIONS OF OPTIMAL FACTOR HIRE

Ironically, Marshall penned his model at the very time when the focus of economic theory had shifted from aggregate growth to individual optimization and allocation. Thus it was not macro but rather micro production functions that began to appear with increasing frequency in the economics literature of the 1890s. Paving the way was the so-called marginal revolution of the early 1870s. That event saw the marginalist triumvirate of William Stanley Jevons, Carl Menger, and Léon Walras apply what in essence was the calculus of constrained optimization to the consumer's utility function. The result was the derivation of the marginal utility theory of consumer's demand. Second-generation marginalists soon realized that those same optimization techniques might be applied to the production function of the individual firm to find the profit-maximizing or cost-minimizing conditions of factor hire. Thünen's marginal productivity theory was born again.

Hermann Amstein

Even before the 1890s, however, a University of Lausanne mathematician named Hermann Amstein had worked out virtually the entire theory of marginal productivity in modern algebraic dress. He did so in response to a request from his colleague Léon Walras, who sought Amstein's help in formulating mathematically the least-cost conditions of factor hire. In a letter of January 6, 1877, Amstein responded with a near-perfect analysis, complete with partial derivatives and Lagrangian multipliers, of cost minimization subject to a production function constraint.

His analysis went as follows. Define unit cost of production U as the ratio of total cost to output. Total cost consists of the sum of the factor inputs each multiplied by its unit price. Let the wage rate w and the interest rate i denote the unit prices of labor and the services of capital, respectively. Competitive firms of course take these and all factor prices as given. Then Amstein argued that the problem is to find the input quantities $L, C, \ldots$ that minimize unit cost

$$U = (wL + iC + \ \ldots)/P$$

for any given level of production $\bar{P}$.

Today economists solve this problem in three steps. First they take the cost function U. Then they subtract from it the production function less the given level of output $f(L, C, \ \ldots) - \bar{P}$, all multiplied by an arbitrary Lagrangian multiplier λ. Finally, they minimize the resulting Lagrangian expression

$$Z = U - \lambda[f(L, C, \ \ldots) - \bar{P}]$$

by setting its first partial derivatives equal to zero.

That is precisely what Amstein did, with one exception. He suppressed the given product term $\bar{P}$ while setting the production function f at naught. Then he minimized the resulting Lagrangian

$$Z = (wL + iC + \ \ldots) - \lambda f(L, C, \ \ldots)$$

by setting its first partial derivatives equal to zero. This operation yielded him the first order conditions

$$w - \lambda(\partial f/\partial L) = 0 \text{ and } i - \lambda(\partial f/\partial C) = 0, \ \ldots .$$

These conditions, when rewritten as

$$w = \lambda(\partial f/\partial L) \text{ and } i = \lambda(\partial f/\partial C), \ \ldots$$

and divided by each other so as to cancel out the λs,

$$w/i = (\partial f/\partial L)/(\partial f/\partial C), \ \ldots$$

identify the least-cost combination of factor inputs as that which equates the ratio of factor prices with the ratio of factor marginal productivities or, alternatively, which renders the marginal product per last dollar spent on each factor

$$(\partial f/\partial L)/w = (\partial f/\partial C)/i = \ \ldots$$

the same across factors.

Amstein's work is a milestone in the history of production functions. Here was the first use of the Lagrangian multiplier technique in economics.[9] Here also was the first rigorous derivation of the least-cost conditions of factor hire from a constrained cost function.

[9] It was not the first to appear in print, however. John Creedy (1980) reports that Francis Edgeworth, in his 1877 *New and Old Methods of Ethics*, employed the Lagrangian multiplier technique to find the distribution of income that maximizes aggregate community satisfaction or utility. And, in his 1881 *Mathematical Psychics*, Edgeworth again used the technique to derive the contract-curve solution of Pareto-efficient allocations according to which each trader maximizes his utility subject to the condition that the other trader's utility remains constant.

These innovations, however, went completely unnoticed. For Walras, who at the time had hardly progressed beyond elementary analytical geometry and was just beginning to teach himself the rudiments of calculus, knew too little mathematics to understand Amstein's formulation and to take advantage of it. And Amstein himself knew too little economics to appreciate the significance of his demonstration and to prepare it for publication. For these reasons his contribution remained unknown until William Jaffé discovered it in the Lausanne archives and published it in 1964. The result was to delay the progress of production theory for at least 12 years. Not until 1889 was a production function employed again in an optimization problem. And not until the 1920s were Lagrange multipliers seen again in production-function analysis. Technique here ran ahead of its potential users until they became convinced of the gain from mastering it.

Francis Y. Edgeworth and Profit Maximization

Amstein derived the conditions of optimal factor hire from the competitive firm's constrained cost function. He solved a cost-minimization problem in which the production function entered as a constraint. By contrast, the next group of writers derived the factor-hire conditions from the firm's profit function. They solved a profit-maximization problem in which the production function entered as a component of gross revenue. Profit, or net revenue, they defined as the difference between gross revenue and total cost. Gross revenue consisted of product price multiplied by output as represented by the production function. Total cost consisted of the sum of factor inputs each multiplied by its factor price.

Thus Francis Edgeworth, in his 1889 *Journal of the Royal Statistical Society* article "On the Application of Mathematics to Political Economy," stated that the entrepreneur acts to maximize the profit or net revenue expression

$$f(C, T) - iC - rT.$$

Here f is gross revenue, or output valued at its given competitive market price (implicitly assigned a value of unity by Edgeworth), C is capital, T is land, i is the interest rate or price of the services of capital, and r is the rent-per-acre price of land. Maximizing this expression by setting its first partial derivatives equal to zero, Edgeworth obtained the conditions

$$\partial f/\partial C = i \text{ and } \partial f/\partial T = r.$$

In short, profit maximization requires hiring factors up to the point where they just pay for themselves, namely where their marginal value products equal their prices.

Arthur Berry and William Ernest Johnson

Cambridge lecturers Arthur Berry, a mathematician, and William Ernest Johnson, a logician, philosopher, and economic theorist, then extended Edgeworth's analysis in four ways. First, they increased the number of inputs in the production function. Thus Berry's function, which appears in his 1891 paper "The Pure Theory of Distribution," contains separate symbols for capital as well as for labor and land, both subdivided into unlimited kinds and qualities. Likewise, Johnson's production function, as presented in his 1891 piece entitled "Exchange and Distribution," embodies a potentially unlimited number of variable factor inputs V_i.

Second, Berry and Johnson incorporated product price into the entrepreneur's profit expression, thus making explicit what Edgeworth had left implicit. Johnson's model is typical of Berry's as well. Let π stand for product price, $P(\ )$ for product quantity (the production function), w for input price, V for input quantity, and the subscript $i = 1,\ 2,\ 3\ \ldots$ for the separate inputs. Then the entrepreneur seeks to maximize the profit expression

$$\pi P(V_i) - \Sigma w_i V_i \qquad (i = 1, 2, 3, \ \ldots),$$

where the first term is gross revenue and the second is total cost. Maximization yields the first-order conditions

$$\pi(\partial P/\partial V_1) = w_1, \quad \pi(\partial P/\partial V_2) = w_2, \quad \pi(\partial P/\partial V_3) = w_3, \text{ etc.}$$

Together, these state that each factor should be hired to the point where its marginal value product equals its price.

Third, suppose the firm operates under imperfect, rather than perfect, competition. Facing a downward-sloping demand curve, the firm finds the selling price of its product now varies inversely with output. Correspondingly, its marginal revenue now always lies below its price. Berry and Johnson noted that this special case necessitates replacing product price with marginal revenue in the first-order factor-hire conditions. Those conditions then read: hire factors up to the point where their marginal revenue products, or marginal physical products multiplied by marginal revenue, equal their factor prices.

Fourth, Berry and Johnson indicated how the factor-hire conditions might be incorporated into simple general equilibrium systems complete with commodity demand functions, labor supply functions, and full-employment conditions. These models no doubt influenced Edgeworth. For, in his 1894 review of Friedrich von Wieser's book *Natural Value*, he incorporated dual production functions into a two-good, two-factor model of general equilibrium. Doing so, he showed that the equality of marginal value product per last dollar spent on each factor must be the same across all goods as well as factors.

Taken together, these contributions constitute what Joseph A. Schumpeter (1954, p. 1032n) termed "a considerable achievement." They show that the

production function already was becoming an essential component of micro models of the business firm by the first half of the 1890s.

6. PRODUCTION FUNCTIONS AND THE ADDING-UP PROBLEM

Production functions continued to prove their worth in the latter half of the 1890s when marginalists employed them to resolve the famous adding-up problem of product exhaustion.[10] At stake was nothing less than the logical consistency of the marginal productivity theory of distribution. Would wages, rent, and interest, if each input is paid its marginal product, just add up to and so exhaust the total product as the theory claimed? That is, would the total product exactly be disposed of without residue or shortage?

A positive answer would confirm the consistency of the theory. But a negative answer would refute it. For if the sum of the payments according to marginal productivity exceeded the total product, the excess would go unrealized since no firm could afford to pay out more than is produced. Some inputs would be forced to accept less than their marginal products, contrary to the theory. Conversely, if less was paid out under marginal productivity than was produced, the remaining surplus would have to be distributed on grounds other than marginal productivity, contrary to the theory. Small wonder that marginalists were eager to prove the answer was yes.

Product Exhaustion under Constant Returns to Scale: Philip H. Wicksteed

First to do so expressly was Philip H. Wicksteed.[11] In his 1894 *An Essay on the Co-ordination of the Laws of Distribution,* Wicksteed proved that product exhaustion holds, provided perfect competition prevails and production functions are linear homogeneous and so exhibit constant returns to scale. Competition ensures that inputs receive their marginal products. And linear homogeneity ensures that the resulting distributive shares sum to the total product.

Had he realized it, Wicksteed could have deduced adding-up directly from Leonhard Euler's famous mathematical theorem on homogeneous functions. That theorem says that any linear homogeneous function can be written as the sum of its first partial derivatives each multiplied by the associated independent

[10] George Stigler (1941, Chapter XII) is the standard source on the history of the product exhaustion problem. See also Steedman's (1987) useful treatment.

[11] Already Knut Wicksell, in his 1893 *Value, Capital, and Rent,* had constructed a marginal productivity model that implied a proof of product exhaustion (see Stigler [1941], pp. 289–95). But he failed to make the proof explicit and was content to see Wicksteed receive credit for its discovery in the following year.

variable. In other words, the production function $P = f(L, C, \ldots)$, if linear homogeneous as Wicksteed thought, can be written as

$$P = (\partial f/\partial L)L + (\partial f/\partial C)C + \ldots,$$

where the terms on the right-hand side are factor incomes determined by marginal productivity. Here at once is the proof, ready-made, that Wicksteed sought. Curiously enough, however, he never used it. Owing to his lack of formal mathematical training, he apparently was unaware of the theorem and so made no mention of it. Instead, he sought to prove adding-up, or product exhaustion, by reconciling marginal productivity with Ricardo's theory of intensive rent.

Wicksteed's Proof of Product Exhaustion

Such reconciliation was necessary. For in Ricardo's theory there is no adding-up problem to solve. Rent, as we have seen, is a pure residual in the Ricardian model. It is what is left of the total product after the other (composite) factor, labor-and-capital, has received its marginal product. And with rent determined residually, it is tautologically true that the sum of factor incomes must just add up to the total product. The residual would always adjust to make it so. To transform Ricardo's theory into one in which the adding-up theorem applied, Wicksteed had to prove that Ricardian residual rent was the same as rent as marginal product. This proof would then imply that remuneration of all factors according to their marginal productivities exhausts the total product.

His demonstration required four steps. First, he did what Ricardo had failed to do. He entered land explicitly into the production function by writing the function in per-acre form. That is, he expressed product per acre of land P/T as an increasing function of labor-and-capital per acre L/T. In symbols, he posited

$$P/T = f(L/T)$$

or

$$P = Tf(L/T).$$

Second, he expressed Ricardian rent R as the residual part of total product remaining after each unit of labor-and-capital L receives its marginal product payment $\partial P/\partial L$. That is,

$$\begin{aligned} R &= P - (\partial P/\partial L)L \\ &= Tf(L/T) - T[\partial f(L/T)/\partial L]L \\ &= Tf(L/T) - T\{f'(L/T)[\partial(L/T)/\partial L]\}L \\ &= Tf(L/T) - T[f'(L/T)(1/T)]L \\ &= Tf(L/T) - f'(L/T)L. \end{aligned}$$

Here is rent income expressed as Ricardian residual.

Third, he expressed land's income alternatively as marginal product. That is, he computed the partial derivative

$$\partial P/\partial T = \partial[Tf(L/T)]/\partial T$$

to represent the marginal product of the last acre in use and multiplied it by the number of acres cultivated T. The result was the expression

$$[f(L/T) + Tf'(L/T)(-L/T^2)]T,$$

which reduces to

$$Tf(L/T) - Lf'(L/T),$$

precisely the same expression as residual rent. Here is his proof that Ricardian residual rent equals rent as marginal product.

Fourth, to this computed marginal productivity payment to land he adds the corresponding marginal productivity payment to labor-and-capital. He gets

$$Lf'(L/T) + Tf(L/T) - Lf'(L/T),$$

which equals

$$Tf(L/T)$$

or total product P. Here is his proof that product exhaustion occurs when both factors are paid their marginal products.

A. W. Flux and Euler's Theorem

After Wicksteed came A.W. Flux. His innovation was to accomplish what Wicksteed had failed to do, namely to deduce the adding-up proposition directly from Euler's theorem. His review of Wicksteed's *Co-ordination,* published in the June 1894 issue of the *Economic Journal,* is absolutely clear on this point.

Let the production function be linear homogeneous such that a scalar increase in all inputs yields the same scalar increase in output. Then, wrote Flux, "Euler's equation gives us at once the result" that output equals the sum of the inputs each multiplied by its partial derivative, or marginal productivity. Put differently, Euler's theorem gives us the result that factor income shares determined by marginal productivity must sum to unity and so absorb the product. Here is the first application of Euler's theorem to production function analysis. Contrary to common belief, it was Flux and not Wicksteed who introduced this theorem to economists.

The Critics: Barone, Edgeworth, Pareto, and Walras

The Wicksteed-Flux proof of product exhaustion received an inhospitable reception. Critics including Enrico Barone, Francis Edgeworth, Vilfredo Pareto, and Léon Walras attacked its homogeneity assumption. They argued that linear

homogeneity renders adding-up a trivial outcome that holds at every point on the production function regardless of the proportions in which the factors are combined. In other words, homogeneity proves too much and is thus too good to be true. Edgeworth's ([1904], 1925, p. 31) caustic remark was devastating: "There is a magnificence in this generalization which recalls the youth of philosophy. Justice is a perfect cube, said the ancient sage; and rational conduct is a homogeneous function, adds the modern savant."

The critics further noted that the homogeneity proposition yields horizontal long-run marginal and unit cost curves. Such curves render firm size indeterminate. They thus cast doubt on the large-numbers property of competition. Why? Because competitive firms possessing horizontal cost curves can minimize cost at any scale of operation. With no cost advantage to being small or disadvantage to being large, a firm could be of any size. But such firm-size indeterminacy in turn implies indeterminacy of the number of firms in the industry. Conceivably, a few firms might be so large as to monopolize the market, contrary to the assumptions of the competitive model.

Vilfredo Pareto (1897) adduced three additional reasons working against linear homogeneity. First, some factors are in fixed supply. They cannot expand equiproportionally with the others as homogeneity implies. One can, for example, replicate all the elements of a restaurant on the Champs Elysées except the location itself. Second, some inputs come in units that are large and indivisible. Such lumpy inputs can hardly be scaled up or down in proportion to output as homogeneity assumes. An example is a train tunnel that can accommodate a quadrupling of the traffic but that cannot be subdivided into smaller tunnels of the same efficiency to handle a fraction of the traffic. Third, some factors are in a fixed technological relation with the product (iron and iron ore) or with each other (trucks and truck drivers). Their lack of independent variation thwarts the freedom of factor substitution that homogeneity assumes. To Pareto, these reasons were enough to render production functions nonhomogeneous so that they exhibit increasing or decreasing returns to scale.

Knut Wicksell's Reconciliation of Nonhomogeneity with Adding-Up

Knut Wicksell clarified, refined, and considerably amplified the foregoing observations. In so doing, he reconciled product exhaustion with nonhomogeneity. Unlike Wicksteed, who saw constant- and nonconstant-returns production functions as mutually exclusive phenomena, Wicksell (1901, 1902) argued that a firm's production function might exhibit successive stages of increasing, constant, and decreasing returns to scale. These stages correspond to the falling, constant, and rising segments of the firm's U-shaped long-run average cost curve. Free entry of rivals in pursuit of profit forces the competitive firm to operate at the minimum point of this curve. Or what is the same thing, competition forces the firm to operate at the zero-profit point, where its production function

is tangent to a linear homogeneous plane. Here, constant returns and therefore adding-up prevail. In short, product exhaustion is an equilibrium condition that holds at the single point where the firm's nonhomogeneous production function behaves as if it were linear homogeneous.

Wicksell noted, however, that nonhomogeneous production functions for firms are perfectly compatible with a linear homogeneous function for the entire industry. Suppose industry output expands and contracts through the entry and exit of identical firms, each operating at the same minimum unit cost. The result is to trace out a horizontal long-run industry supply curve that looks like it came from a constant-returns production function.

Here was Wicksell's contribution. At one stroke, he solved three problems. He reconciled adding-up with nonhomogeneity of production functions at the level of the individual firm. He then reconciled those functions with homogeneous functions for the industry. In so doing, he justified the use of aggregate linear homogeneous functions such as the Cobb-Douglas function. Finally, he reconciled competitive equilibrium with determinate firm size.

Product Exhaustion under Nonconstant Returns

The preceding considerations led Barone, Walras, and Wicksell to formulate an alternative proof of product exhaustion. Dispensing with Wicksteed's assumption of linear homogeneity, they instead posited nonhomogeneous production functions yielding U-shaped long-run unit cost curves. They interpreted product exhaustion as an outcome of competitive equilibrium in which firms operate at the minimum point of these curves and charge a price equal to the minimum unit cost.

Their proof, already anticipated by Amstein in 1877, is straightforward. Into the competitive firm's unit cost equation

$$U = (wL + iC + \ \ldots)/P$$

substitute the cost-minimizing conditions of optimal factor hire. Since these conditions state that factor prices equal factor marginal physical products times product price π, such substitution yields the expression

$$U = \pi[(\partial f/\partial L)L + (\partial f/\partial C)C + \ \ldots]/P.$$

Divide both sides by U and multiply both sides by P to obtain

$$P = (\pi/U)[(\partial f/\partial L)L + (\partial f/\partial C)C + \ \ldots].$$

Note that free entry in long-run competitive equilibrium dictates that firms produce at the minimum, or zero profit, point on their unit cost functions where product price π equals unit cost U. The upshot is that the term π/U equals one and the equation reduces to the product-exhaustion condition

$$P = (\partial f/\partial L)L + (\partial f/\partial C)C + \ \ldots.$$

Wicksell was right. Evidently, competitive equilibrium ensures that even nonhomogeneous production functions deliver product exhaustion with factor shares adding up to unity. It is enough that the functions be momentarily homogeneous at the equilibrium point.[12]

7. WICKSELL'S ANTICIPATION OF THE COBB-DOUGLAS FUNCTION

The adding-up controversy had at least one important unintended consequence. It advanced knowledge of production functions to the point where the Cobb-Douglas equation, heretofore known only to Thünen, was within easy grasp of serious scholars. Wicksell is the key figure here. It was he who essentially transformed Thünen's implicit or disguised version of the Cobb-Douglas function into its exact or final form. And he did so on at least five occasions, the first appearing 27 years before and the last appearing four years before Cobb-Douglas. Owing to him, economists hardly had to wait for the equation to appear in 1928. Instead, they could refer to Wicksell, who was already using it at the turn of the century.

It is easy to trace the evolution of the equation in Wicksell's writings (see Olsson [1971], Sandelin [1976], and Velupillai [1973]). Like Thünen before him, Wicksell begins, in his 1896 *Finanztheoretische Untersuchungen,* by

---

[12] John R. Hicks (1932, pp. 234–39) proved as much without referring to product price. Minimize unit cost

$$U = (wL + iC \ldots)/P$$

by setting its first partial derivatives at zero. Make use of the definition that the sum of factor prices times factor quantities equals total cost or UP. The resulting partial derivatives

$$\partial U/\partial L = (1/P^2)[Pw - UP(\partial P/\partial L)] = (1/P)[w - U(\partial P/\partial L)]$$

and

$$\partial U/\partial C = (1/P^2)[Pi - UP(\partial P/\partial L)] = (1/P)[i - U(\partial P/\partial C)]$$

when set at zero reduce to

$$w = U(\partial P/\partial L)$$

and

$$i = U(\partial P/\partial C).$$

Substitute these into the unit cost equation to get

$$U = U[L(\partial P/\partial L) + C(\partial P/\partial C) + \ \ldots]/P.$$

Multiply both sides by P and divide both sides by U to get the product-exhaustion expression

$$P = (\partial P/\partial L)L + (\partial P/\partial C)C + \ \ldots .$$

presenting a per-worker version of the function.[13] Four years later, in his 1900 *Ekonomisk Tidskrift* piece on "Marginal Productivity as the Basis for Distribution in Economics," he advances to an exact replica of the function. He notes that the Wicksteed product-exhaustion formula for labor and land

$$P = (\partial P/\partial L)L + (\partial P/\partial T)T$$

is a partial differential equation that has the general solution

$$P = Lf(T/L).$$

He then cites as one example of this class of functions the Cobb-Douglas equation

$$P = L^{\alpha}T^{\beta},$$

where the exponents α and β sum to unity.

The Cobb-Douglas formula reappears in Volume 1 of his 1901 *Lectures on Political Economy*. There he adds that if the exponents α and β together exceed or fall short of unity, the factor shares will respectively over- and under-exhaust the product. He then insists that competition, by forcing firms to operate at minimum unit cost where constant returns prevail, ensures that the exponents sum to one as required by product exhaustion. Similarly, in correspondence with his colleague David Davidson in 1902, he uses the Cobb-Douglas function to prove that, with constant returns to scale, the joint marginal product of labor, land, and capital together equals the sum of their separate marginal products (see Uhr [1991]).

Again, in his 1916 article "The 'Critical Point' in the Law of Decreasing Agricultural Productivity," he employs the Cobb-Douglas function

$$P = L^{\alpha}T^{\beta}C^{\gamma}$$

to reconcile constant returns to scale with diminishing returns to proportions. You have constant returns to scale when a 10 percent increase in all inputs increases product by

$$10(\alpha + \beta + \gamma) = 10(1) = 10 \text{ percent.}$$

[13] His function is

$$p = ch^{m}t^{k}b^{v},$$

where p is output per worker, c is a constant, h is the land-to-labor ratio, t and b are the lengths of the investment periods of labor and land, respectively, and the exponents m, k, and v are fractions. Multiplying both sides by labor L yields

$$pL = P = cL(T/L)^{m}t^{k}b^{v} = cL^{1-m}T^{m}t^{k}b^{v}.$$

This function is Cobb-Douglas in labor and land but not in the investment periods. Nor should it be since such periods are hardly factor inputs logically parallel to labor and land.

You have diminishing returns to proportions when a 10 percent increase in both labor and capital, land held constant, increases product by

$$10(\alpha + \gamma) < 10 \text{ percent.}$$

Since output increases by less than the 10 percent increase in labor and capital, it follows that the average product of those inputs decreases. It does so because each unit of augmented labor and capital must work with a smaller amount of cooperating land.

Finally, in his 1923 review of Gustaf Akerman's doctoral dissertation *Realkapital und Kapitalzins*, Wicksell writes the Cobb-Douglas function as

$$P = cL^{\alpha}C^{\beta},$$

with the exponents $\alpha + \beta$ adding up to one. Clearly, if priority of discovery were the criterion, the names of Wicksell and Thünen should precede those of Cobb and Douglas when attached to the function. Credit should go to Cobb and Douglas not for inventing the function itself but for showing that it provides a good description of the aggregate data.

8. SOME FINAL OBSERVATIONS

The preceding discussion has concentrated exclusively on major landmarks in the history of production functions before Cobb-Douglas. In so doing, it undoubtedly has overlooked other milestones.

For example, nothing was said about the fixed-proportion production functions of Richard Cantillon (1755), Léon Walras (1874), and Gustav Cassel (1918). These functions foreshadowed modern Leontief functions and share the same features. Production is characterized by rigidly fixed input coefficients that rule out factor substitution. Such coefficients specify input requirements per unit of output. Thus if one hour of labor L (assisted of course with the required amounts of cooperating inputs) can produce ten units of product P such that each unit of product requires a tenth of an hour of labor, then $L/P = 1/10$ is the input coefficient of labor in producing output. Similar coefficients hold for other required inputs.

Denote the production coefficients of labor, capital, and land as l, c, and t, respectively. Then Cantillon, Walras, Cassel, and Leontief could write the production function as

$$P = \min\ (L/l, C/c, T/t).$$

Here the terms L/l, C/c, and T/t are, respectively, the largest outputs producible from the quantities of labor, capital, and land available. The smallest of these terms determines the level of output. Why? Because with fixed factor proportions, output is limited by the relatively scarcest factor, just as the size of a cake is limited by the recipe ingredient in shortest supply. Given the quantity

of the limitational ingredient, extra units of the other ingredients would not increase output; their marginal products are zero. At the point of limitation, output absorbs inputs precisely in the ratio $l : c : t$.

By 1900, however, most economists, including Walras himself, were balking at the evidently unrealistic notion of fixed proportions and zero factor substitution. Already they were employing variable-coefficient functions rather than fixed-coefficient ones. Fixed production coefficients, however, made a comeback in the 1950s and 1960s in linear programming and input-output models.

The preceding paragraphs also failed to mention the British physicist Lord Kelvin's 1882 engineering production function for the transmission of electric power (see Smith [1968], p. 515). Kelvin's pioneering work (see Appendix) foreshadowed modern engineering production functions for gas and heat transmission, steam power production, metal cutting, and batch reactor chemical processes.

Nevertheless, enough has been said to document the main contention of the article, namely that algebraic production functions long predate Cobb-Douglas. At least 18 economists from seven countries over a span of 160 years either presented or described such functions before Cobb-Douglas. Seen in this perspective, the Cobb-Douglas function and its more recent successors represent the culmination of a long tradition rather than the beginning of a new one.

APPENDIX

Lord Kelvin (William Thompson) related electric power output P (the quantity of electric energy delivered) to two factors, namely power input I and the size S of the copper cable, or conductor, through which the electric current is transmitted. Power output is that part of power input not lost through frictional heating of the cable. Such loss varies directly with the square of power output and inversely with the size (weight or volume) of the cable. The result is the implicit production function

$$P = I - (kP^2/S),$$

where k is a constant that depends on the length and resistance properties of the cable. When solved explicitly for output,

$$P = (S/2k)\{[1 + (4kI/S)]^{1/2} - 1\},$$

Kelvin's function ascribes quantity of electric power delivered to three determinants, namely power input, cable size or weight, and the constant of resistance. From this function derives Kelvin's famous law stating that the conductor cable reaches its optimum size when the annual interest cost of the copper invested in the cable equals the value of the energy lost annually through heating of the cable.

REFERENCES

Amstein, Hermann. Correspondence on Marginal Productivity Theory, 1877, in William Jaffé, "New Light on an Old Quarrel," *Cahiers Vilfredo Pareto*, No. 3, 1964, pp. 94–97, and William Jaffé, ed., *Correspondence of Léon Walras and Related Papers*, 1857–1883, Vol. I. Amsterdam: North-Holland, 1965, pp. 516–20, reprinted in William J. Baumol and Stephen M. Goldfeld, eds., *Precursors in Mathematical Economics: An Anthology.* London: The London School of Economics and Political Science, 1968, pp. 309–12.

Barkai, H. "Ricardo on Factor Prices and Income Distribution in a Growing Economy," *Economica*, vol. 26 (August 1959), pp. 240–50.

Barone, Enrico. "Sur un livre récent de Wicksteed," 1896, French translation by Léon Walras in *Cahiers Vilfredo Pareto*, No. 3, 1964, pp. 68–73.

Berry, Arthur. "The Pure Theory of Distribution," *Report of the Sixtieth Meeting of the British Association for the Advancement of Science*, 1891, reprinted in William J. Baumol and Stephen M. Goldfeld, eds., *Precursors in Mathematical Economics: An Anthology*. London: The London School of Economics and Political Science, 1968, pp. 314–15.

Blaug, Mark. *Economic Theory in Retrospect*, 4th ed. Cambridge: Cambridge University Press, 1985.

Cantillon, Richard. *Essai sur la nature du commerce en général*, 1755, edited with English translation and other material by Henry Higgs. London: Macmillan (for the Royal Economic Society), 1931.

Cassel, Gustav. *Theoretische Sozialökonomie*. Leipzig: C. F. Winter, 1918, translated into English as *Theory of Social Economy.* London: T. F. Unwin, 1923, new revised edition London: E. Benn, 1932.

Cobb, Charles W., and Paul H. Douglas. "A Theory of Production," *American Economic Review*, vol. 18 (Supplement) (March 1928), pp. 139–65.

Cournot, Antoine Augustin. *Researches into the Mathematical Principles of the Theory of Wealth*, 1838, translated by Nathaniel T. Bacon. New York: Macmillan, 1929, reprinted, New York: Augustus Kelley, 1971.

Creedy, John. "The Early Use of Lagrange Multipliers in Economics," *Economic Journal*, vol. 90 (June 1980), pp. 371–76.

Dickinson, H. D. "Von Thünen's Economics," *Economic Journal*, vol. 79 (December 1969), pp. 894–902.

Edgeworth, Francis Y. "The Theory of Distribution," *Quarterly Journal of Economics*, 1904, reprinted in Francis Y. Edgeworth, *Papers Related to Political Economy*, Vol. 1. London: Macmillan (for the Royal Economic Society), 1925.

_____. Review of *Natural Value* by Friedrich von Wieser, *Economic Journal*, 1894, reprinted in Francis Y. Edgeworth, *Papers Related to Political Economy*, Vol. 3. London: Macmillan (for the Royal Economic Society), 1925.

_____. "On the Application of Mathematics to Political Economy," *Journal of the Royal Statistical Society*, vol. 52 (December 1889), pp. 538–76, reprinted in Francis Y. Edgeworth, *Papers Related to Political Economy*, Vol. 2, London: Macmillan (for the Royal Economic Society), 1925.

_____. *Mathematical Psychics*. London: C. Kegan Paul & Co., 1881.

_____. *New and Old Methods of Ethics*. Oxford: James Parker & Co., 1877.

Flux, A. W. Review of Philip H. Wicksteed, *Essay on the Co-ordination of the Laws of Distribution*, in *Economic Journal*, vol. 4 (June 1894), pp. 308–13, reprinted in William J. Baumol and Stephen M. Goldfeld, eds., *Precursors in Mathematical Economics: An Anthology*. London: The London School of Economics and Political Science, 1968, pp. 326–31.

Hicks, John R. *The Theory of Wages*. New York: Macmillan, 1932.

Jaffé, William. "New Light on an Old Quarrel," *Cahiers Vilfredo Pareto*, No. 3, 1964, pp. 61–102.

Johnson, William E. "Exchange and Distribution," *Cambridge Economic Club*, 1891, reprinted with brief commentary in William J. Baumol and Stephen N. Goldfeld, eds., *Precursors in Mathematical Economics: An Anthology*. London: The London School of Economics and Political Science, 1968, pp. 316–20.

Kelvin, Lord (William Thompson). "On the Economy of Metal in Conductors of Electricity," 1882, Report of 51st meeting, held August-September 1881, British Association for the Advancement of Science, *Reports* 51, pp. 526–28.

Leigh, Arthur H. "Thünen, Johann Heinrich von," in David L. Sills, ed., *International Encyclopedia of the Social Sciences*, Vol. 16. Macmillan and Free Press, 1968, pp. 17– 20.

Lloyd, Peter J. "Elementary Geometric/Arithmetic Series and Early Production Theory," *Journal of Political Economy*, vol. 77 (January/February 1969), pp. 21–34.

Malthus, Thomas Robert. *An Essay on the Principle of Population*, 1st ed. London, 1798.

Mitscherlich, E. A. "Das Gesetz des Minimums und das Gesetz des abnehmenden Bodenertrages," *Landw. Jahrb*., vol. 38 (1909), pp. 537–52.

Niehans, Jurg. *A History of Economic Theory*. Baltimore: Johns Hopkins University Press, 1990.

Olsson, Carl-Axel. "The Cobb-Douglas or the Wicksell Function?" *Economy and History,* vol. 14 (1971), pp. 64–69.

Pareto, Vilfredo. *Cours d'économie politique,* II. Paris, 1897.

Pigou, Arthur Cecil, ed. *Memorials of Alfred Marshall*. London: Macmillan, 1925, reprinted, New York: Kelley and Millman, 1956.

Ricardo, David. *On the Principles of Political Economy and Taxation*. London: Murray, 1817, reprinted in Piero Sraffa, ed., *The Works and Correspondence of David Ricardo,* Vol. 1. Cambridge: Cambridge University Press, 1951.

Samuelson, Paul A. "Paul Douglas's Measurement of Production Functions and Marginal Productivities," *Journal of Political Economy,* vol. 87 (October 1979), pp. 923–39.

Sandelin, Bo. "On the Origin of the Cobb-Douglas Production Function," *Economy and History,* vol 19 (No. 2, 1976), pp. 117–23.

Schumpeter, Joseph A. *History of Economic Analysis*. London: Allen & Unwin, 1954.

Smith, Vernon L. "Production," in David L. Sills, ed., *International Encyclopedia of the Social Sciences,* Vol. 12. Macmillan and Free Press, 1968, pp. 511–18.

Spillman, W. J., ed. *The Law of Diminishing Returns*. Yonkers-on-Hudson, N.Y.: World Book Co., 1924.

Steedman, Ian. "Adding-up Problem," in John Eatwell, Murray Milgate, and Peter Newman, eds., *The New Palgrave: A Dictionary of Economics,* Vol. 1. London: Macmillan, 1987, pp. 21–22.

Stigler, George J. "The Ricardian Theory of Value and Distribution," *Journal of Political Economy,* vol. 60 (June 1952), pp. 187–207, reprinted in George J. Stigler, *Essays in the History of Economics*. Chicago: University of Chicago Press, 1965, pp. 156–97.

_____. *The Theory of Price.* New York: Macmillan, 1946.

_____. *Production and Distribution Theories: The Formative Period*. New York: Macmillan, 1941.

Stigler, Stephen M. "Stigler's Law of Eponymy," *Transactions of the New York Academy of Sciences,* 2d series, vol. 39 (April 1980), pp. 147–58.

Thünen, Johann Heinrich von. *Der isolierte Staat in Beziehung auf Landwirthschaft und Nationalökonomie.* 3 volumes. Jena, Germany: Fischer, 1930. A partial translation of *The Isolated State in Relation to Agriculture and Political Economy,* Vol. 2, "The Natural Wage and Its Relation to the Rate of Interest and to Economic Rent" is included in Bernard W. Dempsey, *The Frontier Wage: The Economic Organization of Free Agents.* Chicago: Loyola University Press, 1960.

Turgot, Anne Robert Jacques. *Observations sur le Mémoire de M. de Saint-Péravy,* Limoges, 1767, in E. Daire, ed., *Oeuvres de Turgot,* Paris, 1844, pp. 414–33, translated as "Observations on a Paper by Saint-Péravy on the Subject of Indirect Taxation," in P. D. Groenewegen, *The Ecomomics of A. R. J. Turgot,* The Hague, 1977, pp. 43–95.

Uhr, Carl H. "David Davidson: The Transition to Neoclassical Economics," in Bo Sandelin, ed., *The History of Swedish Economic Thought*. London: Routledge, 1991, pp. 44–75.

Velupillai, Kumaraswamy. "The Cobb-Douglas or the Wicksell Function?—A Comment," *Economy and History,* vol. 16 (1973), pp. 111–13.

Walras, Léon. "Note on Mr. Wicksteed's Refutation of the English Theory of Rent," Appendix III of *Eléments d'économie politique pure,* 3d ed. Lausanne, 1896.

_____. *Eléments d'économie politique pure,* 1st ed. Lausanne: F. Rouge, 1874.

Whitaker, J. K., ed. *The Early Economic Writings of Alfred Marshall, 1867–1890,* Vol. 2. New York: The Free Press, 1975.

Wicksell, Knut. *Über Wert, Kapital, und Rente,* Jena, Germany: G. Fischer, 1893, translated by S. H. Frowein as *Value, Capital, and Rent.* London: Allen & Unwin, 1954, reprinted, New York: Augustus M. Kelley, 1970.

_____. "The 'Critical Point' in the Law of Decreasing Agricultural Productivity," 1916, in E. Lindahl, ed., *Selected Papers on Economic Theory by Knut Wicksell.* London: Allen & Unwin, 1958.

_____. "On the Problem of Distribution," 1902, in E. Lindahl, ed., *Selected Papers on Economic Theory by Knut Wicksell.* London: Allen & Unwin, 1958.

_____. "Marginal Productivity as the Basis for Distribution in Economics," *Ekonomisk Tidskrift,* 1900, in E. Lindahl, ed., *Selected Papers on Economic Theory by Knut Wicksell*. London: Allen & Unwin, 1958, pp. 93–121.

_____. "Real Capital and Interest," 1923, Review of Gustaf Akerman's *Realkapital und Kapitalzins,* in Lionel Robbins, ed., *Lectures on Political Economy,* Vol 1: *General Theory,* Appendix 2. London: Routledge & Kegan Paul, 1934.

_____. *Föreläsningar i nationalekonomi,* Vol. I, Lund, 1901, translated by E. Classen in the third Swedish edition, 1928, and in Lionel Robbins, ed., *Lectures on Political Economy,* Vol. 1: *General Theory.* London: Routledge & Kegan Paul, 1934.

_____. *Finanztheoretische Untersuchungen nebst Darstellung und Kritik des Steurewesens Schwedens.* Jena, Germany: G. Fischer, 1896.

Wicksteed, Philip H. *An Essay on the Co-ordination of the Laws of Distribution.* London: Macmillan & Co., 1894, revised with an introduction by Ian Steedman, Aldershot, England: Edward Elgar, 1992.