History and Theory of the NAIRU: A Critical Review

MARCO A. ESPINOSA-VEGA AND STEVEN RUSSELL

Espinosa is a senior economist at the Federal Reserve Bank of Atlanta. Russell is an assistant professor of economics at Indiana University-Purdue University at Indianapolis. *They thank, without implicating, Eric Leeper, Maurice Obstfeld, Mary Rosenbaum, and Charles Whiteman for preliminary conversations on the topic and Bob Eisenbeis and Robert Bliss for helpful expositional comments on the final draft.*

WHAT CAUSES CHANGES IN THE RATES OF INFLATION AND UNEMPLOYMENT? HOW ARE THE PRICE LEVEL AND THE LEVEL OF EMPLOYMENT RELATED? THESE HAVE BEEN KEY QUESTIONS FACING ECONOMISTS FOR AT LEAST FORTY YEARS. DISCUSSIONS ABOUT THEM IN THE PRESS AND ELSEWHERE OFTEN CENTER ON AN APPROACH TO EXPLAINING THE INFLATION-UNEMPLOYMENT RELATIONSHIP THAT DATES BACK TO THE 1960S AND 1970S. ACCORDING TO THIS APPROACH, INFLATION IS CAUSED BY AN EXCESSIVELY TIGHT LABOR MARKET THAT DRIVES UP WAGES AND FORCES FIRMS TO RESPOND BY RAISING PRICES.

An important element of this approach is the concept of a nonaccelerating inflation rate of unemployment, or NAIRU. As its name suggests, the NAIRU is supposed to be an unemployment rate (or range of unemployment rates) that produces a stable rate of inflation: if the unemployment rate is lower than the NAIRU then the inflation rate will tend to rise, and vice versa.

Recently, both the NAIRU and the theory of the inflation-unemployment relationship on which it is based have received a great deal of attention from the press. From December 1995 to December 1996, for example, there were ten articles on this subject in the *Wall Street Journal*, five articles in the *New York Times*, and three in *The International Economy*. One common feature of all these articles is that they link Federal Reserve monetary policy to the NAIRU. Most of the authors seem to assume that the NAIRU is or should be the Fed's principal guide for conducting monetary policy. According to this view, if the current unemployment rate is below some NAIRU estimate (say, 6 percent) then the Fed should tighten monetary policy to head off a coming increase in the inflation rate.

Despite the extensive press coverage the NAIRU concept has received recently, the theory of the inflation-unemployment relationship that it is part of is quite controversial. Although the NAIRU is alive and well in the media and among economic policymakers, it is no longer very popular among academic economists. It has fallen out of favor partly because its conceptual foundation is

weak and partly because its empirical track record does not inspire confidence. Its survival is due largely to the fact that economists have not been able to reach any consensus about alternative guides for monetary policy.

The purpose of this article is to provide some historical perspective on the "NAIRU theory" and the assumptions behind it. Most of the analysis presented in this article is not original: it has been around for two decades or more. However, the recent resurgence of interest in the NAIRU indicates that there may be a need for a basic review of its origins and a brief explanation of some of the claims surrounding it. Readers interested in additional details should consult the reference list.

The first section of the discussion that follows briefly introduces the Keynesian and classical theories of macroeconomics. Keynesian theory is the macroeconomic theory on which the NAIRU is principally based while classical theory provides the foundation for the monetarist and neoclassical critiques of Keynesian theory that are discussed at length in this article. As we shall see, the concept of a NAIRU grew out of economists' attempts to reconcile the differences between Keynesian and monetarist theories on the subjects of the causes of price level changes and the relationship between inflation and unemployment. The next section discusses the Phillips curve, a description of the inflation-unemployment relationship that provided the empirical and theoretical starting points for the development of the NAIRU. The third section reviews the monetarist critique of analysis based on the Phillips curve and discusses a number of related questions. The next two sections explain how the NAIRU developed as a response to the monetarist critique of the Phillips curve and raise some basic questions about the NAIRU. The final part of the discussion reviews the concept of rational expectations, a theoretical contribution of neoclassical theory that amplified the monetarist critique of the Phillips curve. This section also discusses some neoclassical contributions that may offer alternatives to the Phillips curve approach to the study of inflation, unemployment, and the effects of monetary policy.

Two Economic Traditions

Classical economic theory developed in the early 1900s, at a time when there was no formal distinction between micro- and macroeconomics. The theory was based on the same basic assumptions that had become widely used to study the behavior of individual households and firms. These included the assumptions that individuals usually act in ways that maximize their self-interest, that prices are determined in the marketplace, and that markets operate efficiently. According to classical theory, perfect competition is a good approximation of the operation of most real-life markets. The basic assumptions of classical theory are generally understood to imply that government policies have relatively little importance in determining economic outcomes.

Keynesian theory, which developed in the 1930s and 1940s, was the first macroeconomic theory: it was designed specifically to study economywide phenomena, and it was not simply an extension of the conventional economic theory that continued to be used to study the behavior of individual parts and sectors of the economy. Keynesian theory was based on the work of John Maynard Keynes, a British economist who did most of his work in the 1920s and 1930s. One of the basic goals of Keynes's theory was to explain the persistently high rates of unemployment that appeared across the world during the Great Depression. Most of this unemployment was generally believed to be "involuntary," in the sense that the unemployed people were willing to work at the going wage rates but were unable to find jobs. A closely related goal of Keynes was to identify steps that the government could take to alleviate the high levels of unemployment.

The NAIRU has fallen out of favor among academic economists partly because its conceptual foundation is weak and partly because its empirical track record does not inspire confidence.

Keynesian theory assumes that some important prices are determined or strongly influenced by forces outside the marketplace so that many markets may not be able to "clear" in the sense of successfully reconciling demand with supply. It also assumes that people may not always make the economic decisions that would be best for them. According to Keynesian theory, perfect competition is not a good approximation of the operation of many important real-life markets. The theory implies that government policies can have large, important effects on the economy and that if the policies are carefully devised these effects can be very constructive in nature.[1]

Keynes's ideas and goals placed him in direct conflict with the exponents of the reigning classical theory.

1. The monetarist and neoclassical theories developed later—monetarism in the 1950s and neoclassical theory in the 1970s. These theories were developed as alternatives to Keynesian theory, which was then accepted by most contemporary economists. Both theories drew heavily on the classical tradition. As we shall see, the economic theory behind the NAIRU is basically Keynesian in nature, but it has been influenced heavily by monetarist ideas and to a lesser extent by neoclassical ones.

Classical theory predicted that when unemployment was high wages would adjust downward, stimulating more hiring and reducing the unemployment rate. As a result, high unemployment could not last long. It seemed obvious to Keynes (and many others) that the high, persistent levels of unemployment observed during the Depression were inconsistent with this prediction and that classical theory was incapable of explaining them. In 1933 prominent classical theorist A.C. Pigou published *The Theory of Unemployment*; according to Keynes, this book was "the only detailed account of the classical theory of employment" in existence at the time. In his "General Theory" article, Keynes dismisses Pigou's book as "a non-causative investigation into the functional relationship which determines what level of real wages will correspond to any given level of employment. . . . [It] is not capable of telling us what determines the *actual* level of employment; and on the problem of involuntary unemployment it has no direct bearing" (1964, 275).

Classical economic theory was based on the assumptions that individuals usually act in ways that maximize their self-interest, that prices are determined in the marketplace, and that markets operate efficiently.

According to Keynes, what prevented labor markets from clearing, and explained involuntary unemployment, was that when firms' demand for labor decreased, nominal (money) wages did not fall as fast or as far as classical theory predicted.[2] "Classical theory," he comments, "has been accustomed to rest the supposedly self-adjusting character of the economic system on an assumed fluidity of money-wages" (1964, 257). Keynes believed that sluggish labor demand would not push nominal wages downward, at least in the short run. The logic behind this belief was that organized workers had enough market power to resist employers' attempts to reduce money wage rates. As a result, Keynesian theory is often described as being based on the assumption of "sticky wages."[3] In the classical model, unlike the Keynesian model, money wages and prices are assumed to be perfectly flexible, so labor markets always clear. If temporary unemployment appears because of deficient aggregate demand, then the unemployed workers will bid down nominal wages until they have fallen far enough to eliminate the unemployment.

Keynes also criticized classical theory for failing to provide an integrated analysis of the behavior of different parts of the economy and for making an unwarranted leap from analysis of individual-industry labor markets to analysis of the determinants of aggregate employment. He writes that "if the classical theory is not allowed to extend by analogy its conclusions in respect of a particular industry to industry as a whole, it is wholly unable to answer the question what effect on employment a reduction in money-wages will have. For it has no method of analysis wherewith to tackle the problem" (1964, 257).

Over time, it became clear that both classical and Keynesian theories suffered from some important deficiencies. Classical theorists needed to integrate their microeconomic theories of individual labor markets into a macroeconomic theory of total employment. They also needed to explain how government policies affected the labor market. The Keynesians needed to move in the opposite direction, integrating their macroeconomic theory with a microeconomic theory of labor markets and formalizing their explanation of wage-setting behavior.

The Phillips Curve

Inflation and Unemployment. In 1958 British economist A.W. Phillips published the results of an empirical analysis of historical data from the U.K. labor market. Phillips's study was intended to help answer one of the basic questions in macroeconomic theory, which concerns the cause of inflation. He hoped to find empirical support for the Keynesian view that the rate of wage inflation—that is, the rate of increase in nominal (money) wage rates—depended on the tightness of the labor market. Since the level of unemployment was a readily observable indicator of the tightness of the labor market, Phillips's immediate goal was "to see whether statistical evidence supports the hypothesis that the rate of change of money wage rates in the United Kingdom can be explained by the level of unemployment and the rate of change of unemployment" (1958, 284).

The logic behind Phillips's theory is very simple. If for some reason the demand for labor were high relative to its supply—as in Atlanta during the Olympics, to use a modern example—then equilibrium wage rates would be expected to rise above current wage levels, and there would be upward pressure on nominal wages as firms bid for additional workers. As additional workers were actually hired, moreover, the unemployment rate would fall. The larger the discrepancy between the quantity of labor demanded and the quantity supplied, the stronger the upward or downward pressure on wage rates. The opposite would be true when there was excess supply of labor and rising unemployment.

Phillips found, as he expected, that from 1861 to 1957 the growth rate of nominal wages was negatively correlated with the rate of unemployment—that is, low unemployment rates tended to be associated with rapidly rising wages while high unemployment rates were associated with slowly rising wages. Phillips also found that the strength of the unemployment versus wage-change relationship seemed to depend on the level of unemployment.

When unemployment was low, decreases in unemployment tended to be associated with big increases in wage inflation while when unemployment was high, decreases in the unemployment rate seemed to produce small increases in wage growth rates (see Chart 1 for a hypothetical Phillips curve). These findings appeared to confirm Keynes's theory of the downward stickiness of nominal wages. Tight labor markets seemed to cause employers to bid wages up rapidly while loose markets (high unemployment) seemed to cause workers to bid wages down relatively slowly.

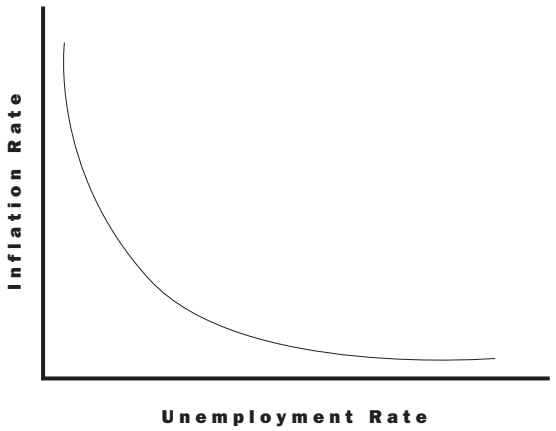
Phillips's findings have had a profound and lasting effect on economists' ideas about the relationship between inflation and unemployment. What made them so interesting is that they seemed to establish a clear linkage between the state of the labor market and the rate of inflation. By the early 1960s, inflation rates in the United States and western Europe had increased to the point that inflation was coming to be regarded as a serious economic problem. As a result, economists and policymakers were eager for information about its possible causes and potential cures. The Phillips curve appeared to link the real and nominal sides of the economy.[4]

One possible objection to the conclusions that Phillips (and others) drew from his findings is that standard economic theory predicts that what matters to workers is not their nominal wages but their real, or inflation-adjusted, wages.[5] Phillips did not attempt to measure real wages or study their statistical relationship to unemployment. Under the Keynesian assumption of predetermined or sticky nominal prices, however, changes in expected real and nominal wages would coincide. In addition, while Phillips's statistical evidence involved changes in current nominal wages, the hypothesis that he was trying to test involved changes in *expected* nominal wages. If workers were slow to adjust their price expectations to actual price changes, changes in current nominal wages could be interpreted as changes in expected real wages.

Another problem with Phillips's findings is that they involve wage inflation while economists were principally concerned about explaining price inflation. Since wages are the biggest single component of firms' costs, however, most economists were willing to assume that persistent increases in wage rates would eventually force firms to begin increasing their prices, producing economywide price inflation. For this explanation for inflation to make sense, however, it was necessary to make even more elaborate assumptions about stickiness: wages now had to be assumed to adjust faster than goods prices, at least when wages were rising. (In conventional Keynesian theory, nominal wages were supposed to be slow to fall when a decrease in aggregate demand put downward pressure on prices; the result was a higher-than-equilibrium real wage and involuntary unemployment.)

How was the Phillips curve related to monetary policy? Keynesian theory held that monetary policy could be used to increase or decrease the economy's aggregate demand—the total nominal demand for goods and services of all types—and through it the aggregate level of employment in the economy. The Phillips curve mechanism explained how aggregate demand management could affect the rate of inflation. Thus, economic policymakers began to think in terms of a trade-off between the unemployment rate and inflation rate. Although government aggregate-demand stimulus was no longer cost-free, as it had been in traditional Keynesian theory (which had viewed the price level as constant), it was still possible for the policy authority to reduce the level of employment if it was willing to tolerate the resulting increase in inflation along the Phillips curve. As the next section will show, another reason for the popularity of

CHART 1
The Phillips Curve



The now-conventional Phillips curve diagram has the unemployment rate on the horizontal axis and the inflation rate on the vertical axis.

2. According to Keynes, the principal source of the observed fluctuations in labor demand was the volatility of aggregate investment. Investment volatility, in turn, was caused by changes in short- and long-term business expectations and variation in interest rates.

3. The discussion will show that the stickiness assumption was also extended to aggregate prices.

4. Phillips was not the first researcher to turn up findings of this general sort. As long ago as 1926 Irving Fisher had found a negative correlation between the rate of goods-price inflation and the level of unemployment.

5. If workers in New York City and rural Mississippi both make $2,500 per month, the worker in rural Mississippi will have a much higher real wage because the cost of living is lower there.

the Phillips curve is that it was seen by some prominent economists as providing a synthesis of competing theories of inflation.

Cost-Push versus Demand-Pull Inflation. At the time the Phillips curve analysis appeared, economists' interest in understanding the relationship between wages, prices, and economic activity had been growing for some time, and there was also growing interest in studying the effects of government policies on this relationship. Samuelson and Solow (1960) provide a comprehensive review of the debate on these questions that took place after the Second World War. The debate centered on two basic theories of the causes of inflation: demand-pull and cost-push. Both theories can be explained using the aggregate-demand/aggregate-supply model of output and price level determination that was developed during the 1950s and remains popular in textbooks. Demand-pull inflation resulted from increases in the level of aggregate demand that occurred at or near the point of full capacity utilization—that is, at points at which the aggregate supply curve was upward-sloping rather than flat. Cost-push inflation, on the other hand, was caused by upward shifts in the aggregate supply curve. These shifts could allow wages and prices to rise even before full employment was reached.[6]

According to Samuelson and Solow, there were really no purists in this debate. Most economists believed that inflation had both demand-pull and cost-push components, but they differed as to which component predominated. Thus, although demand-pull inflation was associated with Keynesian theory, Keynes himself did not dismiss the cost-push hypothesis. He was "willing to assume that attainment of full employment would make prices and wages flexible upward. . . . Just as wages and prices may be sticky in the face of unemployment and overcapacity, so may they be pushing upward beyond what can be explained in terms of levels and shifts in demand" (1964, 180-81).

Samuelson and Solow believed that in order to reconcile the two sides of this debate it would be necessary for economists to improve their understanding of the behavior of money wages with respect to the level of employment. They saw the Phillips curve as a useful tool for analyzing this behavior. Under some conditions, they explained, "movements along the Phillips curve might be dubbed standard demand-pull, and shifts of the Phillips curve might represent the institutional changes on which cost-push theories rest" (1960, 189).

The Monetarist Challenge to the Keynesian Approach

The Acceleration Hypothesis. One prominent U.S. economist who was skeptical of Keynesian theory in general, and of Phillips curve analysis in particular, was Milton Friedman. Friedman was the champion of monetarism, a theory that saw inflation as always and everywhere a monetary phenomenon. He was also rather skeptical of the Keynesian view that demand-management policy could have significant effects on output or employment. Beginning in the mid-1960s, Friedman began to challenge some of the conclusions about the inflation-unemployment relationship that economists writing in the 1960s and early 1970s were drawing on the basis of Keynesian theory.

As we have seen, Keynes's explanation for persistent unemployment was that the prevailing level of real wages was not compatible with labor market clearing and instead produced excess supply of labor. This fact raised the question of why lower, market-clearing real wages could not be produced by reductions in nominal wages. One explanation frequently offered was that workers would oppose nominal wage reductions. Friedman (1976) was very skeptical about this and other explanations that Keynesians put forward to explain supposed nominal wage rigidities. He was willing to concede that there might be some situations in which wages and salaries were rigid; the legal minimum wage, he noted, was an example of such a rigidity. He argued, however, that situations like these were the exception rather than the rule. In most industries, he pointed out, relatively few workers earned the minimum wage: what prevented workers in these industries from reducing their wage requests in order to avoid layoffs? And while unions could conceivably be a factor delaying wage adjustment because of their reluctance to accept wage cuts that would benefit unemployed workers at union members' expense, he did not believe that unions were powerful or perverse enough to keep wages from adjusting to full employment levels in the long run.

A second criticism Friedman raised was that researchers had not been able to construct "decent" empirical Phillips curves for the United States or other countries. In later years this problem got worse, and even ardent Keynesians were forced to acknowledge the weakness of the empirical evidence supporting the existence of stable national Phillips curves. In 1980, for example, prominent Keynesian Arthur Okun, commenting on the U.S. case, wrote that "since 1970, the Phillips curve has been an unidentified flying object and has eluded all econometric efforts to nail it down" (1980, 166).

Friedman's third criticism was outlined in the previous section: Phillips's statistical evidence involved nominal wages, but standard economic theory assumes that households and firms base their employment decisions on real wages. Clearly, Phillips and his successors were assuming that changes in current nominal wages were equivalent to changes in expected future real wages. This assumption, Friedman noted, really amounted to two assumptions. The first was that prices, or at least price expectations, were rigid: people did not expect the price level to change and consequently interpreted changes in

their nominal wages as changes in their real wages. The second assumption was that workers would not resist reductions in their real wages that were caused by inflation rather than by reductions in their nominal wages. Only if both assumptions were true could the relationship between the rate of change in nominal wages and the aggregate level of unemployment be stable enough to then offer policymakers a usable menu of options.

A closely related argument made by both Friedman (1968) and Phelps (1967) involved the long-run implications of the Phillips curve. In order to make this argument, they imagined a situation in which a policymaker was trying to use the hypothesized inflation-unemployment trade-off to achieve a lasting reduction in the unemployment rate. Such a policymaker, they argued, would find that while there might indeed be an inflation-unemployment trade-off in the short run, the trade-off would disappear in the long run. In the long run, they asserted, unemployment tended to return to a "natural rate" (NR) that was determined by real economic forces.[7] Monetary policy, in their view, could do nothing to change the natural rate.

The analysis presented by Friedman and Phelps, which was later summarized by Friedman (1976), involved the relationship between real wages and unexpected inflation. The emphasis on unexpected inflation reflected an attempt on the part of Friedman and Phelps to reconcile the classical principle that labor supply behavior depends on the real wage with Keynes's observation that workers respond differently to different types of real wage decreases: "Every trade union," Keynes writes, "will put up some resistance to a cut in money-wages, however small, . . . but no trade union would dream of striking on every occasion of a rise in the cost of living" (1964, 14-15). According to Friedman, this differential response is due to temporary money illusion: it takes time for workers to recognize that the price level has increased, and until they do so they do not realize that their real wage rates have fallen.

Friedman's discussion can be interpreted as an implicit description of the following hypothetical sequence of events. Suppose the economy starts out in its long-run equilibrium at its normal inflation rate and its natural rate of unemployment. This equilibrium is disturbed when a monetary expansion increases households' aggregate demand for goods and services at current prices. Demand curves will shift to the right throughout the economy, and the market prices of (output) goods and services will rise. The increased market prices of goods will cause the aggregate demand curve for labor, plotted against the nominal wage, to shift to the right.

If workers realize that the price level has increased, then their aggregate labor supply curve will shift to the left, as depicted in the shift from curve D to curve D^1 in Chart 2. Equilibrium will be restored at a higher nominal wage rate but at unchanged levels of employment, output, and real wages. If workers do not realize that the price level has increased—that is, if the increase in prices is both unperceived and unexpected—then employment and nominal wage rates will increase along the old labor supply curve. Workers will now be providing more labor than they would be willing to provide at the current real wage if they knew what that wage really was. At some point, however, workers will figure out that the price level has increased, and the aggregate labor supply curve will begin shifting to the left, as depicted in the shift from curve S to S^1 in Chart 2. The shift in the supply curve will drive nominal wages up further. As nominal wages rise, the supply curves for goods and services will shift to the left, driving the price level up further, and so on. Nominal wages will rise faster than prices, however, as workers catch on to the successive price increases. Eventually a new long-run equilibrium is reached at the original unemployment rate (the natural rate) and the original level of real wages—the point L^0 in Chart 2. Notice that once the process of adjustment to the new long-run equilibrium gets started, prices lead wages upward rather than the reverse.

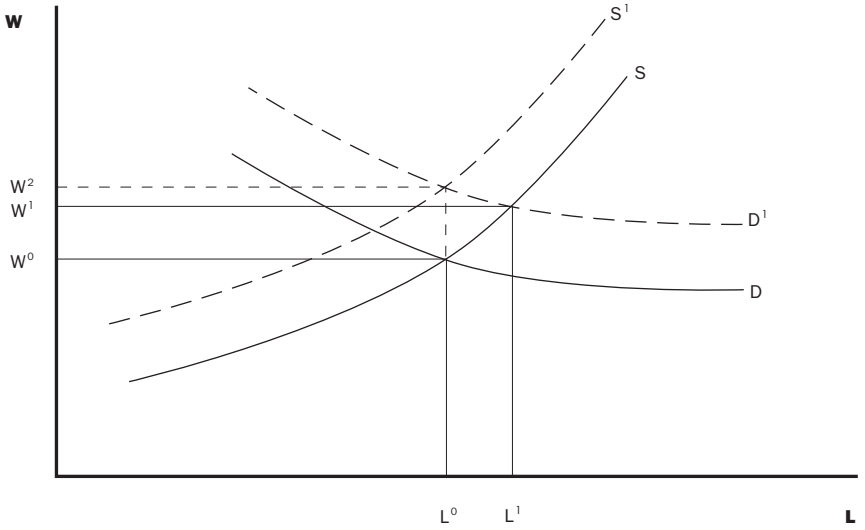
Keynesian theory implies that government policies can have large, important effects on the economy and that if the policies are carefully devised these effects can be very constructive in nature.

To summarize, Friedman and Phelps argued that unexpected inflation can drive the level of unemployment

6. Believers in cost-push inflation often identified unions as one of its main sources. Samuelson and Nordhaus point out, however, that "this view of unions as the clear-cut villain of cost-push inflation does not fit the complex historical facts. Take as an example the depressed year of 1982, when unemployment averaged 9.7 percent of the labor force. During that year, labor costs for union workers rose 7.2 percent, and the cost of nonunion workers rose 6 percent. Both union and nonunion wages rose smartly in spite of high unemployment" (1989, 326).

7. Friedman defines the natural rate of unemployment as the level of unemployment "that would be ground out by the Walrasian system of general equilibrium equations, provided there is imbedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the cost of mobility, and so on" (1968, 8).

CHART 2 Effects of Monetary Policy on the Labor Market



Friedman and Phelps argued that unexpected inflation can drive the level of unemployment below the natural rate, but only temporarily.

below the natural rate, but only temporarily. In the long run, the surprise factor will disappear as workers learn that the price level has increased; as a result, the level of employment will go back to the natural rate. Thus, in the long run there will be no inflation-unemployment trade-off. Stated differently, the long-run Phillips curve is vertical.

At this point, it is necessary to make some important distinctions regarding the term *inflation*. A one-time increase in the price level is sometimes called inflation, but it is very different from a situation in which the price level is increasing over time at a constant rate. Both these situations, moreover, are different from one in which the price level is increasing over time at a rate that is also increasing over time (so that the price level is accelerating upward). The inflation that the Keynesian economists who developed Phillips curve analysis had in mind was the type in which the price level increases at a fixed rate. These economists believed that from the point of view of policymakers, the cost of achieving a lower level of unemployment was that the price level would now increase at a higher rate. Inflation would remain constant at this new, higher rate as long as the unemployment rate remained at its new, lower level.

According to Friedman and Phelps, the actual relationship between inflation and unemployment was quite different. In their minds, at least, the difference between their view of this relationship and the Keynesian view involved the way in which workers were assumed to form their expectations. In describing the difference between his view of this process and the view he attributes to Phillips, Friedman quotes Abraham Lincoln's famous assertion that "you can fool all of the people some of the time, you can fool some of the people all of the time, but you can't fool all of the people all of the time" (1976, 231). To Friedman, Phillips's analysis made sense only if workers could be fooled all the time—only, that is, if a given increase in the price level (beyond some unspecified base inflation rate) always fooled workers to exactly the same extent, regardless of how many times they had been fooled previously. Thus, persistent increases in the price level could hold the labor supply curve fixed in a location to the right of its no-surprises position, producing lower unemployment. Higher inflation rates, moreover, shifted the curve further than lower inflation rates and thus produced lower levels of unemployment.

Friedman and Phelps, in contrast, thought that while it might be possible to fool all the workers some of the time (temporarily), it was not possible to fool all of them all of the time (permanently). Eventually, workers would recognize that the base rate of inflation had increased, at which point the labor supply curve would begin to shift back and the increased inflation rate would gradually lose its power to reduce the unemployment rate. Further declines in unemployment could then be achieved, if at all, only by further increases in the rate of inflation. Thus, "the only way unemployment can be kept below the natural rate is by an ever-accelerating inflation, which always keeps current inflation ahead of anticipated inflation" (Friedman 1976, 227).

The view underlying this "acceleration hypothesis" is that while agents cannot be permanently fooled by inflation at a fixed rate, they can be fooled persistently, if not permanently, by accelerating inflation. One reason to be skeptical about this story is evidence from economies that have experienced hyperinflations (extremely rapid

increases in the aggregate price level): it is not unusual to see hyperinflation and high rates of unemployment go hand-in-hand. It should also be emphasized that nothing in this analysis suggests that any one-time increase in the price level must necessarily be followed by persistent inflation at a fixed rate that will eventually turn into accelerating inflation. The accelerating inflation described by Friedman and Phelps is created by design in order to surprise economic agents. It will not result from forces beyond the control of the policymakers, and it will not be produced by policymakers that implement a stable monetary policy—even if that policy involves a high money growth rate.

The NIRU (aka NAIRU): A Response to the Monetarists. Although the introduction to this article focused on the NAIRU, the analysis presented so far has concentrated on the Phillips curve. The reason for this attention is that the Phillips curve is a key element of the theory of the inflation-unemployment relationship that includes the NAIRU.

As the discussion has shown, during the 1960s Keynesian theorists came to regard the inverse (downward-sloping) empirical relationship between inflation and unemployment—the Phillips curve relationship—as a stable menu of options from which policymakers could choose. The apparent concreteness of this menu helped produce widespread confidence in the potential effectiveness of Keynesian-inspired countercyclical demand management. To Keynesians, the job of macroeconomists was to design demand-management policies that would strike the right balance between the competing problems of unemployment and inflation. Monetarists did not share the Keynesians' faith in the effectiveness of demand management, and during the 1960s and the 1970s there were fierce debates between the two schools. These debates sometimes took the form of disputes about the slope of the Phillips curve. Keynesians believed that the Phillips curve was quite flat, particularly at high unemployment rates. It followed that when unemployment was high, the unemployment rate could be reduced at little cost in terms of increased inflation. Monetarists, on the other hand, believed the curve was quite steep, so expansionary demand management was likely to produce a significant amount of inflation without providing much benefit in terms of lower unemployment. The monetarist challenge to Keynesian ideas about the Phillips curve culminated in the Friedman-Phelps hypothesis that the curve was vertical in the long run.

During the severe recession of 1974-75 both the inflation rate and the unemployment rate reached some of the highest levels in postwar U.S. history. This experience shook public faith in Keynesianism and played a key role in shaping the subsequent debate about inflation. The warnings of Milton Friedman and other monetarists that attempts to "ride the Phillips curve" might lead to accelerating inflation began to be heeded by more and more people, both inside and outside the ranks of professional economists. The credibility of the monetarist alternative to Keynesian theory was greatly strengthened.

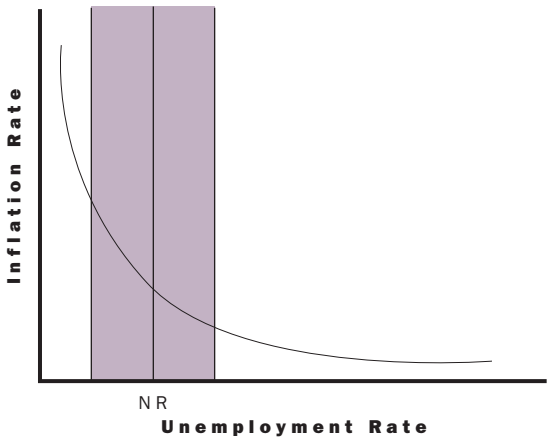
Despite the credibility gains of the monetarists, however, the events of the mid-1970s did not result in the demise of Keynesian macroeconomics or even of analysis based on the Phillips curve. Many economists continued to use the Phillips curve as the basis for forecasting and policy advice. As Okun recalls, "It was hard to cast aside a tool that had traced the United States record so well from 1954 through the late sixties. And it was easy to ignore the Friedman and Phelps attack on the stability of the short-run Phillips curve, and their prophetic warning (issued at a time when the Phillips curve was still performing admirably) that the curve would come unstuck in a prolonged period of excess demand. Unfortunately, most of the profession (including me) took too long to recognize that" (1980, 166).

Some Keynesians reacted to the events of 1974-75 by attempting to reinterpret the Phillips curve in a way that reconciled the Keynesian and monetarist views of the inflation-unemployment relation but preserved considerable scope for activist demand management. To do so was necessary to acknowledge that there might indeed be limits to the exploitability of the Phillips curve relation: in particular, attempts to use it to keep the unemployment rate below a threshold level might indeed result in accelerating inflation. As early as 1975, for example, Keynesians Franco Modigliani and Lucas Papademos asserted that "the existence of NIRU [the noninflationary rate of unemployment] is implied by both the 'vertical' and the 'nonvertical' schools of the Phillips curve" (1975, 142).[8]

What exactly was the NIRU? In the now-conventional Phillips curve diagram, which has the unemployment rate on the horizontal axis and the inflation rate on the vertical axis, the NIRU was the unemployment rate at which the Keynesians' downward-sloping Phillips curve intersected a vertical line at Friedman's natural rate of unemployment. Thus, the NIRU was equal to the natural rate. But while monetarists believed that the existence of a natural rate implied that there was no useful trade-off between inflation and unemployment, Modigliani and Papademos interpreted the NIRU as a constraint on the

8. The NIRU was later renamed the NAIRU, or nonaccelerating inflation rate of unemployment. This name makes it clear that sufficiently low unemployment rates are believed to be associated with accelerating inflation, not just higher fixed rates of inflation.

CHART 3
The Natural Rate



Keynesians believed that the economy spent most of its time in a range of unemployment rates well to the right of the natural rate and that unemployment rates to the left of the shaded area implied that inflation was likely to accelerate.

ability of policymakers to exploit a trade-off that remained both available and helpful in the short run.

Perhaps the most striking thing about the Modigliani-Papademos argument is that while it incorporated many aspects of Friedman's critique of Keynesian theory, it stood Friedman's principal policy recommendation on its head: Friedman was strongly opposed to activist monetary policy. One of the reasons that was possible was that most expositions of the monetarist view of the inflation-unemployment relationship—including Friedman's—did not seem to resolve the question of the strength or persistence of the short-run effects of monetary policy. After all, Friedman's inflation-acceleration theory did seem to suggest that monetary policy could produce temporary reductions in the level of unemployment—but these reductions could be sustained only at the price of continually increasing inflation rates.

The remaining difference between the Keynesians and the monetarists was actually quite fundamental: it involved the direction of the causal relationship between inflation and unemployment. This difference continued to allow members of the two schools to hold contrasting views about the sensitivity of the unemployment rate to changes in the inflation rate (or vice versa) and thus about whether the short-run Phillips curve trade-off was potentially useful to policymakers.

Monetarists saw the level of unemployment as determined largely through the process of labor market clearing. The economy, in their view, was never far from the full-employment equilibrium of the classical model. Monetarists believed that monetary policy had a direct and powerful influence on the price level and the inflation rate. While the channels through which it obtained this influence might involve the goods and labor markets, these markets adjusted and cleared so quickly that policy changes had little effect on them. In particular, monetary policy could affect the level of unemployment only marginally and only by producing inflation surprises whose impact would decrease rapidly over time. Since unexpected changes in the inflation rate could produce only small changes in the level of unemployment, the Phillips curve was quite steep even in the short run. The rate of unemployment could never stray far from the natural rate, and continued efforts to keep it below the natural rate would result mostly in accelerating inflation.

Keynesians, on the other hand, continued to believe that the economy could and often did operate at "equilibrium" positions in which aggregate demand was deficient—positions in which there was massive excess supply of labor and large-scale involuntary unemployment. The level of unemployment, they believed, determined the rate of inflation by determining the growth rate of nominal wages (see above). Thus, changes in unemployment caused changes in inflation, rather than the reverse.

It was their belief that the level of aggregate demand could be and often was deficient that allowed Keynesians to believe that policies that influenced its level could play an important role in determining the current level of employment. As long as there was "slack" (unemployed labor and other resources) in the economy, monetary ease, for example, would not start a wage-price spiral because the initial round of goods-price increases it produced (see above) would not place substantial upward pressure on nominal wage rates. Thus, Keynesians believed that the economy spent most of its time in a range of unemployment rates well to the right of the natural rate/NIRU—a range within which the Phillips curve was very flat. If demand stimulus pushed the unemployment rate too low, however, then labor market tightness would put persistent upward pressure on the inflation rate. This was the range where the short-run Phillips curve was steep; it was also the range within which the long-run increases in the inflation rate predicted by Friedman were a serious potential problem. Thus Modigliani and Papademos wrote that "unemployment rates left of the shaded area [the area displaying the current range of NIRU estimates] imply a high probability that inflation will accelerate" (1975, 147) (see Chart 3).

Despite the fundamental differences between the monetarists and the Keynesians, the NIRU was seen by many contemporary economists as helping build a consensus about the nature of the inflation-unemployment relationship. According to James Tobin, the "consensus macroeconomic framework, vintage 1970" held that "the nonagricultural business sector plays a key role in determining the economy's rate of inflation. . . . According to the standard 'augmented Phillips curve' view, rates of

price and wage increase depend partly on their recent trends, partly on expectations of their future movements, and partly on the tightness . . . of markets for products and labor. Variations in aggregate monetary demand, whether the consequences of policies or other events, affect the course of prices and output, and wages and employment, by altering the tightness of labor and product markets, and in no other way. . . . Inflation accelerates at high employment rates because tight markets systematically and repeatedly generate wage and price increases. . . . At the Phelps-Friedman 'natural rate of unemployment,' the degrees of resource utilization and market tightness generate no net wage and price pressure up or down and are consistent with accustomed and expected paths, whether stable process or any other inflation rate. The consensus view accepted the notion of a nonaccelerating inflation rate of unemployment (NAIRU) as a practical constraint on policy" (1980, 23).[9]

Most current descriptions of the Phillips curve relationship and the NAIRU are not very different from Tobin's description. One difference is that most modern descriptions see changes in monetary policy as the principal source of changes in the economy's aggregate demand—a view that Tobin ascribes to the monetarists. Otherwise, the accounts are similar to Tobin's in the sense of asserting (1) that the current stance of [monetary] policy can be determined by looking at the unemployment rate and comparing it with its natural rate and (2) that the current level of the unemployment rate provides a good indication of the direction and strength of future changes in the inflation rate: low unemployment indicates that the rate of inflation will increase in the short run and accelerate in the long run.

As Tobin pointed out, the macroeconomic consensus about the nature of the inflation-unemployment relationship did not extend to the question of whether policymakers could or should exploit that relationship. In a recent column in the *Wall Street Journal*, Friedman recounts: "I introduced the concept of the natural rate in 1968 as part of an article on 'The Role of Monetary Policy.' . . . The natural rate is a concept that does have a numerical counterpart—but that counterpart is not easy to estimate and will depend on particular circumstances of time and place. More important, an accurate estimate is not necessary for proper monetary policy. I introduced the concept in a section titled 'What Monetary Policy Cannot Do.' It was part of an explanation of why, in my opinion, the monetary authority cannot adopt 'a target for employment or unemployment . . . ; be tight when unemployment is less than the target; be easy when unemployment is higher than the target'" (*WSJ*, September 24, 1996).

To summarize, the NAIRU was born out of an attempt by proponents of the Phillips curve to address the monetarist critique of policy prescriptions based on the curve. In the minds of many Keynesians, the NAIRU theory successfully reformulated the natural rate hypothesis as a relatively minor qualification of Keynesian theories about the usefulness of the Phillips curve as a guide to monetary (or fiscal) policy. From the monetarist perspective, however, the NAIRU was simply another name for the natural rate. The NAIRU theory, moreover, was based on a fundamental misunderstanding of the natural rate hypothesis—a hypothesis that demonstrated the ineffectiveness of government demand-management policy.

There is a sense in which it is hard to blame the NAIRU proponents for ignoring monetarist assertions that monetary policy was inherently neutral. After all, monetarists such as Friedman had long argued that activist monetary policy was in fact the principle source of short-run economic fluctuations. The monetarists gave further ground to the Keynesians by maintaining a distinction between the short run and the long run and by speaking of money illusion as a channel that gave policymakers access to a short-run inflation-unemployment trade-off. To the dismay of the leading monetarists, proponents of the NAIRU were quite successful in capitalizing on its appeal as a simple, intuitive guide for giving policy advice.

Some Questions about the NAIRU

There are additional problems with using the NAIRU concept to formulate policy rules that are not directly connected to the Keynesian-monetarist debate. One of these involves the relationship between changes in relative prices and changes in the aggregate price level. Relative price changes signal degrees of relative scarcity in the economy: they reveal how highly the economy values different goods and services and are often associated with changes in the quantities of those goods and services produced or employed. One very important relative price is the real (or relative) wage, which is the purchasing power of the nominal wage in terms of goods and services and can be loosely defined as the average nominal wage divided by the average price level. Changes in real wages reflect changes in the value of labor services relative to the value of other goods and services. They are often associated with changes in the level of employment.

Both conventional monetarist theory and the Keynesian/monetarist synthesis of the 1970s predict that the mechanism by which monetary policy creates inflation involves repeated increases in both nominal and real wages and temporary decreases in the rate of

9. Tobin seems to have been the first writer to actually use the term NAIRU; *recall that Modigliani and Papademos used the acronym NIRU (noninflationary rate of unemployment).*

unemployment. According to these theories, the changes in nominal wages caused by monetary policy do not result in permanent changes in the level of real wages because the price level eventually adjusts to offset the nominal wage changes.

These macroeconomic theories are often "inverted" to produce rules for conducting monetary policy that are based on current levels of unemployment or current rates of change in nominal wages. The simplest rule of this type is that when the unemployment rate is lower than the NAIRU, monetary policy has become too "easy" and should be tightened to head off the coming inflationary spiral. Unfortunately (or perhaps fortunately), the fact that there is little agreement on the precise value of the NAIRU makes this rule hard to implement. An alternative rule that does not suffer from this problem is to tighten policy whenever nominal wages begin growing more rapidly than prices so that real wages begin to rise.

Friedman and Phelps argued that unexpected inflation can drive the level of unemployment below the natural rate, but only temporarily; in the long run there will be no inflation-unemployment trade-off.

As the introduction to this article noted, one of the reasons that the NAIRU has attracted so much attention recently is that the level of unemployment has been low—lower than many NAIRU estimates. As a result, some economists have called on monetary policymakers to move to tighten policy, and others have suggested that they begin watching nominal wage changes closely and tighten policy as soon as there is any sign that real wages are rising.

Non-Policy-Induced Changes in Real Wages. Do policy rules of this sort make sense, even if we accept the underlying theory of the effects of monetary policy? One fundamental problem with these rules is that they implicitly assume that any change in labor market conditions that produces lower unemployment or higher real wage rates must have resulted from monetary policy.[10] Of course, virtually every economist acknowledges that changes in labor productivity produce persistent increases in the relative price of labor, thus causing real wage rates to change independently of monetary policy changes. Consequently, it is often suggested that policymakers should respond only to increases in nominal wages that result in real wage increases that cannot be attributed to gains in productivity.

Is it reasonable to assume that every increase in real wages that cannot be directly linked to an increase in productivity has been caused by a change in monetary policy and will eventually be followed by an increase in inflation? Two big problems with this assumption are that labor productivity is notoriously difficult to measure and that productivity data become available only after a considerable time lag. However, even if labor productivity could be measured in a timely and accurate manner, it would not follow that increases in wages that were not associated with productivity gains were necessarily caused by monetary policy. Not all changes in the demand for goods and services come from changes in monetary policy, or even government policy, and some of these changes may affect both the relative price of U.S. labor (that is, the real wage) and the level of U.S. employment. Examples include changes in foreign demand for U.S. exports—particularly exports of goods that are labor- or human capital-intensive—or changes in domestic tastes favoring goods of the same type.

As we shall see, if wages and prices are perfectly flexible then changes in relative prices—including relative wages—should have no effect on the aggregate price level. If wages or prices are sticky, however, then relative wage or price changes may appear to produce aggregate price level changes and may even appear to produce persistent inflation.

Why is the possibility of non-policy-induced changes in the relative price of labor important? Most economists would agree that it makes sense to use monetary policy to resist real wage or unemployment rate changes if these changes are simply a lane on the road to a permanent increase in the rate of inflation. Most economists would also agree that policymakers should not resist real wage or unemployment rate changes that are associated with permanent (or persistent) increases in the relative price of labor—even if these changes appear to produce temporary increases in the inflation rate. Resisting changes of this sort would risk letting monetary policy interfere with the important job of relative price changes, which is to ensure that inputs and outputs continue to be used and produced efficiently.

Unfortunately, it is not always easy to distinguish temporary changes in the inflation rate from permanent ones. As a result, the fact that there may have been many occasions in the past when increases in the relative price of labor produced temporary increases in the inflation rate may reinforce some economists' present tendency to advocate tightening in response to current increases in nominal and real wages. Thus, real wage changes that are not caused by policy-induced changes in aggregate demand can create a great deal of confusion for policymakers who are trying to use wage growth rates or unemployment rates as guides to monetary policy.

Relative Price Changes and the Aggregate Price Level. The most common method for measuring changes in the aggregate price level involves taking a fixed basket of goods and determining how the money cost of that bas-

ket has changed over time. The price index produced by this method is equal to the market value of the basket at a particular point in time divided by the market value of the same basket in a fixed base year—typically, the year in which the basket was chosen. The consumer price index (CPI), which is the most closely watched price index, is constructed in this manner. The following example illustrates the impact of a relative price change—a change in the price of a single good relative to the prices of other goods—on a price index like the CPI.

Imagine a household that consumes (1) directly provided labor services (for example, cleaning services), (2) the services of durable goods (such as personal computers), and (3) food (bread). Now suppose that the demand for directly provided labor services increases—perhaps because foreign tourism in the United States has increased and hotels and condo owners are hiring people to clean the rooms and condos foreign tourists have rented. Standard microeconomic theory predicts that this increase in demand will lead to an increase in the price of these services. Assume, for the moment, that the prices of the two other classes of consumption goods do not change (an assumption that will have to be abandoned later). Thus, both the absolute and relative prices of direct labor services have increased.

How will a change in the price of direct labor services—a relative price change—affect the CPI, which is a measure of the overall price level? The fixed-market-basket method for constructing the CPI amounts to assigning different fixed weights to the prices of the different items in the basket. For the purposes of this example, assume that cleaning services, personal computer services, and bread are the only items in the basket and that their initial prices are \$10 per hour for cleaning services, \$10 for computer services, and \$10 per loaf of bread. Also assume that a typical individual allocates 10 percent of his or her spending to cleaning services, another 10 percent to computer services, and the remaining 80 percent to buying bread. Finally, assume that the initial prices of these items are the same as the ones from the base year. We can then construct the initial value of our hypothetical price index:

$$CPI_{initial} = \frac{0.1(\$10) + 0.1(\$10) + 0.8(\$10)}{0.1(\$10) + 0.1(\$10) + 0.8(\$10)} = \frac{\$10}{\$10} = 1.00.$$

Now suppose that the price of cleaning services doubles but the other two prices remain unchanged. The CPI would then be

$$CPI = \frac{0.1(\$20) + 0.1(\$10) + 0.8(\$10)}{0.1(\$10) + 0.1(\$10) + 0.8(\$10)} = \frac{\$11}{\$10} = 1.10.$$

In this case, the reported inflation rate would be 10 percent.

An important question, however, is whether it is really reasonable to hold the weights of the three goods/services fixed in light of the large increase in the price of one of them. From elementary microeconomics, we know that the price increase is likely to produce a "substitution effect" on spending: people will respond to the relative price increase by substituting out of market-delivered cleaning services, either by accepting slightly messier homes or doing more cleaning themselves. They may also buy additional durable goods (such as carpet-cleaning machines) to help them do their own cleaning. With this likelihood in mind, and ignoring for the moment the possibility of further adjustment in relative prices, let's imagine the effects of allowing the quantity weights to adjust. Assume that U.S. households change their spending patterns so that they purchase fewer hours of cleaning services (labor), whose weight falls from 0.1 to 0.05, and more durables services, whose weight rises from 0.1 to 0.15). (Note that the government agency that constructs the actual CPI does not make these kinds of adjustments, except quite infrequently—see below.) Our "revised" June CPI would look like this:

$$CPI_{(rev)} = \frac{0.05(\$20) + 0.15(\$10) + 0.8(\$10)}{0.1(\$10) + 0.1(\$10) + 0.8(\$10)} = \frac{\$10.50}{\$10} = 1.05,$$

in which case the rate of inflation would now be only be only 5 percent.

Clearly, the increase in the value of the price index—that is, in the aggregate price level—is smaller when the quantity weights are allowed to adjust to changes in expenditure patterns. In other words, the substitution effect acts to restrain the "inflationary" effects of relative price increases.

10. One noteworthy aspect of Friedman's explanation of the Phillips curve mechanism was that he was as willing as most other economists to accept the notion that increases in wage rates were essentially equivalent to increases in the price level. In Friedman's words: "Fisher talked about price changes, Phillips about wage changes, but I believe that for our purposes that is not an important distinction. Both Fisher and Phillips took for granted that wages are a major component of total cost and that prices and wages would tend to move together. So both of them tended to go very readily from rates of wage change to rates of price change, and I shall do as well" (1976, 218). Of course, Friedman may have taken this approach not because he agreed with the assumption that all wage-rate changes necessarily produce proportional price level changes but because he was able to make his point about the natural rate of unemployment without worrying about this distinction.

A second effect of relative price changes on the price level is the often-overlooked "income effect." An increase in the absolute (dollar) price of cleaning services reduces households' purchasing power: they are no longer able to afford the quantities of the three goods that they were purchasing initially. This loss of purchasing power will typically cause them to reduce their purchases of all goods—even goods that are not closely related to the goods whose prices have changed. In our example, households are slightly poorer because of the increased price of cleaning services, and they react by reducing their purchases of bread. Bakers may be forced to respond by reducing the price they charge for bread, which we will assume falls to $9.38 per loaf.

What does the newly revised CPI look like after accounting for the income effect?

$$CPI_{(rev^2)} = \frac{0.05(\$20) + 0.15(\$10) + 0.8(\$9.38)}{0.1(\$10) + 0.1(\$10) + 0.8(\$10)}$$
$$= \frac{\$10}{\$10} = 1.00.$$

Thus, after the substitution and income effects have worked their way through the economy, the increase in the price level caused by an increase in a relative price—in this case, something like a relative wage—is zero.

Unfortunately, the CPI as currently calculated does not capture the substitution effect in a timely fashion: while the quantity weights are periodically changed to reflect changes in spending patterns, this revision happens only once every five years. Income-effect-induced price changes will be captured as soon as they occur, but these often take a long time to work their way through the economy. It may take households some time to realize that their real income has decreased and some additional time to adjust to the decrease; until they do adjust, they may dig into their savings to finance higher-than-normal expenditures. Consequently, relative price or wage increases may produce increases in the measured price level in both the short run and the medium run, even though they may have no long-run price level effects once the income and substitution effects work their way through the economy.

Menu Costs. Ball and Mankiw (1995) have developed a theory that provides a more detailed and specific explanation of the process by which increases in relative prices produce temporary increases in the aggregate price level. Their key postulate is that there are "menu costs"—costs of changing prices—that prevent nominal prices from being fully flexible. Suppose, for example, that veal is a key ingredient in many of the items on a restaurant's menu and that its market price has gone up by a small amount. The restaurant owner is consequently faced with an uncomfortable choice: increase the prices of veal-based dishes to reflect the new veal price, which will require an expensive reprinting of all the menus in the restaurant, or simply absorb the price increase.

Changing announced prices may be costly for many firms other than restaurants. The Ball-Mankiw theory predicts that these costs will produce a "range of inaction"—a range of input-price increases small enough that they will not cause producers to increase the prices of outputs. They explain that "when a firm experiences a shock to its desired relative price, it changes its actual price only if the desired adjustment is large enough to warrant paying the menu cost. . . . In this setting, shifts in relative prices *can* affect the price level" (1995, 162). To understand the latter point, imagine a no-menu-cost situation in which the prices of a small number of goods rise substantially but the aggregate price level does not rise because the income effect of these price increases reduces the demand for a large number of other goods and causes their prices to decline slightly. When there are menu costs, however, it may not pay the producers of these other goods to cut their prices in response to small demand decreases. As a result, there may not be a large number of small price decreases to offset the small number of large price increases, and the aggregate price level may rise.

The Ball-Mankiw theory can help explain how a one-time increase in real wage rates or other relative prices can produce a temporary increase in the aggregate price level (as measured by a price index) and how repeated increases in real wages or relative prices can produce a temporary increase in the inflation rate.[11] As the authors note, this explanation presumes that the relative price increases are concentrated in particular industries, and thus require large price adjustments, while the resulting income-effect-driven demand decreases are spread across many different industries and consequently require relatively small adjustments. As applied to wage rates, the theory predicts that the increases in real wages that are most likely to result in temporary increases in inflation are increases that are concentrated, at least initially, among workers in particular industries. These wage increases will produce cost increases in these industries that exceed their ranges of inaction and will consequently impel the industries to increase their product prices substantially.

Price Stickiness: The Empirical Evidence. Menu costs are one possible example of a "nominal rigidity"—a source of friction that prevents money prices from adjusting in the perfectly flexible manner assumed by classical theory. Much of Keynesian theory, including the theory behind the NAIRU, is based on the assumption that the economy is afflicted by many other price rigidities of this general type. As a result, one natural strategy for convincing skeptics of the validity of the theory would be to describe the nature and source of these rigidities as pre-

cisely as possible. It would be helpful, for example, to be able to identify the rigidities that are severe enough to prevent nominal wages from adjusting to eliminate a persistent excess supply of labor—the rigidities, that is, that allegedly permit persistent involuntary unemployment. Similarly, it would be helpful to be able to identify the frictions that allegedly make firms slow to adjust their prices to increases in demand and workers slow to adjust their wage demands to increases in prices. This information would make it much easier for skeptics to understand how aggregate demand stimulus could produce significant (if temporary) increases in output and employment.

Given the wealth of NAIRU-based advice that is currently being offered to policymakers, it may seem reasonable to infer that there is plenty of good evidence supporting the claim that nominal rigidities are widespread and substantial. In reality this is not at all the case. In a recent paper, Wynne (1995) reports the results of a systematic search for empirical studies documenting price stickiness. Despite the widespread acceptance of theories based on sticky prices, he was able to find only a small number of studies, including only three that used data from the post-World War II period. Wynne also points out that these studies would not stand up well against some elementary objections to their methodology. For example, the goods and services whose prices are examined in these studies account for a very small fraction of GDP; they also include, in many cases, goods whose prices are known a priori to be relatively inflexible or which "exhibit little or no quality changes over time." Wynne goes on to point out that "many hi-tech products have remarkably flexible prices" (1995, 7).

What about the assumption that is widely considered absolutely fundamental to Keynesianism—the assumption that nominal wages are sticky downward? Zarnowitz notes that "the average annual money earnings from wages declined in about half of the business contractions of 1860-1914 and in all of those of 1920-38, according to the data compiled in Phelps Brown 1968. . . . In contrast, they kept rising through the period 1945-60, which witnessed four moderate or mild recessions. . . . Data for 1889-1914 from Rees 1961 show that peaks and troughs in annual earnings matched nearly two thirds of the like business cycle turns of the period, but those in hourly earnings fewer than half. . . . The conclusion is that most of the major business downturns and some of the minor ones have historically been associated with declines in nominal wage earnings" (1992, 146).

The NAIRU's Empirical Record. As Okun (1980) explains in a passage quoted above, for roughly fifteen years ending in the late 1960s U.S. inflation and unemployment data seemed to line up along a stable Phillips curve. The stagflation of the 1970s destroyed this empirical relationship. During the last twenty years, econometricians have not had much success identifying a stable, reliable relationship between inflation and unemployment.

Of course, econometricians' inability to construct an empirically reliable Phillips curve makes it impossible for them to produce a reliable estimate of the NAIRU. Recently this problem has become a serious one for economists who think monetary policy should be based on the NAIRU. During the past two years, for example, the U.S. unemployment rate has been quite low—lower than many widely publicized NAIRU estimates. However, the inflation rate has shown no signs of increasing (to say nothing of accelerating) in the way the NAIRU theory predicts. As Fred Bleakley reports in a recent article in the *Wall Street Journal* (February 1996), the failure of relatively low unemployment rates to produce higher inflation rates has led several prominent economists to revise their estimates of the NAIRU downward. Ex post revisions of this sort are probably very frustrating for policymakers who are seeking a reliable guidepost for monetary policy.

Despite the credibility gains of the monetarists, the events of the mid-1970s did not result in the demise of Keynesian macroeconomics or even of analysis based on the Phillips curve.

As frustrating as the current situation may be, economists and policymakers definitely prefer it to the 1970s, when inflation rates and unemployment rates were high simultaneously rather than low simultaneously. By the end of the decade even inveterate Keynesians had begun to lose faith in the usefulness of the NAIRU concept. At the close of the 1970s, Tobin warned that "as for the shape of the short-run trade-off [between inflation and unemployment], Murphy's Law of macroeconomics assures us that it is an *L* with the corner wherever it happens to be. . . . It is possible that there is no NAIRU, no natural rate, except one that floats around with actual

11. The theory does not imply that changes in relative prices can produce permanent price level increases. If the restaurant owner believes that the change in the price level is permanent then it will make sense for him to revise his menu immediately since he will have to revise the menu eventually and the longer he waits the greater his losses will be. Similarly, if relative prices rise gradually over a period of time then the theory predicts that the inflation rate may increase during the same period of time but not that the inflation rate will increase permanently.

history. It is just as possible that the direction the economy is moving is at least as important a determinant of acceleration and deceleration as its level. These possibilities should give policymakers pause as they embark on yet another application of the orthodox demand-management cure for inflation" (1980, 61-62).[12]

Neoclassical Macroeconomics

The 1970s: Theory and Evidence Collide. Economically, the decade of the 1970s was dominated by major "supply shocks"—principally, the OPEC oil embargo and the resulting increases in world oil prices. Supply shocks were not easily incorporated into Keynesian theory. Historically, Keynesian theorists had concentrated on studying the effects of changes in aggregate demand and had implicitly assumed the existence of a stable aggregate supply schedule. As a result, the supply shocks of the 1970s caused forecasts based on Keynesian predictions to generate huge errors. As Tobin pointed out, "the inflationary components of the expansions, 1971-73 and 1975-79, were unexpectedly and distressingly large. The disinflationary consequence of the first contraction, 1969-71, was distressingly small. Indeed, money wages 'exploded' while unemployment was rising. . . . The major economic events of the decade were the extraordinary changes in world supplies and prices of specific commodities. Their interaction with macroeconomic indicators and events confronted both policymakers and analysts with problems for which they were unprepared. . . . No one foresaw in 1970 the main economic events of the decade or the formidable challenges those surprises would pose for macroeconomics and stabilization policy. We macroeconomists were caught unawares. It was not simply that our models, theoretical and econometric, now had to be applied to novel situations. Worse than that, the shocks of the 1970s required some fundamental rethinking and rebuilding" (1980, 21-23).

From the monetarist perspective, the NAIRU theory was based on a fundamental misunderstanding of the natural rate hypothesis—a hypothesis that demonstrated the ineffectiveness of government demand-management policy.

Although Tobin acknowledged that Keynesian theory faced problems, he was not at all ready to abandon the Keynesian ship. In his view, the "consensus model" of the early 1970s was in need of extension and refinement rather than replacement. As noted earlier, however, the high inflation rates of 1974-75 pushed many other economists in the direction of the monetarists.

In retrospect, it is clear that the record of 1974-75 posed big problems for both Keynesians and monetarists. While Keynesians could try to explain the high unemployment as a consequence of insufficiently aggressive management of aggregate demand, they could not explain how the inflation rate had become so high when the labor market was clearly the opposite of tight. Monetarists, on the other hand, could blame accelerating inflation on overly aggressive demand management but could not explain how a too-expansionary policy could have produced such high unemployment. To make matters worse, monetarism held that recessions were almost always caused by monetary tightening (see Friedman and Schwartz 1963), but if a major tightening had occurred then the inflation rate should have fallen.

A New (and Old) Approach to Macroeconomics. The inability of Keynesian and monetarists theories to explain the key macroeconomic events of the 1970s caused these theories to become discredited in the minds of many economists. This widespread disenchantment with traditional macroeconomic theory left the field open for a group of young economists who were attempting to develop a new approach to macroeconomics on the foundation provided by the classical paradigm. The research program of these economists came to be known as neoclassical economics.[13]

The neoclassical attempt to build on classical principles involves formalizing many of the concepts that have been used informally by classical and monetarist economists. Neoclassical economics is based on the classical assumption that individual households and firms make the decisions that maximize their well-being subject to their budget and technological constraints. Neoclassical economists extend this assumption to intertemporal decisions—an extension that forces them to study the interaction between current choices and future choices and to attempt to trace out the consequences of these choices over time. They prefer to conduct these investigations in general equilibrium settings—that is, in formal models that try to take into account the complex and often simultaneous interactions among different economic variables in both the short run and the long run.

A key principle of neoclassical economics is that in order to determine the economic impact of a hypothetical change in government policy—a tax cut or an increase in the money supply growth rate, for example—it is necessary to consider the possibility that individual households and firms may react to government policy changes by changing the ways in which they make their own economic decisions. Neoclassical economists' effort to describe the nature of these changes in individual "decision rules" focuses on the manner in which the individuals formulate their economic expectations. More specifically, a fundamental and formative assumption of

neoclassical economic theory is that the economic expectations of households and firms are formulated in the most accurate possible manner, given the information available to them—including information about changes in government policy. This assumption is known as rational expectations.

As we have seen, the question of how workers formed their expectations about future prices became a key issue in the debate between the Keynesians and monetarists over the inflation-unemployment relationship. The analysis used by the original Keynesians did not include any formal description of the way expectations of this sort were formed. The expectational assumption behind the Friedman-Phelps natural rate hypothesis—a hypothesis that was (as we have also seen) partially incorporated into early-1970s Keynesianism—was "adaptive expectations." Adaptive expectations is the assumption that people base their expectations about the future values of economic variables on the past values of these variables, emphasizing values from the recent past. In the case of inflation, one specific adaptive expectations assumption that was commonly used in econometric studies was that next year's rate of inflation was expected to be equal to a weighted average of the values of past inflation rates, with the weight of a particular past inflation rate declining as it receded further into history. As we note below, because adaptive expectational assumptions do not take into account the systematic changes in ways the public forms its expectations that may occur when the government changes policy, results obtained using them will be very different from those obtained using the assumption of rational expectations.

The following two examples illustrate the potential impact of rational expectations on the effects of government policy. First, imagine that the Smith family is considering buying a house in a particular neighborhood. The family wants to make sure the house will bring a good price if they have to sell it in the future. The Smiths will probably use the price information from recent sales of comparable homes to estimate the future resale price of the home they are looking at. Suppose, however, that the Smiths learn that the government has decided to build an interstate highway extension that will, when completed, come within a thousand feet of their prospective home. Will they take this change in government policy into consideration when estimating the future sale price of the home, or will they continue to concentrate exclusively on past sale price information?

For a second example, imagine that during a mild recession the government decides to try to stimulate the economy by giving temporary tax breaks to families who buy new homes. Suppose, for the sake of argument, that this policy really does succeed in influencing potential home buyers and that the economy actually improves as a result. Now suppose that the government, emboldened by the apparent success of its new policy, makes the decision to use it to combat future recessions. What will happen the next time the economy begins to slow down? Will people remember the tax break that was offered during the previous recession and decide to hold back on their new-home purchases until the government decides to offer another tax break? If they do, then the recession may come sooner and be more severe than it would have been otherwise, and the effects of the tax break policy will be almost exactly the opposite of what the government intended.

Real wage changes that are not caused by policy-induced changes in aggregate demand can confuse policymakers who are trying to use wage growth rates or unemployment rates as guides to monetary policy.

These examples illustrate two important things about the ways in which rational, forward-looking individuals are likely to respond to changes in government policy. First, in projecting the consequences of their economic decisions individuals are likely to consider not only the consequences of similar past decisions but also all the other relevant information that may be available—including information about the effects of government policies. When it comes to predicting inflation, for example, people will not look exclusively at inflation rates from the recent past, as adaptive expectations assumed. Instead, they will also try to make use of any information available to them about the motives and behavior of monetary policymakers. Second, just as people will try to learn from the results of their own past decisions, they will also try to learn from their past observations about the effects of government policy. In particular, people will try to distinguish unsystematic variation in government behavior from systematic changes in government policy. Suppose, for example, that people discover that every time the unemployment rate is above a certain percentage, monetary policymakers react by increasing the money supply in an effort to reduce the rate of unemployment. It will not be long

12. For a closer look at the question of the estimation and empirical usefulness of the NAIRU, see Staiger, Stock, and Watson (1997) and Chang (1997).

13. For summaries of some of the innovations this research program produced, see Lucas and Sargent (1979) and Miller (1995).

before both employers and employees begin to take into consideration the effects of this policy in their wage and salary negotiations. If the unemployment rate is above the threshold percentage at the time of the negotiations, then the wage and salary levels that emerge from the negotiations may include upward adjustments for expected price increases. As a result, the final negotiated salary may be the same, in real terms, as it would have been if the government had not acted, and the government's actions may not end up having any effect on the level of employment.

How did neoclassical theory view the Phillips curve? To neoclassical economists the Friedman-Phelps critique of Keynesian notions about the effects of monetary policy was a step in the right direction, but only a rather tentative step. As we have seen, Friedman and Phelps forced Keynesians to accept the natural rate as a long-run constraint on demand-management policy but did not succeed in suppressing their belief in the existence and exploitability of a short-run Phillips curve relationship. Neoclassical economists, however, argued that even if a statistical relationship between inflation and unemployment did exist in the short run, it might be impossible for the government to exploit the relationship because people might respond to government demand-management policy in ways that would frustrate the goals of the policy.

Neoclassical economics is based on the classical assumption that individual households and firms make the decisions that maximize their well-being subject to their budget and technological constraints.

The first economist to make this point was Robert Lucas, who is generally regarded as the founder of the neoclassical school. The formal model Lucas (1972) developed and analyzed had three basic features that have become characteristic of neoclassical macroeconomic theory. First, the model integrated microeconomics and macroeconomics by studying the impact of the decisions of individual households and firms on the values of economic aggregates. Second, the model was dynamic—that is, it took intertemporal considerations into account, including the expectations of households and firms. Third, the model was stochastic—that is, it accounted for the fact that many decisions had to be made under uncertain circumstances and that the decisions of the households and firms played a role in determining the nature of this uncertainty.

In Lucas's model, individuals are "farmers" who simultaneously provide labor, produce goods, and consume goods. These individuals face fluctuations in prices that are caused partly by changes in "real" economic conditions—good or bad crops—and partly by unsystematic changes in monetary policy. The latter take the form of random deviations from a systematic path of the money supply. Each period, the change in the price of any particular good is caused partly by a change in real economic conditions and partly by a change in monetary policy.

Individuals would like to respond differently to price fluctuations that come from different sources. If the relative prices of the particular goods they produce increase, then they want to work harder and increase their production of these goods, for standard microeconomic reasons. If, however, the increase in the price of the good a particular individual produces is simply part of an increase in the overall price level (that is, in absolute prices)—so that the relative price of this good has not changed—then there is no reason for that individual to increase his production or work effort. Thus, if individuals could distinguish relative price changes from absolute price changes with 100 percent accuracy, then they would never increase their work effort in response to absolute price changes. As a result the Phillips curve for this economy would be vertical, even in the short run.

In Lucas's model, as in most real-life situations, individuals do not possess complete information about the current state of the economy. In particular, individuals are assumed to be unable to observe the current prices of any goods other than the goods they produce. Consequently they cannot tell for certain whether changes in the prices of "their goods" represent absolute or relative price changes. However, individuals do know the statistical properties of the two different types of price fluctuations. They can use this information to calculate the average part of each price change that represents a relative price movement and then respond only to that part of the price change.[14] This is the key place where the assumption of "rational expectations" is used in the model.

Now suppose that, during a particular period, relative prices happen to remain entirely unchanged because there have been no changes in real economic conditions. At the same time, the absolute price level rises by a larger-than-normal amount because there has been a larger-than-normal increase in the money supply. Individuals will have no way of knowing that this particular price change is all absolute; consequently, they will proceed under the assumption that some part of it represents a relative price change. As a result, they will increase their work effort in response to the price increase. The larger the absolute price increase, moreover, the larger their work-effort increase will be. Thus, monetary-policy-induced changes in the price level will have real effects of a type consistent with a Keynesian-looking short-run Phillips curve.[15]

Can monetary policymakers use this short-run Phillips curve to increase the levels of employment and output? Suppose that in an effort to do so they increase the average money growth rate by some fixed percentage. If individuals are aware of this change in policy they will realize that prices are now going to increase at a higher average rate. As a result, the fact that the price level increases at a higher rate next period or in subsequent periods will not surprise or confuse them, and the policy-induced increase in the inflation rate will have no effect on work effort. People will still respond to unsystematic price level changes in the same way they did previously, but they will now expect a higher average rate of inflation. The statistical Phillips curve will shift up by the amount of the increase in the average inflation rate, but the Phillips curve facing policymakers will be vertical.

Lucas's 1972 paper had a tremendous impact on the economics profession: it is arguably the most influential single contribution by a macroeconomist in the last fifty years. There are two basic reasons for its significance. The first reason, already noted, is that the paper represented a huge methodological advance in macroeconomic theory, combining as it did general equilibrium theory, dynamic analysis, and rational expectations.[16] The second reason is that he provided a qualitative explanation for two phenomena that were both puzzling and troubling to macroeconomists—the fact that the seemingly reliable Phillips curve of the 1950s and 1960s had begun shifting upward erratically at just about the time that policymakers began to try to use it to guide monetary and fiscal policy and the (closely related) fact that deliberate changes in monetary and fiscal policy did not seem to be having the effects on employment and output that were predicted by Keynesian theory.

What does Lucas's theory predict about the natural rate and the NAIRU? In his model, systematic changes in monetary policy have no effect on the level of employment, and the labor market does not play any special role in the mechanism by which a monetary expansion produces inflation. As a result, in the context of the model it would not make sense for the government to focus on unemployment rates or wage changes as guides for monetary policy.

Lucas's paper also makes two broader points whose potential applicability extends far beyond the specific features of his model. The first point, discussed earlier, is that theories of the effects of government policy that are based on the assumption that people make systematic forecasting errors are not very sensible: since people have strong economic incentives to correct such errors, the changes in their behavior induced by changes in policy are likely to disappear very quickly as they revise their forecasting schemes. This point had already been made by Friedman and Phelps, but Lucas's analysis reinforced it in an exceptionally stark and rigorous way. The second point, which was an entirely new contribution, is that the existence of statistical relationships between variables of interest to policymakers is no guarantee that these relationships can be exploited by policymakers, regardless of how reliable the relationships may seem to be. In Lucas's model, the Phillips curve is, by construction, a very reliable statistical relationship—a relationship in which the levels of employment and output fluctuate around long-run averages that can be thought of as the analogues of the natural rate or NAIRU. However, the short-run component of the Phillips curve relationship, which is the only

A fundamental assumption of neoclassical economic theory is that the economic expectations of households and firms are formulated in the most accurate possible manner, given the information available to them.

14. For example, individuals may know that, on average, one-third of the increase in the price of a good represents an increase in the relative price of that good while two-thirds represents an increase in the absolute price level. In this case, if individuals observe that the price of their good has increased by, say, 3 percent, then they will estimate that the relative price of the good has increased by 1 percent and will increase their work effort accordingly.

15. Workers in Lucas's model can be viewed as displaying a type of "money illusion": they supply additional labor in response to expansionary monetary policy because for a time after the policy is implemented they believe, incorrectly, that the purchasing power of their income is higher than it will actually turn out to be. Unlike the analysts who preceded him, however, Lucas provided a rigorous explanation for the source of workers' money illusion. This extra step was crucial because it enabled him to ask (and answer) the question of whether the mechanism generating the money illusion would allow it to be exploited by policymakers. As we shall see, he concluded that it would not.

16. The rational expectations assumption was developed and first used by Muth (1961). However, Lucas (1972) was the first economist to accomplish the conceptually and mathematically challenging task of including rational expectations in a dynamic stochastic general equilibrium model. Sargent and Wallace (1975) illustrated the central importance of rational expectations by inserting this expectational assumption into a simple macroeconomic model of an otherwise-conventional (that is, non-neoclassical) type. The results were similar to those reported by Lucas: the model generated a Phillips curve-type relationship that government policy was powerless to exploit.

component that involves changes in the levels of employment and output, is generated by forces that have nothing to do with the systematic (policy-determined) component of monetary policy. As a result, deliberate, policy-induced changes in the inflation rate have no power to influence the unemployment rate in Lucas's model.[17]

Neoclassical Economics in Perspective. In the quarter-century since Lucas published this seminal paper, neoclassical theory has become the dominant school of thought among academic macroeconomists. To be sure, the neoclassical school has not escaped criticism. The rational expectations assumption, in particular, has been criticized as requiring unrealistically high levels of economic knowledge and forecasting ability on the part of households and firms and also because the econometric restrictions it implies are regularly rejected by the data. As a result, in recent years there has been a renewed interest in the implications of adaptive expectations, especially relatively sophisticated adaptive mechanisms such as least squares learning, Bayesian updating, and genetic algorithms (see, for example, Marcet and Sargent 1989 and Arifovic 1995). The goal of this research program is to try to better replicate the way in which real-world individuals learn from their mistakes and adjust their expectations to changes in the economic environment.

The significance of neoclassical macroeconomics is not that it has provided anything like a definitive macroeconomic model but that it has imposed more rigorous scientific discipline on macroeconomic theorizing.

Neoclassical use of general equilibrium models has been criticized on the grounds that the existing versions of these models are too simplistic and restrictive to capture the complex and diverse behavior of real-world households and firms. A closely related criticism is that neoclassical models simply cannot explain important macroeconomic phenomena. For example, although the "policy ineffectiveness" prediction of the original Lucas article has remained a fundamental part of the neoclassical message, a great many economists continue to believe that monetary policy has substantial real effects, and there is a good deal of empirical evidence supporting this position.[18]

In hindsight, it is clear that the significance of neoclassical macroeconomics is not that it has provided anything like a definitive macroeconomic model but instead that it has imposed more rigorous scientific discipline on macroeconomic theorizing. Stated differently, neoclassical macroeconomic theory is at an early stage of development, and there are many basic questions to which it has not yet been able to provide definitive answers. However, it has been very successful at identifying the logical and conceptual problems with the Keynesian and monetarist theories that preceded it.

Neoclassical macroeconomics has made a second major contribution to macroeconomic thought—a contribution that is less direct but perhaps equally important. By creating skepticism among economists that monetary or fiscal policy is responsible for business cycle fluctuations, it has forced them to recognize the possibility that the fluctuations may be caused by real forces—that is, by changes in technology, tastes, or resource costs of the type that cause supply and demand curves to shift in conventional microeconomic theory. In recent years, one of the fastest-growing branches of neoclassical macroeconomics has been real business cycle theory, which tries to attribute cyclical fluctuations to random changes in technological productivity. Kydland and Prescott (1982) pioneered in the development of this theory, and Nelson and Plosser (1982) provided empirical evidence that is widely viewed as indicating the importance of real as opposed to nominal factors in driving the business cycle.[19]

One basic prediction of real business cycle theory is that the observed changes in real wages and hours worked represent fluctuations in the relative value of labor—a prediction that has been emphasized in real business cycle studies by Hansen (1985) and Prescott (1986). As we have seen, this implication of the theory provides another argument against conducting monetary policy using rules of thumb based on the unemployment rate or the rate of wage inflation. Another interesting implication of neoclassical theory (though not necessarily of real business cycle theory) is that monetary policy and fiscal policy interact so that the effects of changes in monetary policy may depend partly or wholly on the response of fiscal policy. The first neoclassical economists to make this point forcefully were Sargent and Wallace (1981), who constructed a simple model in which the inflationary implications of a change in monetary policy depended critically (and dramatically) on how the government managed its debt. Again, this implication of the theory suggests that any reasonable set of rules for monetary policy guidance must be multidimensional in nature.[20]

Conclusion

Economic commentators regularly urge the Fed to use the level of unemployment or the rate of change in wages as leading indicators of inflation and as guides to whether they should ease or tighten monetary policy.[21]

The logic behind this approach is based on modern (post-1970s) Keynesian macroeconomics and, more specifically, on the Phillips curve and the NAIRU.

According to this view, inflation is caused by excessive "aggregate demand," and changes in aggregate demand show up first in the labor markets. Low levels of unemployment—levels below the natural rate/NAIRU—reflect the fact that excessive aggregate demand has produced a tight labor market. A tight labor market will put upward pressure on wages. Increases in wages will force firms to increase their prices and will consequently produce a higher rate of inflation. Since modern Keynesianism sees the state of monetary policy as the principal determinant of the level of aggregate demand, a tight labor market also reflects excessively expansionary monetary policy and indicates the need for corrective Fed tightening.

This article has attempted to provide some basic information about this NAIRU theory of the causes of inflation and the role of monetary policy. We began by describing the Phillips curve, an apparent empirical relationship between wage increases and unemployment that Keynesian economists used as the basis for a theory of the inflation-unemployment relationship. The theory implies that policymakers could use demand stimulus or restraint to produce lower or higher unemployment at the cost of higher or lower inflation. Monetarist economists, who were deeply skeptical of Keynesian views about the effectiveness of demand management, developed a critique of the Phillips curve that was based on the concept of a "natural" rate of unemployment. According to the monetarists, attempts to use monetary or fiscal policy to keep the unemployment rate below the natural rate might have limited success in the short run but in the long run would produce continually increasing inflation.

The stagflation (simultaneous high inflation and high unemployment) that afflicted the U.S. economy during the 1970s shook economists' faith in the existence of a stable Phillips curve and greatly increased the credibility of the monetarist "acceleration hypothesis." The proponents of Keynesian theory weathered the monetarist critique by accepting the natural rate—which they rechristened the NAIRU—as a long-run constraint on demand-management policies that, in their view, remained effective in the short run. Although the monetarists were not satisfied with the Keynesians' response to their critique, the fact that the two schools of macroeconomic thought were working with a common set of theoretical weapons prevented the monetarists from overwhelming the Keynesians' defenses. As a result, the modified Keynesian theory of the 1970s became the standard theory taught to economics students and used by policymakers. A basic feature of this theory was a simple rule of thumb for monetary policy: tighten policy when the unemployment rate is below the NAIRU or when real wages are rising, and ease policy when the reverse is true. The low unemployment rates observed in the mid-1990s have caused many commentators to urge the Fed to consider using this rule of thumb as a justification for preemptive monetary tightening.

After describing the historical development of the NAIRU theory, the discussion raises some practical questions about the validity of the theory and its usefulness as the basis for policy advice. Perhaps the most important question involved the difficulty of distinguishing policy-induced changes in absolute wages from changes in relative wages associated with real changes in the economy—changes that it would not make sense for monetary policymakers to attempt to oppose. A second question focused on the fact that there is very little empirical evidence supporting the notion of sticky prices on which

17. In his paper, Lucas imagines a researcher who tries to use a statistical analysis of data from his model to provide advice to policymakers. The researcher runs a linear regression with output or employment as the dependent variable and the inflation rate as the independent variable. He finds that the coefficient estimate for the inflation rate is positive and consequently advises policymakers that using monetary policy to increase the inflation rate is likely to succeed in increasing the levels of employment and output. As we have seen, however, this policy advice is incorrect.

18. Strictly speaking, neoclassical theory does not preclude monetary policy from having real effects: it simply rules out real effects that rely on frictionless markets not clearing or on the public being systematically fooled. Thus, although it is arguably fair to describe monetary policy ineffectiveness as a characteristic feature of neoclassical theory, there are an increasing number of neoclassical models in which monetary policy has short-run real effects. Leeper and Gordon (1992) and references therein are examples of key contributors to the rapidly growing "liquidity effects" literature, which uses real business cycle models (see below) to study the short-run effects of changes in monetary policy. There are also a few neoclassical models in which monetary policy has long-run real effects. Examples of the latter type include Wallace (1984), Bhattacharya, Guzman, and Smith (1996), Espinosa and Russell (1997a, 1997b), and Bullard and Russell (1997).

19. For a more detailed description of real business cycle theory and a review of the formative developments in the theory see Prescott (1986).

20. The following statement by former Federal Reserve Governor Lawrence Lindsey provides a good example of unidimensional reliance on the NAIRU: "The NAIRU is a useful theoretical construct . . . sufficient for making quick 'on your feet' estimates of likely economic performance. . . . If I knew with certainty that the NAIRU was 5.837. . . I would have the information I needed to know with certainty that I should tighten" (1996, 10).

21. Ironically, experts who specialize in studying the properties of the business cycle classify wages as lagging rather than leading indicators. See Moore (1961) and Zarnowitz (1992).

Keynesian theory is based, and a third involved the empirical weakness of the Phillips curve relationship that provides the basis for the NAIRU.

The discussion also includes neoclassical economics, a relatively new school of macroeconomic thought that has provided a second, more fundamental challenge to Keynesian thought. We described the fundamental principles of neoclassical theory and went on to explain how Robert Lucas, one of the theory's founders, used these principles to construct a groundbreaking theoretical model whose properties cast doubt on the short-run effectiveness of monetary policy and thus on the usefulness of monetary policy rules based on the NAIRU. Neoclassical theory still has a large number of basic macroeconomic questions to answer. However, it has produced huge logical and methodological improvements in macroeconomic analysis, and it has left the Keynesian and monetarist theories that preceded it largely discredited—including the modern form of Keynesian theory that provides the basis for the NAIRU. Recent developments in neoclassical theory indicate that business cycle fluctuations in employment and output may be caused primarily by real forces—a situation that, if true, increases the danger that monetary policy based on the NAIRU may interfere with the proper functioning of the price system.

Our own view is that proponents of the NAIRU have never provided anything like a satisfactory answer to the neoclassical critique, or even to the questions raised in this article. Given that this is the case, it is hard to give much credence to the commentators who urge the Fed to base its monetary policy on the NAIRU. Unfortunately, neoclassical economists have yet to provide monetary policymakers with reliable policy rules to replace NAIRU-based rules. Until they do, monetary policy decision making will remain a difficult task.

REFERENCES

Arifovic, Jasmina. 1995. "Genetic Algorithms and Inflationary Economies." *Journal of Monetary Economics* 36 (December): 219-43.

Ball, Laurence, and N. Gregory Mankiw. 1995. "Relative-Price Changes as Aggregate Supply Shocks." *Quarterly Journal of Economics* (February): 161-93.

Bhattacharya, Joydeep, Mark G. Guzman, and Bruce D. Smith. 1996. "Some Even More Unpleasant Monetarist Arithmetic." Cornell University, Center for Analytic Economics, CAE Working Paper 95-04, April.

Bullard, James, and Steven Russell. 1997. "How Costly Is Sustained Low Inflation for the U.S. Economy?" Federal Reserve Bank of St. Louis and IUPUI Working Paper.

Chang, Roberto. 1997. "Is Low Unemployment Inflationary?" Federal Reserve Bank of Atlanta *Economic Review* 82 (First Quarter): 4-13.

Espinosa, Marco A., and Steven Russell. 1997a. "Can Higher Inflation Reduce Real Interest Rates in the Long Run?" *Canadian Journal of Economics* (forthcoming).

———. 1997b. "Conventional Monetary Policy Wisdom in the Diamond Model." Federal Reserve Bank of Atlanta, working paper, forthcoming.

Fisher, Irving. 1926. "A Statistical Relationship between Unemployment and Price Changes." *International Labor Review* 13 (June): 785-92.

Friedman, Milton. 1968. "The Role of Monetary Policy." *American Economic Review* 68 (March): 1-17.

———. 1976. "Wage Determination and Unemployment." Chap. 12 in *Price Theory*. Chicago: Aldine Publishing Company.

Friedman, Milton, and Anna J. Schwartz. 1963. *A Monetary History of the United States, 1867-1960.* Princeton, N.J.: Princeton University Press.

Hansen, Gary. 1985. "Indivisible Labor and the Business Cycle." *Journal of Monetary Economics* 16 (November): 309-27.

Keynes, John M. 1964. Reprint. *The General Theory of Employment, Interest, and Money*. San Diego: Harcourt Brace and Company. Originally published, 1953.

Kydland, Finn, and Edward Prescott. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica* 50 (November): 1345-70.

Leeper, Eric M., and David B. Gordon. 1992. "In Search of the Liquidity Effect." *Journal of Monetary Economics* 29 (June): 341-69.

Lindsey, B. Lawrence. 1996. "NAIRU Disrobed." *International Economy* (March/April): 8-13.

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

Lucas, Robert E., Jr., and Thomas Sargent. 1979. "After Keynesian Macroeconomics." Federal Reserve Bank of Minneapolis *Quarterly Review* 3 (Spring): 1-16.

Marcet, Albert, and Thomas Sargent. 1989. "Convergence of Least-Square Learning Mechanisms in Self-Referential Linear Stochastic Models." *Journal of Economic Theory* 48 (August): 337-68.

Modigliani, Franco, and Lucas Papademos. 1975. "Targets for Monetary Policy in the Coming Year." *Brookings Papers on Economic Activity*, no. 1:141-63.

MOORE, GEOFFREY. 1961. *Business Cycle Indicators*. Vol. 1 of *Contributions to the Analysis of Current Business Conditions*, National Bureau of Economic Research. Princeton, N.J.: Princeton University Press.

MUTH, J.F. 1961. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29 (July): 315-35.

NELSON, CHARLES R., AND CHARLES I. PLOSSER. 1982. "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications." *Journal of Monetary Economics* 10:139-62.

OKUN, ARTHUR M. 1980. "Postwar Macroeconomic Performance." In *The American Economy in Transition*, edited by Martin Feldstein, 162-69. Chicago: University of Chicago Press.

PHELPS, EDMUND. 1967. "Phillips Curves, Expectations of Inflation, and Optimal Unemployment over Time." *Economica* 34 (August): 254-81.

PHILLIPS, A.W. 1958. "The Relationship between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861-1957." *Economica* 25 (November): 283-99.

PIGOU, A.C. 1933. *The Theory of Unemployment*. London: Macmillan.

PRESCOTT, EDWARD. 1986. "Theory Ahead of the Business Cycle." *Carnegie-Rochester Conference Series on Public Policy* 25 (Autumn): 11-44.

SAMUELSON, PAUL A., AND WILLIAM NORDHAUS. 1989. *Economics*. 13th ed. New York: McGraw-Hill.

SAMUELSON, PAUL A., AND ROBERT M. SOLOW. 1960. "Analytical Aspects of Anti-Inflation Policy." *American Economic Review* 50 (May): 177-194.

SARGENT, THOMAS, AND NEIL WALLACE. 1975. "Rational Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *Journal of Political Economy* 83 (April): 241-54.

———. 1981. "Some Unpleasant Monetarist Arithmetic." Federal Reserve Bank of Minneapolis *Quarterly Review* 5 (Fall): 1-17.

STAIGER, DOUGLAS, JAMES STOCK, AND MARK WATSON. 1997. "The NAIRU, Unemployment and Monetary Policy." *Journal of Economic Perspectives* 11 (Winter): 33-50.

TOBIN, JAMES. 1980. "Stabilization Policy Ten Years After." *Brookings Papers on Economic Activity*, no. 1:19-71.

WALLACE, NEIL. 1984. "Some of the Choices for Monetary Policy." Federal Reserve Bank of Minneapolis *Quarterly Review* 8 (Winter): 15-24.

WYNNE, MARK. 1995. "Sticky Prices: What Is the Evidence?" Federal Reserve Bank of Dallas *Economic Review* (First Quarter): 1-12.

ZARNOWITZ, VICTOR. 1992. *Business Cycles: Theory, History, Indicators, and Forecasting*. Vol. 27 of *Studies in Business Cycles*, National Bureau of Economic Research. Chicago: University of Chicago Press.

Identifying Monetary Policy: A Primer

TAO ZHA
The author is an economist in the macropolicy section of the Atlanta Fed's research department. *He is grateful to Roberto Chang, Frank King, Eric Leeper, Larry Wall, and especially Jerry Dwyer and Mary Rosenbaum for valuable comments.*

THE POPULAR PRESS AND UNDERGRADUATE ECONOMICS TEXTBOOKS HAVE LONG CONCLUDED THAT AN INCREASE IN THE FEDERAL FUNDS RATE TARGET BY THE FEDERAL OPEN MARKET COMMITTEE (FOMC) TENDS TO SLOW GROWTH OF NATIONAL OUTPUT AND REDUCE INFLATIONARY PRESSURES. ECONOMISTS GENERALLY AGREE ON THIS POINT, BUT THEY DISAGREE CONSIDERABLY ABOUT THE QUANTITATIVE IMPACT OF MONETARY POLICY. FOR EXAMPLE, A GROUP OF ECONOMISTS CALLED MONETARISTS ARGUE THAT "IN THE SHORT RUN, WHICH MAY BE AS LONG AS THREE TO TEN YEARS, MONETARY CHANGES AFFECT PRIMARILY OUTPUT" BUT NOT PRICES (FRIEDMAN 1992, 48) WHILE OTHER ECONOMISTS SUCH AS REAL BUSINESS CYCLE THEORISTS POSTULATE THAT MONETARY CHANGES AFFECT ONLY PRICES BUT HAVE LITTLE OR NO EFFECT ON OUTPUT (COOLEY AND HANSEN 1995).

As it stands, economists' beliefs about the quantitative importance of monetary policy stem largely from theoretical models through which the policy effects of changing monetary policy are inferred. It is no surprise, then, that different conclusions arise from different experiments or theories. The actual economy, however, is not the result of any such controlled experiment. Obviously, a central bank cannot change policy for the sake of examining its effect on the economy. In the real world, inferences about the quantitative effect of monetary policy must rely on observations of actual economic activity in which many variables are changing simultaneously. What can be observed is the equilibrium outcome of interaction among all players in the economy—the central bank, financial market participants, producers, and consumers. On this playing field, sorting out the central bank's behavior from that of the many other participants is the first and critical step in attempting to estimate the actual impact of monetary policy. This sorting-out process is known in technical parlance as identification.

Identification of monetary policy is partly a conceptual (economic) issue and partly an empirical (technical) one. Conceptually, the process requires that one understand the economics of the demand and supply of money, or, in other words, the interaction between the central bank's reaction to economic conditions and the private sector's response to policy actions. Empirically, one needs sophisticated mathematical tools to isolate the central bank's behavior from all other behaviors in the observed data and examine its consequences.

The purpose of this article is to explain these two issues: the conceptual one of why identification of monetary policy is important and the empirical issue of how difficult it is in practice. The article focuses on these two issues exclusively because of how vital careful identifica-

tion is for an accurate assessment of policy effects. Given this purpose, the article refrains from discussing how to resolve the disparate views about the actual quantitative effect of monetary policy on a given country's economy. The discussion first explores identification of monetary policy as having much in common with issues familiar to us from basic economics principles. The article then discusses the identification issue special to the analysis of monetary policy and illustrates the process with a few examples of identifying monetary policy in different countries.

Demand and Supply

The abstract concept of money is clear: money is something the public accepts in exchange for goods and services. In reality, however, the measure of money is not so well defined: money can be currency in circulation, reserves, the monetary base (the sum of currency in circulation and reserves), M1 (currency plus checkable deposits), or M2 (M1 plus other assets). Whichever monetary aggregate is used, the analysis of monetary policy inevitably encounters the two blades of the monetary scissors: demand for and supply of money. Thus it is appropriate to begin exploring the importance of identifying monetary policy with an analysis of the demand and supply of money.

A simple, familiar example is instructive: the demand-supply relationship in the market of goods and services, in this case the wine market. If one has the data on the price of wine (p) and the quantity bought and sold (q), the bivariate demand-supply relationship can be described by the following two equations:

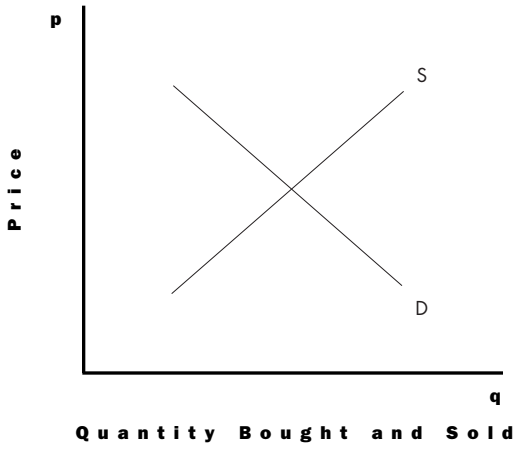
$$q = \alpha_1 p + \alpha_2 X + \epsilon_d \quad \text{(demand)}, \qquad (1)$$

$$q = \beta_1 p + \beta_2 Y + \epsilon_s \quad \text{(supply)}, \qquad (2)$$

where X is a set of variables (such as the government's excise tax and consumers' income) that affect the demand for wine, Y is a set of variables (such as the government's excise tax and weather condition) that affect the supply of wine, the α coefficients describe the behavior of consumers, and β, the behavior of producers.

Before proceeding, explaining a few common notations and notions will lay the groundwork for discussion of these and additional equations. The notations q, p, X, and Y in equations (1) and (2) are variables while α_1, α_2, β_1, and β_2 are coefficients. The sharp distinction between the "coefficient" and "variable" is an important one. A variable has a quantitative value observed in the data so that, for example, the variable q represents the price of wine bought and sold. A coefficient does not come directly from the data; rather, its quantitative value must be obtained by statistical methods. The process of obtaining the value of a coefficient is called estimation, a concept

CHART 1
Demand for and Supply of Wine

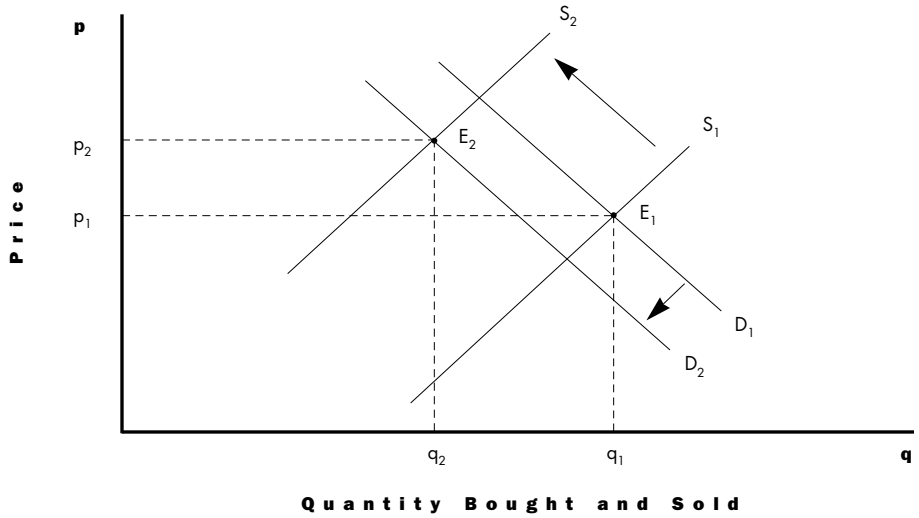


important throughout the article. For instance, coefficients such as α_1 are to be estimated through equations (1) and (2). Finally, the notation ϵ_d represents a random change that cannot be described by normal demand behavior, and this article calls ϵ_d a demand shock. The word *shock* has its familiar meaning of referring to something unpredictable. Similarly, the supply shock ϵ_s in equation (2) indicates an unpredicted change in the supply of wine.

Economic theory implies that price is inversely related to quantity demanded (that is, $\alpha_1 < 0$). It also tells us that when the price is higher, the firm is willing to produce more wine ($\beta_1 > 0$). These relationships can be depicted in a two-dimensional figure like Chart 1, where the downward-sloping curve represents demand behavior and the upward-sloping curve represents supply behavior. Chart 1 is drawn under the assumption that variables other than p and q are held fixed. Therefore, if there are any changes in X or Y, the curves in Chart 1 will shift from the original equilibrium position. For example, when the government raises the tax on wine, both demand and supply will fall, their curves will shift to the left, and the equilibrium will change from E_1 to E_2 (Chart 2). How much the quantity of wine will be reduced (from q_1 to q_2) in the market depends on the behavior of consumers and producers, which is described in equations (1) and (2). In other words, it depends on how far the demand and supply curves in Chart 2 will shift. The policy analyst, to assess the tax's effect on the behavior of consumers and producers, must understand (correctly identify) both the demand function (1) and the supply function (2).

The argument about the importance of identifying the demand for and supply of wine can be carried over to the money market, although monetary analysis is of course far more complicated (see, for example, Leeper 1992). To begin with, one can think of the quantity of

CHART 2 Effect of the Government's Tax on Wine



money as resembling the quantity of wine and the opportunity cost of holding money (the interest rate) as the price of wine. Let M represent money and R, the nominal interest rate. Thus, analogous to the wine market, where q and p are jointly determined, M and R are determined by both demand and supply in the money market. Assume that all deposits in M do not bear interest.[1] The demand function for money derived in standard textbooks can be expressed as

$$M - P = \alpha_1 y + \alpha_2 R + \epsilon_{MD} \text{ (money demand)}, \quad (3)$$

where y represents national output; P, the general price level; α_1, the coefficient of y; α_2, the coefficient of R; and ϵ_{MD}, the money demand shock.[2] The coefficient α_1 measures the percent change in demand for money in response to a percent change in output y, and the measure is known as the income elasticity of money demand.[3] As consumer income rises, the demand for goods and services will increase, and in turn their demand for money will rise so that they can purchase goods and services. Thus, the coefficient α_1 is expected to be positive. The coefficient α_2, known as the interest elasticity of money demand, measures the percent change in money demand in response to a percent change in the interest rate. Since the public is willing to hold less (real) money $(M - P)$ when the cost (R) of holding money increases, the interest elasticity α_2 is expected to be negative. If one depicts the demand curve in the (M, R) plane (see Chart 3), the curve of demand for money has a negative slope (analogous to the downward-sloping curve in the demand for wine in Chart 1).

When broad monetary aggregates such as M1 or M2 are used, the term *money supply* in general involves not merely the behavior of the central bank but also the behavior of banks and other financial institutions whose liabilities (such as checking deposits) serve as part of the medium of exchange as well as the behavior of depositors who decide how much currency to hold in relation to deposits. A central bank, through its open market operations or discount window lending, can affect monetary aggregates through the banking system. To see how a monetary aggregate such as M2 is affected, suppose that the Federal Reserve decides to increase the monetary base (the sum of the currency in circulation and reserves) by buying Treasury securities worth, say, $\$\chi$ from a seller. Suppose the seller, now becoming a depositor, decides to deposit the full amount of $\$\chi$ in Bank A, creating $\$\chi$ in deposits in the bank. After meeting the reserve requirement (that is, the certain percentage of $\$\chi$ that must be kept in the bank), Bank A lends part of the deposit to households who, now becoming depositors, decide to deposit the loans in, say, Bank B. The process can continue. Eventually, such a chain of deposit expansions through bank loans makes an increase in M2 a multiple of the initial increase in the monetary base. Thus the term *money multiplier*, defined by the ratio of the monetary aggregate (like M2) to the monetary base, is used to indicate the extent to which money is created or multiplied through the participation of both banks and depositors.

The incentive to increase deposits in the banking system lies in the prospect of making profitable loans. If the prospect is dim or the demand for loans falls off, banks may not create deposits up to the limit the reserve requirements allow. Thus they may, from time to time, have excess reserves in addition to required reserves. Furthermore, because of the uncertainty about deposit flows and transaction clearing within a day and from day to day, banks typically hold some excess reserves although there are incentives to minimize them. Clearly,

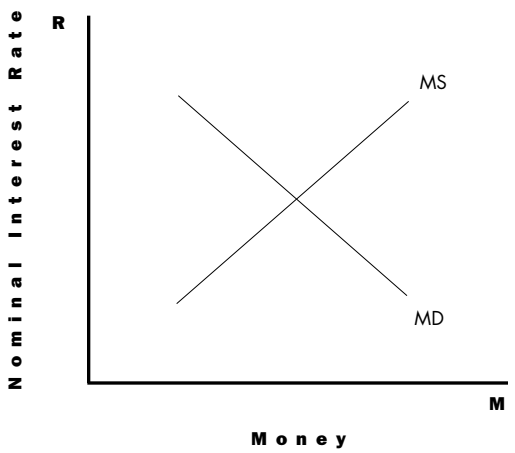
banks' decisions about how much needs to be held as excess reserves, combined with depositors' decisions about how their portfolio should be allocated, can cause the supply of money to change. The central bank is not the only player whose behavior influences the money supply or the money multiplier. Taking all these behaviors into account, the supply function for money can be derived from the money multiplier (see Box 1 on page 32) and usually has the following form:

$$M = \alpha_3 R + \alpha_4 X_s + \epsilon_{MS} \text{ (money supply)}, \qquad (4)$$

where the coefficient α_3 is the interest elasticity of money supply, the coefficient α_4 is the elasticity in relation to X_s, which is a set of variables (such as reserves and output) that can influence the supply of money, and ϵ_{MS} is the money supply shock, which is uncorrelated with the money demand shock ϵ_{MD}. The α coefficients are to be estimated, and the sign of α_3 is expected to be positive. One interpretation for the positive sign of α_3 follows the logic in Box 1: since each dollar of excess reserves is costly to hold because of forgone interest, the amount of excess reserves tends to decline as the rate of interest rises. Through the money multiplier effect, deposits and monetary aggregates tend to increase. Thus, one would expect an upward-sloping curve for the money supply function as depicted in Chart 3.

From the viewpoint of the policymaker the question is, Why is it important to separate the demand for money from the supply of money? Remember that in the example of the wine market, distinguishing the demand behavior of consumers and the supply behavior of producers is important for assessing the effect of the government's tax on the behavior of consumers and producers (recall Chart 2). Here, the central bank needs to assess policy effects in order to attain its objective. When policy actions shift the money supply curve (for example, from MS_1 to MS_2 in Chart 4), the change in the equilibrium quantity of money and the equilibrium rate of interest (from E_1 to E_2 in Chart 4) depends on two factors: the slope of the money demand curve as well as the slope of the money supply curve. Thus this section has shown that understanding the demand and supply of money is crucial for assessing (identifying) policy effects on, say, the quantity of money (M) and the rate of interest (R).

CHART 3
Money Demand and Money Supply



Central Banks' Behavior

The distinction between the demand for and supply of money, explained in the monetary model (3)-(4), is analogous to the intuitive demand-supply analysis in the wine market. What the model does not show is a problem unique to policy analysis: separating the central bank's behavior from the behavior of the banking system and depositors. It is the purpose of this section to discuss in detail the behavior of the central bank.

For simplicity of analysis, textbooks usually use the money supply curve in Chart 3 to represent the central bank's behavior under either of the following two assumptions. One is that the central bank is in complete control of a broad monetary aggregate like M2, in which case the money supply curve is vertical (Chart 5). The other assumption is that unpredicted changes in reserves are caused solely by unpredicted monetary policy shifts (ϵ_{MS}) and that these changes are the only random sources affecting (that is, shifting) the money supply curve in Chart 3.

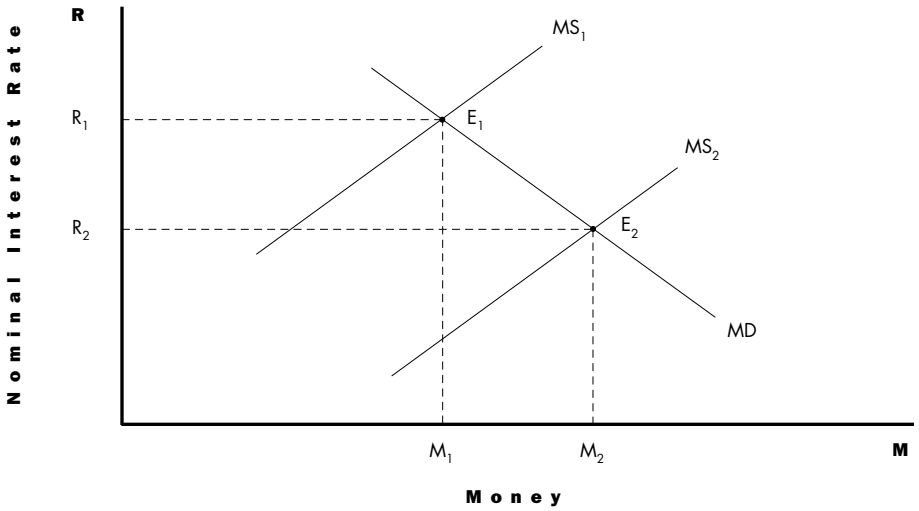
Two caveats are in order. First, these two assumptions are generally not a good description of what actually happens. The central bank can heavily influence broad monetary aggregates such as M2 but cannot control them completely. Moreover, unpredicted changes in reserves can be caused by shifts in banks' demand for reserves or by depositors' portfolio adjustment and thus

1. This assumption is made for the convenience of analysis. Some deposits such as savings accounts in M2 do pay interest on their balances. Then, the cost of holding these assets is reflected by the difference between the interest rate on these assets and the interest rate on other assets such as government bonds. This feature would complicate the analysis but not alter the basic conclusions.

2. All variables discussed in this article except for interest rates are logarithmic. Thus, M – P *is the log of real balances (the money stock deflated by the general price level).*

3. Note the adherence to the convention of using the term income *instead of* output; *both names denote the variable* y. *The concept of elasticity, the percent change in one variable in response to a percent change in another variable, is frequently used in the article.*

CHART 4 Shift of the Money Supply



do not necessarily indicate policy shocks. Second, even if these two assumptions are reasonable, the money supply function discussed does not to this point describe the real world of the central bank's behavior.

What is the real world behavior of the central bank, after all? More important, to what extent can a policy analyst write down a function (functions) or equation that gives a good approximation of that behavior? The macroeconomic policy aspect of many central banks' behavior reflects both their responsibility for controlling inflation and their attention to policy's effect on overall economic activity. In the day-to-day implementation of U.S. monetary policy, for example, the Federal Reserve sets a target for the federal funds rate according to its objective. Its attempts to meet the target require tracking the amount of reserves and subsequently of deposit flows and monetary aggregates. In choosing its target, the Federal Reserve regularly examines economic forecasts prepared by its staff. The staff frequently explore the historical relationships between key macroeconomic variables (such as inflation and output) and policy instruments (such as reserves and the federal funds rate) and provide alternative economic outlooks under different assumptions about policy instruments such as different levels of the federal funds rate. Policymakers then decide what actions to take in order to attain their objectives.

The process of such policymaking is common across different industrial countries. For example, for the Bank of France, senior management "assesses the reserve position of the banking system and evaluates whether current market interest rates, especially the interbank rate, are consistent with the current stance of monetary policy and foreign rates. Instructions are then given to the money market trading room at the Bank of France to intervene in the interbank market on the basis of the evaluations of money market and general macroeconomic conditions" (Batten and others 1990, 78). The Bank of Canada "uses economic projections to translate the Bank's objectives into suggested paths for the instruments of policy, and uses various economic and financial indicators, notably monetary aggregates, to monitor progress and help the Bank to act in a timely fashion when necessary" (Duguay and Poloz 1994, 197).

In short, a central bank tries to achieve its objective subject to the constraints imposed by the private sector's activity. As a result, the central bank comes out with a strategy or plan by reviewing the state of the economy.[4] This article refers to this strategy as the policy reaction function(s) and henceforth uses it to mean monetary policy or the central bank's behavior throughout. The reaction function is therefore composed of two components: the systematic reaction of policy to economic conditions and unexpected shifts in policy (policy shocks).

The discussion first turns to the systematic component of monetary policy because it is the essence of the specification (that is, description) of a reaction function. For illustration, suppose the Federal Reserve's objective is to stabilize inflation at some low level with the federal funds rate as a policy instrument. Since the Federal Reserve has no direct control over the general price level, it uses the federal funds rate to influence intermediate targets such as the three-month Treasury bill rate and M2. Unfortunately, there is no simple linkage of the federal funds rate with M2 and of M2 with the general price level, at least in the short run. The price level today is affected not only by current and past movements in other variables such as the federal funds rate, M2, and output but also by previous changes in the price level itself. The changes in all the variables reflect the interaction

between policy actions and economic activity in the current and previous periods. To attain price stability, the Federal Reserve will adjust its federal funds target in response to changes in all crucial economic variables such as M2, the price level, and output. The reaction function can therefore be summarized as

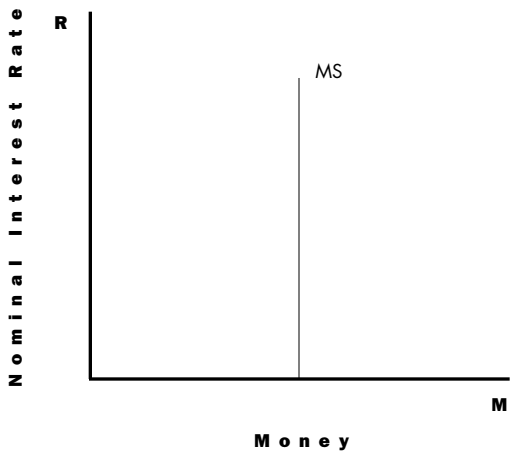
$$FR = \beta_1 M + \beta_2 R + \beta_3 X + \epsilon_{MS} \quad \text{(policy reaction)}, \quad (5)$$

where *FR* stands for the federal funds rate; *M*, for M2; *R*, the three-month Treasury bill rate; and *X*, a set of other crucial variables that are used by the Federal Reserve to predict fluctuations in the general price level.[5] To give an example of which variables are contained in *X*, suppose monthly data are used to estimate the β coefficients in the Federal Reserve's reaction function (5). The set *X* should include all the crucial variables the Federal Reserve uses in making its policy decisions. The variables that may be excluded in *X* are the price level and output in the present month, on the grounds that the data on these variables are available only after the end of the month. Similarly, one should include in the reaction function not only variables (such as the federal funds rate, commodity prices, M2, exchange rates, output, and the price level) in the previous months but also variables such as commodity prices, M2, and exchange rates in the present month. Current data on commodity prices and financial variables provide the Federal Reserve with information about the market's expectation of future inflation while the data from the previous months help predict future economic activity.

Given the systematic component $\beta_1 M + \beta_2 R + \beta_3 X$ in the policy reaction function (5), the sign of the coefficient β_1 is particularly interesting because it indicates how the federal funds rate responds to a change in the monetary aggregate. Suppose there is an increase in the monetary aggregate. If the central bank believes that such an increase will lead to a rise in future inflation, it will tend to increase the federal funds rate in order to offset the rising monetary aggregate. The sign of β_1 is therefore expected to be positive.

The second component of equation (5) is the random shock ϵ_{MS}, which reflects an unpredicted shift in monetary policy. The notion of randomness here is the same as when newspapers use the term *shock* to refer to an oil shock, which appears random because it is unpredictable. Likewise, policy shocks occur when the central bank's instrument changes unpredictably. To explain further, consider the Federal Reserve's policy. Suppose the Federal Reserve's objective is to keep inflation low in the long run and its policy instrument is the federal funds rate. In the short run (say, three to ten years), however, the dynamic relationships between output, unemployment, inflation, and the federal funds rate are complicated, and the trade-off between inflation and output may be substantial and is uncertain. A policy decision reached by sifting through such uncertain relationships can be as unpredictable as any other economic condition.[6] Such an unpredicted movement in the federal funds rate is called a policy shock—ϵ_{MS} in equation (5)—while the predicted movement (systematic reaction) is characterized by $\beta_1 M + \beta_2 R + \beta_3 X$.

CHART 5
Perfectly Inelastic Money Supply (α_3=0)



Note the close connection between the functional forms (4) and (5): the reaction function (5) can be rearranged to have the same functional form as the money supply function (4), and X_s in (4) can then be thought of as including both *X* and the federal funds rate (*FR*) in (5). Is the sign of α_3 in the newly derived function (4) positive as in the original money supply function? Recall the argument for the positive sign of β_1 in equation (5): the Federal Reserve tends to increase the federal funds rate in order to offset the rising monetary aggregate (leaning against the wind, so to speak). When the federal funds rate (*FR*) is expected to rise, the three-month Treasury bill rate (*R*) will tend to rise because there is a strong positive relationship between *R* and expected *FR*.[7] Thus one should expect the positive sign of

4. Formally, one can think of the strategy coming from the first-order conditions in the central bank's maximization problem in a theoretical model.

5. The term ϵ_{MS} will be discussed later.

6. See Leeper, Sims, and Zha (1996) for further discussions.

7. Such a relationship is known as the expectation theory of the term structure in the economic literature. For details of the theory, the reader can consult any standard textbook of monetary economics or finance—for example, Mishkin (1992).

Traditional Approach to the Supply of Money

In standard textbooks, "money supply" typically refers to the joint behavior of the central bank, commercial banks, and depositors. The derivation of the money supply function or equation in this box draws directly from McCallum (1989, 55-73). Suppose that M1 is the definition of money and the central bank resembles the Federal Reserve. By definition,

$$M = C + D, \tag{A1}$$

where C stands for the currency in circulation (held by the nonbank private sector) and D the checkable deposits. The monetary base MB that is heavily influenced by the central bank is

$$MB = C + TR. \tag{A2}$$

Note that TR stands for the total reserves, which can be further broken into two components:

$$TR = RR + ER, \tag{A3}$$

where RR is the required reserves and ER is the excess reserves.[1] The key relationships between the deposits and the other variables are

$$C/D = cr, \tag{A4}$$

$$TR/D = rr, \tag{A5}$$

$$RR/D = k, \tag{A6}$$

$$ER/D = e(R), e'(R) < 0, \tag{A7}$$

where the ratios cr, rr, and k are assumed to be constant (or are not influenced by other variables if changing over time). The assumption that $e(R)$ decreases with R is based on the belief that the banks will hold less excess reserves when the interest rate R rises. A rise in R indicates the opportunity cost of holding the excess reserves.

Using (A4), one can rewrite equation (A1) as

$$M = (cr + 1)D. \tag{A8}$$

Combining (A2), (A3), (A4), (A5), (A6), and (A7) leads to

$$MB = [cr + k + e(R)]D. \tag{A9}$$

The money multiplier, defined by the ratio of money to the monetary base, can be derived from equations (A8) and (A9):

$$\frac{M}{MB} = \frac{cr + 1}{cr + K + e(R)}. \tag{A10}$$

Using the expression $\mu(R; k, cr)$ to summarize the right-hand term of equation (A10) yields the simple money supply function:

$$M = \mu(R; k, cr)MB. \tag{A11}$$

It is obvious from (A7) and (A10) that the money supply function (A11) implies the upward-sloping curve of money supply. Thus, the function (A11) can be written in the form of equation (4), where the condition $\alpha_3 > 0$ reflects the upward-sloping curve of the money supply.

1. Some central banks, like Canada's, no longer have the legal reserve requirement. But such banks still hold reserves in response to the withdrawal of deposits. In the case of Canada, RR *can be thought of as the desired reserves—the amount the banks desire to hold (see Barro and Lucas 1994).*

α_3 and the upward-sloping curve of the newly derived function (4) when depicted in the two-dimensional (*M, R*) chart. Now, however, the new function (4) has a different interpretation: it describes the policy behavior, not a joint behavior of the central bank, commercial banks, and depositors. Despite the nuances in interpreting the same functional form (4), the practice of calling the reaction function "money supply" is very common because it is always intuitive to think of demand and supply. Accordingly, this article shall continue to interchange the terms.

In summary, this section discusses the behavior of the central bank, emphasizes the significance of understanding the policy's systematic response to economic conditions, and shows how such behavior can be modeled or approximated by the policy reaction function (5). Moreover, it reinterprets the traditional money supply function discussed in the previous section as the policy reaction function but does so without changing the characteristics of the money supply function (for example, the upward-sloping curve of the supply function). Given this reinterpretation, one is able to analyze the effect of monetary policy in the intuitive framework of demand and supply, as will be shown in later sections.

Other Points about Identifying Monetary Policy

The antecedent sections establish the importance of identifying monetary policy (separating money demand from money supply) and describe how the money supply function (4) can be used to approximate a central bank's behavior. Even so, the point about the importance of identifying monetary policy is still often misunderstood. Two popular contentions merit further discussion.

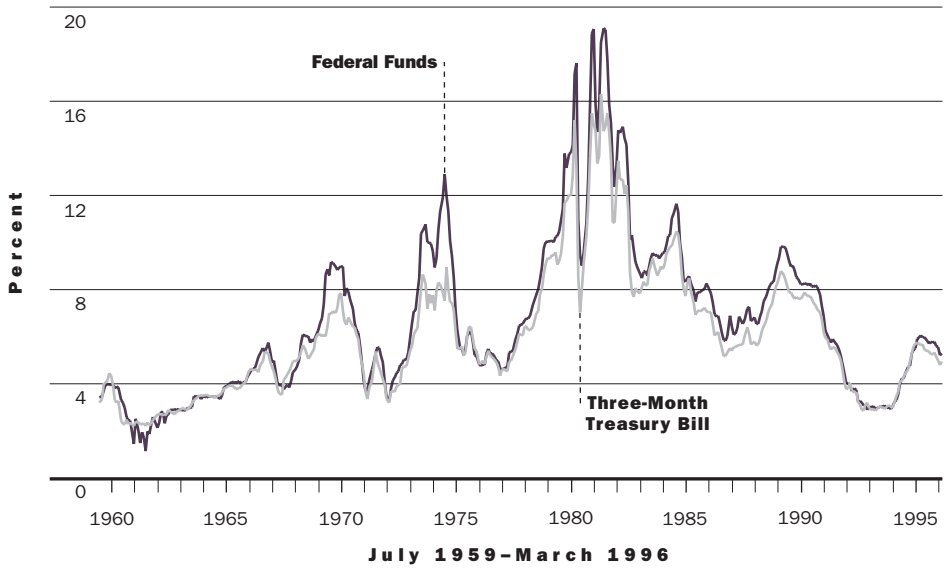
One position is that central banks know exactly what their monetary policy or behavior is and from their viewpoint there is no need to separate the money demand and money supply. For example, if a central bank's objective is a commitment to price stability, monetary policy is to make the general price level stable, and thus the private sector's behavior (such as the demand for money) is not the policymaker's concern. While this notion seems prima facie sensible, it is simply incorrect. It confuses the central bank's objective with its policy, which, as the discussion in the last section argues, is a strategy designed to achieve the objective. The real issue is not whether the central bank knows its objective; the real issue is whether it knows how to form a strategy (monetary policy) to attain its objective. The formation is difficult and requires a thorough understanding of the interaction between the central bank's behavior (money supply) and the private sector's activity.

Consider, for example, Canadian monetary policy. Analysis of Canadian monetary policy is instructive not only because most countries resemble Canada in the sense that they are small and open relative to the U.S. economy but also because it is of interest to U.S. policymakers as the U.S. economy has become increasingly integrated with the rest of the world, especially with other major industrialized countries. Suppose the price of world commodities suddenly drops while other conditions or variables do not change. Since Canada is an exporter of raw materials and commodities, Canadian residents' income will decline. Falling Canadian income means a decrease in the demand for Canadian money, which by itself would lower the exchange value of Canadian currency. If a falling Canadian currency has a direct positive impact on Canadian prices in the short run (see Dornbusch and Krugman 1976), the Bank of Canada will try to stabilize the exchange rate in hopes of stabilizing the price level. In the process of formulating such a monetary policy, the Bank of Canada must have a fairly accurate idea of the demand for Canadian money. It also must have a strong sense of its own behavior (money supply) in order to predict the equilibrium quantity of money and the equilibrium rate of interest (the intersection of demand and supply in Chart 3), as changes in both the money stock and the interest rate will affect the exchange rate and the price level. This example is important because monetary policy in most countries, unlike in the United States, resembles Canadian monetary policy in the sense that the domestic economy is heavily influenced by foreign economies and the exchange rate plays a considerable role in policy formation.

To assess the effect of monetary policy requires understanding the interaction between the central bank's behavior and the private sector's activity.

In the case of the United States, some would say that monetary policy is easy to formulate: it calls for simple adherence to the federal funds rate target the Federal Reserve itself chooses. Again, this argument is a sophism. The federal funds rate target is not set arbitrarily; it reflects the Federal Reserve's concern about its own objective of, say, price stability. When fluctuations in economic activity or the repercussions of past policy choices threaten such an objective under the current rate of federal funds, a new target for the federal funds rate will be chosen. Indeed, as shown in Chart 6, the federal funds rate has changed over time, sometimes frequently. How the target is set reflects how the Federal Reserve reacts to the changing state of the economy, which is described by the reaction function (5) or (4). The Federal Reserve's

CHART 6
Time Series Pattern of Federal Funds Rate and Three-Month Treasury Bill Rate



reaction function can be complicated because there is no simple relationship between the federal funds rate and the general price level, at least within three to ten years.

The other popular contention that questions the role of separating money demand and money supply argues that although a central bank's decision is based on its staff's forecasts of a wide range of macroeconomic variables, such forecasts do not identify distinctive behaviors of the central bank and the private sector. For example, in most models that forecast real gross domestic product (GDP), monetary aggregates, interest rates, and prices, the central bank's reaction function is not explicitly specified or sorted out. But it does not follow that the policymaker has no idea of money demand and money supply. In fact, during the process of decision making, the central bank's behavior and the private sector's behavior are closely examined by looking into the past movements of various key macroeconomic variables (such as M2 and the general price level). When conducting monetary policy, the policymaker always wants to know how much changes in M2 are influenced by present monetary policy (the money supply side) and how much those are merely caused by portfolio shifts in the private sector (the money demand side). If, say, the money demand curve shifts to the right from MD_1 to MD_2 and if the central bank desires to have the money stock at M^* (Chart 7) in order to keep inflation in check, an economic model that explicitly separates the central bank's behavior and the private sector's behavior can undoubtedly aid the policymaker in deciding how the money supply curve needs to be shifted accordingly (from MS_1 to MS_2 in Chart 7). Moreover, such a model allows one to forecast different paths of macroeconomic variables conditional on different policy actions in the future. For example, the Federal Reserve may be interested in deciding whether the federal funds rate in the next two years should be 5 percent or 6 percent or 4 percent. If the economic model distinguishes policy behavior and the private sector's behavior, it can be used to examine how policy actions in the future would lead to different forecasts of the price level, M2, the unemployment rate, and other variables.

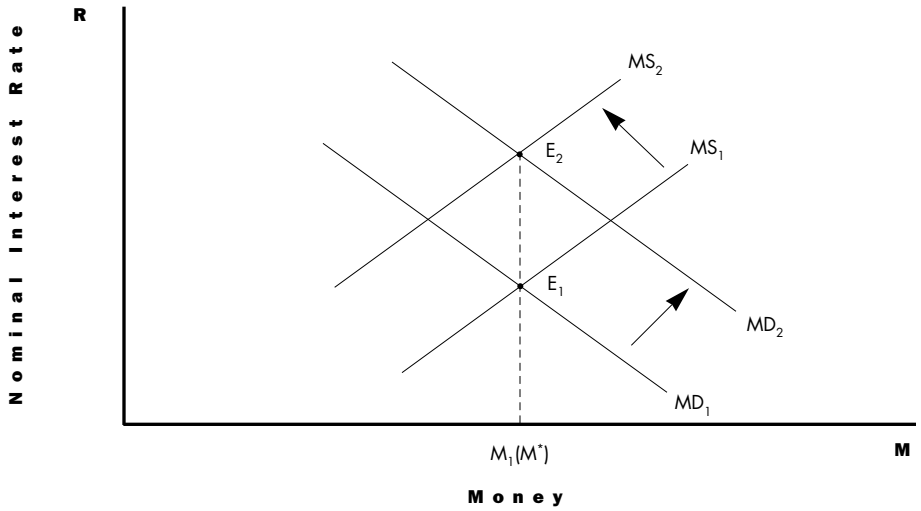
The discussion in this section replies to some prevalent naive thinking about the issue of identifying monetary policy. It reinforces the point that the actual formation of monetary policy in any country is a complicated process. To assess the effect of monetary policy requires understanding the interaction between the central bank's behavior and the private sector's activity.

More on Demand and Supply

So far, the discussion has been concerned with why identification of monetary policy is important in policy analysis. It has not yet answered the question of how we identify monetary policy in practice. How difficult is it to estimate the money demand function (3) and the money supply function (4) or to obtain from the observed data the values of α coefficients in both equation (3) and (4) so that the actual curves of money demand and supply can be plotted? This section turns to this "how" question, which is the essence of identification integral to all empirical study in economics. It begins with the familiar wine market example and then discusses how to estimate the demand and supply of money.

As shown in the dots in Chart 8, the data on the quantity and price of wine are the equilibrium outcome from movements in both demand and supply. These

CHART 7 Monetary Policy Reaction



movements, caused by either the ϵ shocks or other factors such as the government's tax and the consumers' preference for wine over beer, or by both, make the estimation of the demand and supply curves a challenge. A popular approach uses the data to estimate the relationship of quantity to price with the equation

$$q = \gamma_1 p + \epsilon. \qquad (6)$$

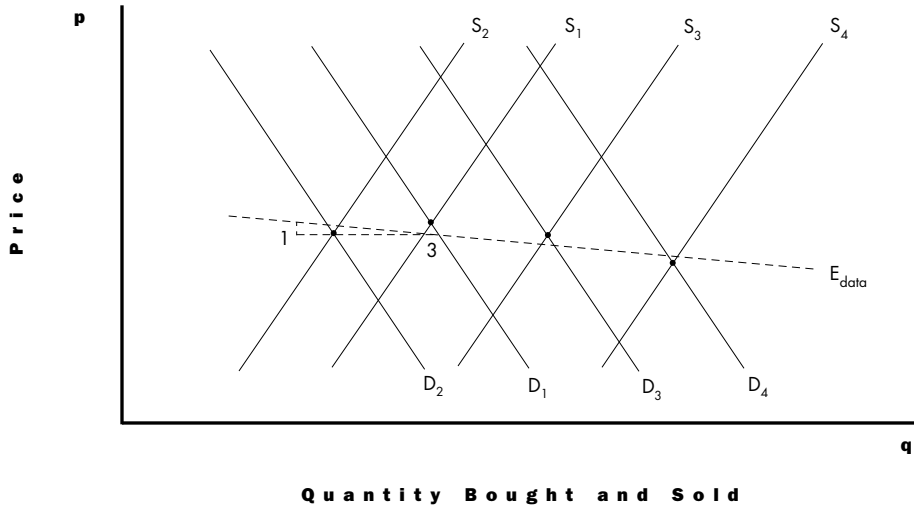
The obvious problem with this approach is that one cannot be sure whether equation (6), after being estimated, is a demand function, a supply function, or a combination of the two. Suppose γ_1 is estimated to be –3 as indicated by the curve E_{data} in Chart 8, implying that the quantity of wine increases by 3 percent when the price falls by 1 percent. This estimated relationship between q and p does not mean that the actual demand for wine (indicated by the curve D in Chart 8) is as elastic as –3. The reason γ_1 and α_1 are not equal is that γ_1 represents the coefficient in the relationship (6) that is directly observed in the data while α_1 represents the coefficient in the wine demand function (1) that is not directly observable.[8] Suppose that the policy analyst mistakenly took the estimated function represented by E_{data} in Chart 8 as the demand function. The analyst would anticipate that the quantity demanded will rise substantially when the price of wine drops. But since the actual demand curve represented by D in the chart is much steeper than the curve E_{data}, the actual demand is less elastic than the one estimated and the quantity actually demanded will not rise so substantially. Then, any conclusions based on this estimate of consumer behavior can be misleading.

The wine market example illustrates that even when demand and supply have simple, uncomplicated relationships, identifying (that is, estimating) each of them is difficult. The same identification problem exists if one tries to estimate both money demand and money supply from the observed data because the data themselves are not sufficient to distinguish supply from demand. Suppose one wants to identify the money supply function. The question is, How can one distinguish the specific behavior of the central bank from the observed data, or how can one figure out the money supply curve in Chart 3? Clearly, one cannot follow the practice of estimating equation (6) in the wine market example and use the data on M and R to estimate such a relationship. Some factor or factors are needed that will shift the money demand curve but not the supply curve. Then, if one traces out the observed data as illustrated by the dots in Chart 9, the time series pattern will precisely reveal the money supply curve that describes the central bank's behavior. Thus, the basic idea of achieving identification is to isolate factors that are in one of the relationships, such as the money demand function (3), but not shared with the others, such as the money supply function (4). An assumption about which factor influences which equation is called an identifying assumption.

Are there such factors? As discussed above, since the central bank is unable to observe the data on output (y) and the general price level (P) within the present month, the set of variables X_s in (4) does not contain y and P. Thus, changes in current output and the current price level serve as the shifters that move the money

8. For similar reasons, γ_1 and β_1 are not equal.

CHART 8 Equilibrium Quantity and Price of Wine



demand curve but not the money supply curve (Chart 9) and help estimate the money supply curve from the data.

The money demand function can also be estimated (identified) in similar spirit. Recall that the information set X_s in the money supply function (4) contains the exchange rate or the commodity price index or both, which are excluded in the money demand function (3). Movements in the exchange rate or commodity price index will shift the money supply curve but not the money demand curve (Chart 10). The money demand curve can then be traced out when there are enough changes in the exchange rate or commodity price index.

Charts 9 and 10 demonstrate the basic idea of achieving identification, but the actual estimation of both the money demand function (3) and the money supply function (4) is far more complicated. When both output (shifting the money demand curve) and the exchange rate (shifting the money supply curve) change at once, the money demand and money supply curves will shift simultaneously (Chart 11).[9] Therefore, one cannot estimate the demand function (3) or the supply function (4) in isolation, as Charts 9 and 10 seem to suggest. Indeed, both functions must be estimated jointly, not in isolation. See Box 2 (page 38) for a discussion of the technical difficulties involved. Nonetheless, the basic idea of achieving identification is clear: conceptually, to do so one needs factors that shift the demand curve independent of the supply curve or vice versa; technically, the demand and supply functions must be estimated simultaneously.

Different Empirical Approaches

Economic research has historically used a variety of empirical approaches to uncover (identify) policy effects from the observed data. To circumvent both the conceptual and technical difficulties in identifying monetary policy, many approaches invoke implausible identifying assumptions. This section considers a few examples or approaches taken from economics journals and argues that assumptions that are convenient for statistical purposes but are not sensible in economics terms are likely to generate misleading results.

Example 1. One traditional approach, which has been exploited at least since Friedman and Schwartz (1963), is to use a single variable (such as a monetary aggregate, an interest rate, or an exchange rate) as an indicator of monetary policy. For example, unpredicted changes in a monetary aggregate—be it reserves or an *M* variable—are often attributed mainly to monetary policy shocks. The common practice is to estimate the relationship of the monetary aggregate to the other variables and then interpret the residuals calculated from such an estimation as policy shocks (for example, Barro 1977 for the United States and Wogin 1980 for Canada).

The assumption underlying this practice is that this relationship represents the central bank's behavior. Unfortunately, this single-equation approach, analogous to the estimation exercise along the line E_{data} in Chart 8, fails to take account of the fact that the data on monetary aggregates are what is observed at equilibrium (the point intersected by the demand and supply curves). Indeed, monetary aggregates such as reserves or M2 are influenced by not only the central bank's behavior but also the demand behavior of other sectors in the economy. Suppose that the equilibrium outcome is such that the monetary aggregate and the interest rate are positively correlated as indicated by the single estimation curve E_{data} in Chart 11. Apparently, it is a mistake to interpret the curve E_{data} as the actual money supply curve *MS*.

The point that the policy behavior inferred through the time series patterns of a single variable can be mis-

CHART 9 Tracing Out the Money Supply Function

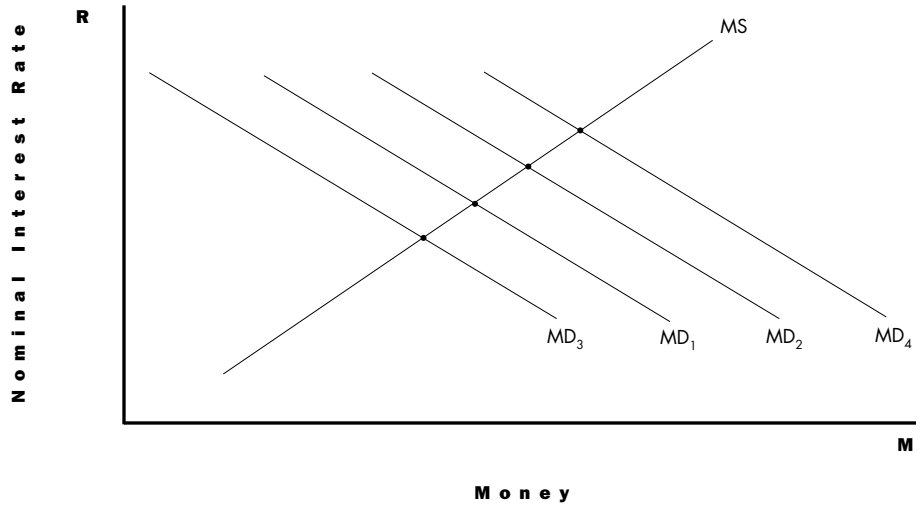
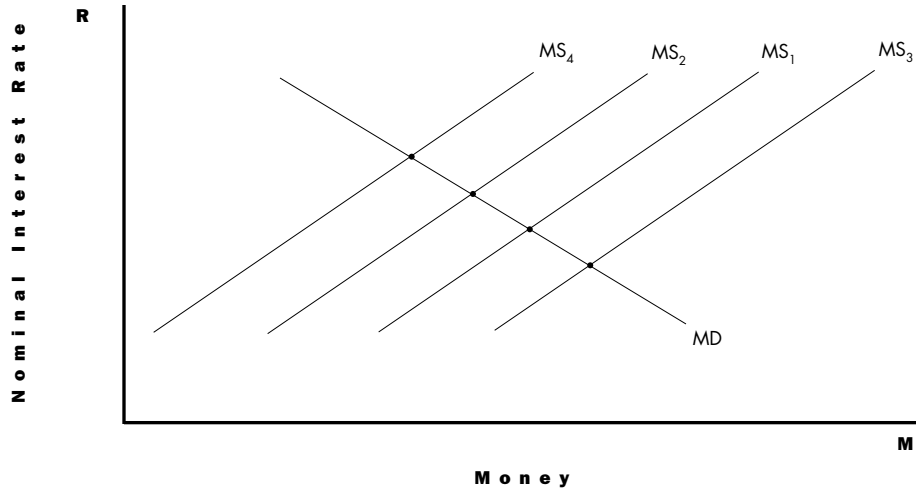


CHART 10 Tracing Out the Money Demand Function



leading was made long ago by Tobin (1970). Tobin presented a dynamic general equilibrium model to show that the same time series evidence used by Milton Friedman and other monetarists could lead to a completely different interpretation of monetary policy effects. Despite Tobin's warning, however, researchers oftentimes continue to use the single-equation approach to modeling monetary policy, at least in part because identifying monetary policy is conceptually difficult and mathematical tools are only now being developed to address the identification issue seriously.[10]

Example 2. In recognition of both the inadequacy of single-equation approaches and the nature of a central bank's reaction to the state of the economy, recent research on identification of monetary policy has developed ways of handling the complex relationships of multiple economic variables (see Box 2). One approach is to include both policy instruments (such as an interest rate) and other macroeconomic variables (such as the general price level) in the same framework (as in Sims 1992, Grilli and Roubini 1995, Eichenbaum and Evans 1995, and Dungey and Pagan 1997). This approach is certainly a

9. The situation is analogous to that depicted in Chart 3.

10. Romer and Romer (1989, 1990), following the spirit of Friedman and Schwartz, invent a single dummy variable indicating the changes in U.S. monetary policy. But as Leeper (1997) forcibly argues, the Romers' dummy variable does not identify the Federal Reserve's behavior.

Empirical Methods

This box, focusing on the case of a small open economy, is largely drawn from Zha (1996). Assume the structural model is of a linear, dynamic form called vector autoregression (VAR):

$$A(L)y(t) = \epsilon(t), \tag{B1}$$

where $A(L)$ is an $m \times m$ matrix polynomial in lag operator L, $y(t)$ is an $m \times 1$ vector of observations of m variables, and $\epsilon(t)$ is an $m \times 1$ vector of i.i.d. structural shocks so that

$$E\epsilon(t) = 0, E\epsilon(t)\epsilon(t)' = I. \tag{B2}$$

The reduced form of (B1) can be obtained by multiplying A_0^{-1} through (B1).

A natural way of estimating the model is to explore the shape of a likelihood function (which describes how likely the model parameters are to lie within a certain range of values) and to obtain the values of parameters that are most likely to occur (the values so obtained are called maximum likelihood [ML] estimates). If the likelihood function is complicated, finding ML estimates may become problematic. For the reduced form of (B1), however, the ML estimates turn out to be simply the ordinary least squares (OLS) estimates in each equation. The OLS estimation is straightforward and can be easily computed using any statistical software package.

To see how the system (B1) can be used to model a small open economy, break (B1) into two blocks—the first block concerns the home (small) economy, and the second block concerns the foreign (the rest of the world) economy. To be specific, let

$$A(L) = \begin{pmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{pmatrix},$$

$$y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix},$$

$$\epsilon(t) = \begin{pmatrix} \epsilon_1(t) \\ \epsilon_2(t) \end{pmatrix}.$$

The matrix A_s is the coefficient matrix of L^s in $A(L)$, where L^s is the lag operator L raised to s power. In most works of the identified VAR literature, the restrictions are imposed only on A_0—the contemporaneous coefficient matrix. The ML estimates of A_0 depend only on the estimated covariance matrix ($\hat{\Sigma}$) of reduced-form residuals; this can be easily seen by writing out the concentrated likelihood function of A_0 (see Sims and Zha 1995 for details):

$$|A_0|^T \exp\left[-\frac{T}{2} trace\left(\hat{\Sigma} A_0' A_0\right)\right]. \tag{B3}$$

Because of (B3), obtaining the estimates of and inference about model parameters is in principle straightforward (Sims and Zha 1995), and when A_0 follows a successive representation, the estimation is straightforward even in practice (Sims 1980).

If the small open economy framework is taken seriously, one will impose the restriction that $A_{21}(L) = 0$, meaning that the small country takes changes in foreign economic conditions as given or exogenous. This small-economy restriction makes the easily implemented procedure developed by Sims and Zha (1995) invalid, mainly because the concentrated likelihood (B3) no longer holds. In principle, various iterative procedures can be used. For example, one begins with the unrestricted $\hat{\Sigma}$ to solve the ML estimate of A_0 with the restriction $A_{0_{21}} = 0$ imposed. The estimates of other structural parameters (A_s, $s \geq 1$) can then be recovered, and a new reduced form covariance matrix is accordingly formed.[1] Use this new matrix to replace the previous $\hat{\Sigma}$ and repeat the procedure until $\hat{\Sigma}$ converges.

Since the size of a small open economy model is typically large relative to closed economy models, this iterative procedure not only is cumbersome but can be computationally prohibitive as well, especially when one computes the inference of the ML estimates. Consequently, previous researchers have not accounted for the small open economy features in their models. The method developed in Zha (1996), which allows for more general cases than the small open economy example here, provides a practically feasible way of obtaining the ML estimates as well as their inference.

1. The details of how the estimates are recovered are discussed in Zha (1996). The idea of this iterative procedure is also mentioned in Dias, Machado, and Pinheiro (1996).

CHART 11 Simultaneous Changes in Money Demand and Money Supply

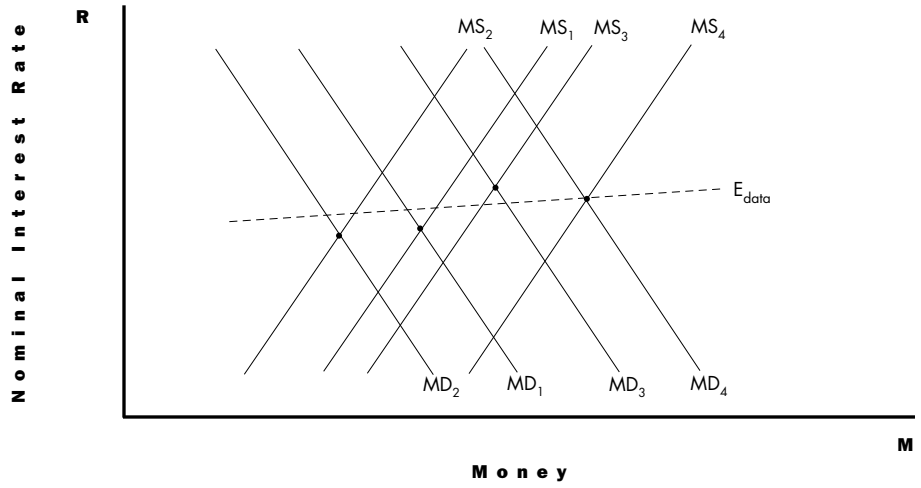
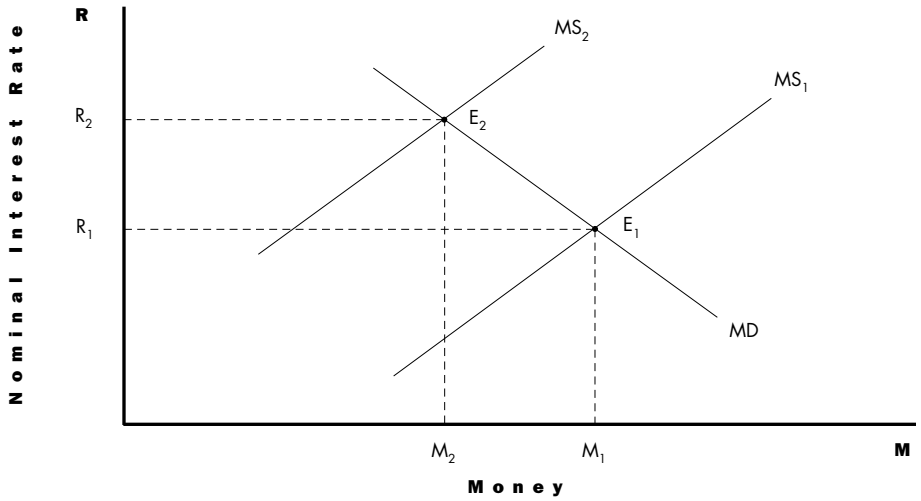


CHART 12 Effect of Contractionary Shock (E_{MS})



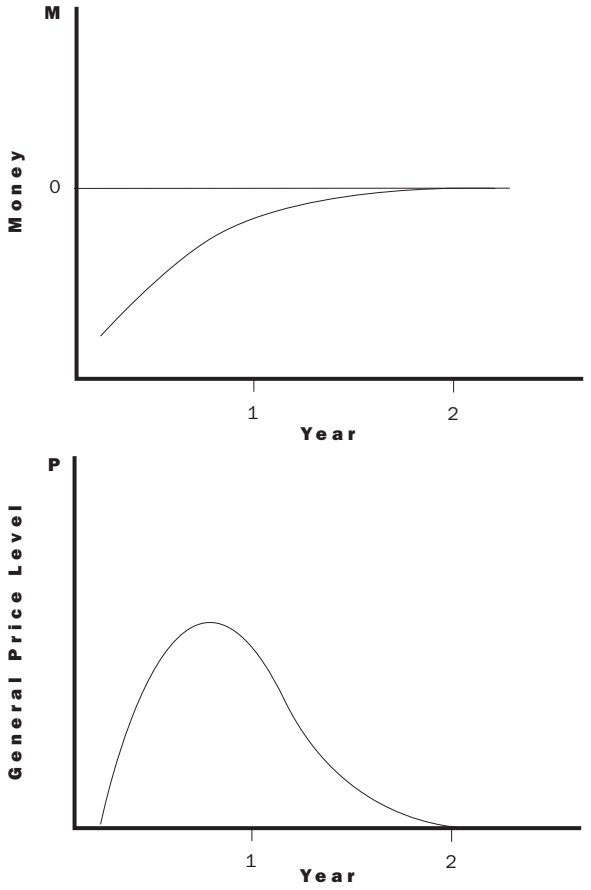
considerable improvement over that outlined in Example 1. Nonetheless, all these works suffer from a common problem: they make identifying assumptions that seem implausible in the description of a central bank's behavior.

As explained above, identifying assumptions are those that help distinguish different behaviors (for example, demand and supply) in the actual economy. They are necessary because, analogous to the example of demand and supply, one needs some factors that shift only the supply curve in order to identify the demand function (Chart 10) and other factors that shift only the demand curve in order to identify the supply function (Chart 9). While the aforementioned works use assumptions that are convenient for statistical computations, the important difference in the approach called for in the previous sections is that it argues for economically sensible assumptions.

Specifically, the common assumption used in the works cited is that different behaviors follow successive relationships. Although the successive assumption makes it convenient to estimate the model (see Box 2), it seldom represents the structure of the actual economy. For example, in Eichenbaum and Evans (1995) the money stock M influences the interest rate R contemporaneously but not vice versa, an assumption that essentially takes α_3 in the money supply equation (4) to be zero, implying that the money supply is perfectly inelastic (Chart 5).[11] In their study of the U.S. economy, for example, Eichenbaum and Evans use nonborrowed reserves as M and the federal

11. The money demand function is not explicitly specified in their papers.

CHART 13
Perverse Dynamic Responses of Price after Monetary Contraction: The Price Puzzle



funds rate as R. Thus, an inelastic money supply ($\alpha_3 = 0$) means that the Federal Reserve does not respond, within the month, to fluctuations in the federal funds rate. In fact, the Federal Reserve frequently influences the federal funds rate in pursuit of its objective, so the assumption that money supply is perfectly inelastic seems at odds with the Federal Reserve's targeting of the federal funds rate. It is therefore not surprising that this extreme assumption would lead to results that are inconsistent with views widely held by both policymakers and economists (Gordon and Leeper 1994 and Leeper 1995).

Before reviewing an example of these inconsistent results, it is important to explain the concept of contractionary monetary policy shock that is often used in economics journals. Recall that this article uses the phrase *policy shocks*—ϵ_{MS} in the money supply equation (4)—to describe unpredicted shifts in monetary policy. Thus, the shock ϵ_{MS} in equation (4) is said to be contractionary if it shifts the money supply curve to the left (from MS_1 to MS_2 in Chart 12), moving the equilibrium outcome from point E_1 to point E_2. The word *contractionary* is adopted because, subsequent to this shock, the money stock M contracts from M_1 to M_2 while the interest rate R rises from R_1 to R_2.

Now, to present an example, consider one of the firmly established views in policy analysis: the price level falls after an unpredicted contraction in monetary policy. When one uses Eichenbaum and Evans's successive assumption to model several industrial countries such as the United States, Japan, and Germany, the model generates the inconsistent result (often termed the price puzzle) that the price level would rise, not fall, in response to a contractionary monetary policy shock (Chart 13).[12] If the model is intended to be useful for policy decisions, such a puzzle is indeed troublesome because it implies that monetary policy must expand the money stock (or lower the federal funds rate) in order to lower inflation. Would one recommend such a policy? Does anyone really believe that inflation will fall if the central bank increases the supply of money (or lowers interest rates)?

When a model produces inconsistent results such as the price puzzle, one needs to examine carefully the underlying (identifying) assumptions to see if they make good economic sense. If a central bank reacts quickly to changes in the interest rate, it makes no economic sense to assume that the interest elasticity α_3 in the money supply equation (4) is zero. If one insists on a successive representation by letting α_3 be zero, equation (4) is then no longer the policy reaction function (or the money supply function).

Example 3. The above example suggests that a reasonable identification of the central bank's behavior inevitably leads to a breakdown of the successive representation commonly used in economics journals. A recent work of Cushman and Zha (1997) argues for a better representation of policy's systematic behavior and makes progress in the specification and estimation of behavioral relationships. In that study, both the money demand equation and the money supply equation are in the same form as equations (3) and (4) in this article. Using Canada as a study case, the paper devotes special attention to Canada's relationship with the U.S. economy and the systematic component ($\alpha_3 R + \alpha_4 X_s$) in the policy reaction function (4). In particular, the set X_s contains a wide variety of macroeconomic variables to which the Bank of Canada would react. Some information, such as output and the general price level in both Canada and the United States, is not readily available to the Bank of Canada in a timely fashion (because data such as industrial output and the consumer price index for a given month are not released until after the end of the month). These pieces of information are therefore excluded from X_s in the money supply function (4). The Bank of Canada, however, can react quickly to changes in other key macroeconomic conditions—the exchange rate, the U.S. interest rate, and commodity prices—for which data are available daily. These economic conditions convey infor-

mation about the current state of both the Canadian economy and the U.S. economy, about possible actions in U.S. monetary policy, and about future inflation.

The estimated money demand and money supply curves by Cushman and Zha (1997) are depicted in Chart 14. The money supply is almost inelastic to the domestic interest rate but, as shown in Chart 15, very elastic to the exchange rate. This condition implies that the Bank of Canada, unlike the Federal Reserve, responds mainly to changes in the exchange rate rather than to the domestic interest rate. Evidently, systematic behavior of central banks may differ across countries (such as the United States and Canada). For the U.S. economy, it is a mistake to assume the interest elasticity α_3 in the money supply function (4) to be zero because the Federal Reserve targets the federal funds rate. In other words, assuming $\alpha_3 = 0$ may be expedient statistically, but it yields results that are not sensible. For the Canadian economy, it is a mistake to assume the exchange elasticity in the money supply function to be zero because, as shown in Chart 15, the Bank of Canada responds to changes in the exchange rate. Indeed, if the exchange rate elasticity were assumed to be zero, the price puzzle, which does not exist in Cushman and Zha's original model, would be present.

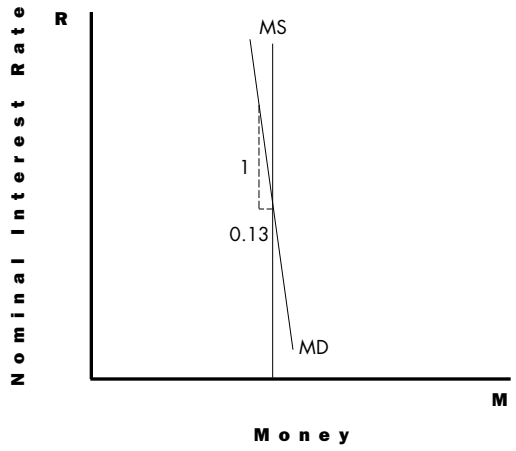
The above example of identifying Canadian monetary policy shows the importance of using sensible identifying assumptions even though such assumptions may raise empirical difficulties. Identifying the U.S. monetary policy involves a similar task of separating the Federal Reserve's behavior from the private sector's behavior. Leeper, Sims, and Zha (1996) discuss how difficult it is to achieve reasonable identification of U.S. monetary policy. Understanding each country's relationship with the rest of the world and each central bank's systematic behavior is a necessary step when one makes identifying assumptions.

This section reviews several identification approaches used in policy analysis. Some, such as the single-equation approach, fail to separate policy's systematic response to the state of the economy from the response of the economy to policy (supply from demand). Those that attempt such separation have often imposed extreme assumptions that would lead to inconsistent or puzzling results. All the examples discussed echo the same message: avoiding extreme or unreasonable identifying assumptions when modeling monetary policy in a given country is crucial for eliminating puzzling results, achieving correct identification, and producing sensible monetary policy analysis.

Conclusion

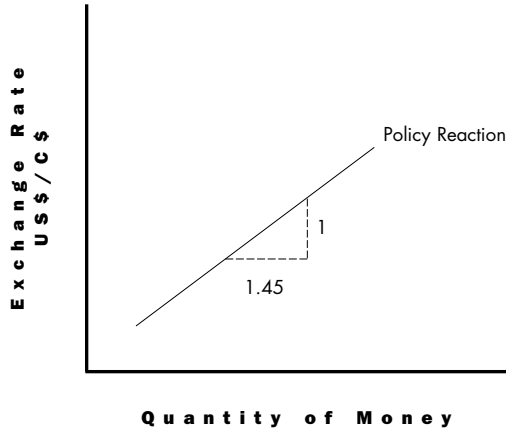
The monetary policy reaction in any actual economy is complicated, and "the policy framework is a pragmatic one. There are no simple rules" (Duguay and Poloz 1994, 197). The discussion here establishes the importance of identifying monetary policy, explains the difficulties involved in identification, cautions about the potential danger of making extreme assumptions about the behavioral relationships, and sheds some light on the progress there has been toward adequately identifying monetary policy (as in Cushman and Zha 1997). In particular, the article uses the simple examples of demand and supply to illustrate how quickly the difficulty in identification of monetary policy can become overwhelming if one wishes to separate the central bank's behavior from

CHART 14
Estimated Money Demand and Money Supply Functions



Source: Adapted from Cushman and Zha (1997).

CHART 15
Relationship between Money Supply and Exchange Rate



12. The author thanks Roberto Chang for suggesting such an exercise.

others' behavior in the economy. The difficulty contributes to the disparity and uncertainty in economists' views on the effects of monetary policy.

The essential point is that because of the complexity inherent in monetary policy reaction unique to different countries, an economic model usable for policy analysis in a given country requires both cogitable reasoning conceptually and serious effort empirically. Notwithstanding significant progress in both theory and econometrics, the gap between economic theory and empirical observations is still wide because theoretical models have not yet produced the same time series pattern of macroeconomic variables as those that characterize the actual economy. The challenge in future research is to narrow the gap and move toward a good economic model usable for a discussion of actual policy effects in different countries.

REFERENCES

Barro, Robert J. 1977. "Unanticipated Money Growth and Unemployment in the United States." *American Economic Review* 67 (March): 101-15.

Barro, Robert J., and Robert F. Lucas. 1994. *Macroeconomics*. Burr Ridge, Ill.: Richard D. Irwin, Inc.

Batten, Dallas S., Michael P. Blackwell, In-Su Kim, Simon E. Nocera, and Yuzuvu Ozeki. 1990. "The Conduct of Monetary Policy in the Major Industrial Countries: Instruments and Operating Procedures." International Monetary Fund Occasional Paper 70, July.

Cooley, Thomas F., and Gary D. Hansen. 1995. "Money and the Business Cycle." In *Frontiers of Business Cycle Research*, edited by Thomas F. Cooley, 175-216. Princeton, N.J.: Princeton University Press.

Cushman, David O., and Tao Zha. 1997. "Identifying Monetary Policy in a Small Open Economy under Flexible Exchange Rates." *Journal of Monetary Economics* 39 (June, forthcoming).

Dias, Francisco C., Jose A.F. Machado, and Maximiano R. Pinheiro. 1996. "Structural VAR Estimation with Exogeneity Restrictions." *Oxford Bulletin of Economics and Statistics* 58 (May): 417-22.

Dornbusch, Rudiger, and Paul Krugman. 1976. "Flexible Exchange Rates in the Short Run." *Brookings Papers on Economic Activity*, no. 3:537-75.

Duguay, Pierre, and Stephen Poloz. 1994. "The Role of Economic Projections in Canadian Monetary Policy Formulation." *Canadian Public Policy* 20, no. 2:189-99.

Dungey, Mardi, and Adrian Pagan. 1997. "Towards a Structural VAR Model of the Australian Economy." Australian National University, unpublished manuscript.

Eichenbaum, Martin, and Charles Evans. 1995. "Some Empirical Evidence on the Effects of Shocks to Monetary Policy on Exchange Rates." *Quarterly Journal of Economics* 110 (November): 975-1009.

Friedman, Milton. 1992. *Money Mischief: Episodes in Monetary History*. New York: Harcourt Brace Jovanovich.

Friedman, Milton, and Anne J. Schwartz. 1963. *A Monetary History of the United States, 1867-1960*. Princeton, N.J.: Princeton University Press.

Gordon, David B., and Eric M. Leeper. 1994. "The Dynamic Impacts of Monetary Policy: An Exercise in Tentative Identification." *Journal of Political Economy* 102 (December): 1228-47.

Grilli, Vittorio, and Nouviel Roubini. 1995. "Liquidity and Exchange Rates: Puzzling Evidence from the G-7 countries." Yale University, unpublished manuscript.

Leeper, Eric M. 1992. "Facing Up to Our Ignorance about Measuring Monetary Policy Effects." Federal Reserve Bank of Atlanta *Economic Review* 77 (May/June): 1-16.

———. 1995. "Reducing Our Ignorance about Monetary Policy Effects." Federal Reserve Bank of Atlanta *Economic Review* 88 (July/August): 1-38.

———. 1997. "Narrative and VAR Approaches to Monetary Policy: Common Identification Problems." *Journal of Monetary Economics* 40 (forthcoming).

Leeper, Eric M., Christopher A. Sims, and Tao Zha. 1996. "What Does Monetary Policy Do?" *Brookings Papers on Economic Activity*, no 2:1-78.

McCallum, Bennett T. 1989. *Monetary Economics: Theory and Policy*. New York: Macmillan.

Mishkin, Frederic S. 1992. *The Economics of Money, Banking, and Financial Markets*. 3d ed. New York: Harper Collins.

Romer, Christina D., and David H. Romer. 1989. "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." In *NBER Macroeconomics Annual 1989*, edited by Olivier J. Blanchard and Stanley Fischer, 121-70. Cambridge, Mass.: MIT Press.

———. 1990. "New Evidence on the Monetary Transmission Mechanism." *Brookings Papers on Economic Activity*, no. 1:149-98.

Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica* 48 (January): 1-48.

———. 1992. "Interpreting the Macroeconomic Time Series Facts: The Effects of Monetary Policy." *European Economic Review* 36 (June): 975-1000.

SIMS, CHRISTOPHER A., AND TAO ZHA. 1995. "Error Bands for Impulse Responses." Federal Reserve Bank of Atlanta Working Paper 95-6, September.

TOBIN, JAMES. 1970. "Money and Income: *Post Hoc Ergo Propter Hoc*?" *Quarterly Journal of Economics* 84 (May): 301-17.

WOGIN, GILLIAN. 1980. "Unemployment and Monetary Policy under Rational Expectations: Some Canadian Evidence." *Journal of Monetary Economics* 6 (January): 59-68.

ZHA, TAO. 1996. "Identification, Vector Autoregression, and Block Recursion." Federal Reserve Bank of Atlanta Working Paper 96-8, August.

International Settlements: A New Source of Systemic Risk?

ROBERT A. EISENBEIS
The author is senior vice president and director of research at the Atlanta Fed. *He thanks Clyde Farnsworth, Craig Furfine, Diana Hancock, Pat Parkinson, and Alice P. White for helpful comments.*

THE VERY REAL SIGNIFICANT SOCIAL COSTS OF SYSTEMIC RISK HAVE LONG SERVED AS AN IMPORTANT RATIONALE FOR A FEDERAL PRESENCE IN THE DOMESTIC PAYMENTS SYSTEM.[1] RECENT MARKET DEVELOPMENTS HAVE HEIGHTENED CONCERNS ABOUT THE POTENTIAL FOR SYSTEMIC RISK IN THE PAYMENTS SYSTEM. FIRST, THE SHEER GROWTH IN LARGE-VOLUME PAYMENTS HAS RAISED THE POTENTIAL COSTS SHOULD A NUMBER OF INSTITUTIONS FAIL. SECOND, TECHNOLOGY AND TECHNOLOGICAL CHANGE SEEM TO BE REDEFINING THE KINDS OF TRANSACTIONS TAKING PLACE AS WELL AS INCREASING THE SPEED WITH WHICH THESE TRANSACTIONS CAN BE COMPLETED AND FUNDS TRANSFERRED. FOR EXAMPLE, BOTH COMPUTER AND OPTIONS PRICING TECHNOLOGIES NOW PERMIT THE UNBUNDLING, RESTRUCTURING, AND CREATION OF TRANSACTIONS (SUCH AS SWAPS AND DERIVATIVES) WHOSE RISKS, LEGAL STATUS, AND RELATED CHARACTERISTICS ARE JUST NOW BEGINNING TO BE UNDERSTOOD.

A third dynamic behind the increased concern about systemic risk in the payments system is the globalization of financial markets, which is tying economies and markets together in ways that introduce additional issues about the mechanisms by which traditional clearing (the notification and transfer of documents and orders to purchase and sell assets) and settlement (the transfer of final payment) take place. Finally, the fears about the supposed potential for systemic risk associated with clearing and settlement loss have been given greater credence by the lack of internal controls within major institutions, which have been exposed by the actions of rogue traders in Kidder, Bankers Trust, and Barings (see Edwards 1996). As the problems in these institutions have been unwound, greater appreciation has emerged of just how complex and segmented the institutional arrangements for clearing and settling transactions have become.

Despite the fact that securities, futures and options, and derivatives are increasingly cleared under a variety of institutional arrangements, final settlement usually takes place in the interbank market. In general, clearing of transactions—be they securities, derivatives, or other assets—is almost exclusively done by the private sector while settlement can take place in the wholesale banking sector or through central banks (see BIS 1997b).[2] Markets have become tiered as more and more transactions are cleared through several layers of institutions before they are ultimately settled (see Corrigan 1990). Equally important, the introduction of new instruments, such as swaps, collateralized mortgage obligations, and off-exchange derivatives, and their associated methods for transferring cash flows and settlement relationships have resulted in seemingly unrelated markets and institutions being linked together in ways that both create and may de facto transfer risks from one market to another. Increasingly, these transactions and markets are assuming an international dimension that can also have significant domestic market implications (see BIS 1997b).

This article examines whether internationalization has changed the nature of and potential vulnerability of the financial system to systemic risks and looks at a method to mitigate them. The Lamfalussy Report, which examined and proposed standards for payments systems settlement and risk control features, indicates that system vulnerability is critically linked to the length of time that participants are exposed to credit and liquidity risks (see BIS 1993b). The analysis presented here suggests that regulatory and legal structures can also have systemic risk dimensions. As to the fundamental question of whether new risks are being introduced, the answer seems to be no. Moreover, recent institutional and regulatory developments may act to reduce the potential scope and the size of these risks and limit implicit taxpayer liabilities should these risks be realized.

Risks in Payments

Regardless of the institutional arrangements, there are four generally accepted generic types of payments system risks that have been identified and have been the focus of much attention. These include operational, legal, credit, and liquidity risks (see Eisenbeis 1995 or BIS 1997b). While it is easy to differentiate these risks conceptually, in reality they tend to be interrelated. The realization of one can lead to occurrences of the others, and this dynamic has not changed with the evolution of the new instruments and markets just described. These interrelationships among risks can be illustrated by considering credit risk, which arises when the purchaser of an asset defaults by failing to settle any or all of its obligations. Credit risk arises as a logical by-product of separating the clearing and settlement functions, which under current institutional arrangements nearly always involves an extension of temporary credit.

Credit risk is a function of the potential loss exposure when a buyer initiates a transaction ordering its bank to transfer funds but then cannot make payment without going into an overdraft situation. The buyer's bank, which is attempting to settle on behalf of the buyer, is faced with essentially three alternatives. First, it can provide credit to the customer until funds are received. Second, the transaction can be canceled, or the bank can complete the transaction itself. If the buyer's bank takes the place of the customer and completes the transaction, it may then take possession of the goods or asset (or any other of the customer's available collateral) and proceed to unwind the transaction. Finally, in the extreme, the buyer's bank can default on its own obligation to settle if the time for settlement has not yet occurred.

Recent market developments have heightened concerns about the potential for systemic risk in the payments system. . . . As to the question of whether new risks are being introduced, the answer seems to be no.

If the buyer has good collateral and a sound credit rating, then extension of credit may be the best alternative. Canceling the transaction may not be an option, especially when delivery of the good or service has already taken place and there is no available collateral.

Settlement failure in this example could be controlled if the buyer's bank were to put a hold on the buyer's funds at the time payment is initiated, collateralizing the transaction. Organized futures markets effectively accomplish this control through the use of margins, mark-to-market accounting, and settlement requirements. For good customers, however, collateralization may not be necessary, practical, or efficient, especially if both the probability of default and the expected loss are small relative to the bank's resources. The lack of a hold or similar type of collateral policy illustrates that an institution's vulnerability and exposure to credit risk often results from the underlying conventions, practices, and

1. See Benston and Kaufman (1995) for a review of the evidence on fragility and systemic risk.
2. For a discussion of the risks and recent developments in exchange-traded derivatives markets, see BIS (1997a).

structures of the markets involved rather than from the realization of performance risks associated with the underlying projects and investments.

As markets have become increasingly global, differences in timing and clearing and settlement conventions and differences in bankruptcy laws can add important temporal and other dimensions to credit risks not always found in domestic markets. This consideration was clearly demonstrated in 1974 when Herstatt Bank failed and was closed by German authorities. Herstatt had entered into agreements to exchange deutsche marks for dollars. The mark leg of the transaction was settled, but the dollar portion was not settled in New York at the time Herstatt was closed since the deadline on CHIPS (Clearinghouse Interbank Payments System) for final settlement was approximately 4:30 P.M. eastern standard time. This difference in settlement times for the two sides of the transaction left the counterparties to the foreign exchange transaction thinking that they had more funds than they did. When the dollar transactions failed to settle, the result was large losses to the U.S. counterparties. This temporal dimension to credit/systemic risk has come to be known as Herstatt risk and can be very large.[3]

As markets have become increasingly global, differences in timing and clearing and settlement conventions and differences in bankruptcy laws can add important temporal and other dimensions to credit risks.

A more recent example of this type of event is the closing of the Bank of Credit and Commerce International (BCCI) in 1991. The Industrial Bank of Japan had paid 44 billion yen into BCCI's branch in Tokyo, for which payment was to be received in New York from BCCI's New York branch. When BCCI was closed, the dollar portion of the transaction was never completed, and Industrial Bank of Japan became a creditor for $30 million.

These examples may at first look like ordinary credit risk in that loss exposure resulted from the inability of Herstatt and BCCI to pay. But the incidence of the losses and ultimate position of the banks' creditors was determined by both home country laws and the intervention policies of their regulatory authorities, whose actions usually cannot be easily predicted or priced.[4] The losses to dollar counterparties in the Herstatt case were the consequence of the timing of the closure of the institution rather than the realization of estimable default risk. Had the German authorities waited until the U.S. dollar markets had settled, then the losses to those expecting dollar transfers would not have occurred and the risks would not have been realized. Such exposure is better characterized as settlement uncertainty rather than settlement or credit risk since it is not possible to estimate reliably and cost out the implications associated with the vagaries of sometimes untested statutes governing transactions and of regulatory actions and policies. Note, too, that although the size of the losses may not have been affected by the closure timing, the distribution of the losses was significantly affected by legal structures and governmental action. At the same time, numerous initiatives by governmental bodies such as the Federal Reserve and the Bank for International Settlements (BIS) are continually seeking to identify and institute policies to limit these problems (see Bank of England 1994 and BIS 1989, 1990, 1993a-c, 1997a, b).

Herstatt-type risk can also be involved solely in dollar clearing systems. In Asia the Chase Manhattan Bank operates a dollar clearing and settlement service through its Tokyo branch. The system provides a limited overdraft facility and promises finality of settlement guaranteed by Chase Manhattan. Participants are permitted to settle overdrafts in New York across the Tokyo/New York business day. Furthermore, Tokyo balances at the end of the day may be transferred to New York through the New York offices of Chase or Tokyo banks or through CHIPS. In this system any problems that may arise in this satellite settlement and clearing system quickly have the potential to transmit liquidity and credit risk from Asia to New York, and ultimately to the Federal Reserve, if it affects CHIPS, Chase, or significant New York correspondents. A failure to settle in New York on payments guaranteed in Japan by Chase creates a form of Herstatt risk that would end up having to be resolved in New York. At present, concern about such clearing and settlement systems stems from the sheer size of the potential losses rather than from a true understanding of well-articulated scenarios on how the risks would be played out.[5]

Sources of Payments Uncertainty

Whenever clearing and settlement of financial assets are separated in the international arena, a given country's rules usually establish the exact point in time that a transaction has been completed and the obligation satisfied. The issue centers on transaction finality and the legal criteria for when debts are discharged and who bears the losses in the event of default. Finality usually occurs when the party selling the asset actually has "good funds" and the transaction is both irrevocable and unconditional. Importantly, since many central bank settlement systems can involve the extension of intraday credit, finality may or may not correspond to the time that the buyer actually settled. For example, because Fedwire provides finality as a matter of Federal Reserve policy, acceptance of a payment order

carries with it the "guarantee" of good funds to the receiver and also discharges the debt, since the sender's reserve account is debited and the receiver's bank account is credited, even though the sender's bank may default on the settlement of its reserve account with the Fed at the end of the day. When the settling institutions are located in two separate countries, the specifics of the transactions in terms of settlement, discharge of debt, and so forth may sometimes be governed by the laws of two separate countries and, if transactions involve clearinghouses, the laws where they are located as well.

The legal status of claims can quickly become very murky when the problems involved in settlement failures in cross-border bilateral and multilateral netting arrangements are examined, especially those transactions involving forward-dated contracts in foreign exchange, derivatives, and other cross-border markets (see BIS 1997b). Under netting systems, debt and credit orders are cumulated, and only the net difference is transferred at an agreed-upon time. This procedure contrasts with real-time gross settlement systems (RTGS), which continually process and settle transactions as the orders are received. Final disposition of the liability under netting systems depends critically on the legal rules governing the disposition of debts and transactions in the event of a default or bankruptcy.

As an example, if two institutions have entered into a bilateral netting arrangement, then completion of all the transactions subject to the arrangement is contingent on settlement of the net position. Should one of the parties fail to settle because of a bankruptcy, all the gross transactions subject to netting may have to be undone. The determining factor here depends upon the legal rules affecting the markets in which the transaction was settled. Since the legal rules may differ according to where settlement takes place, and this location may be beyond the receiver's control, settlement uncertainty may exist.

The exact status of cross-border transactions, therefore, is determined by several sets of laws. These include the laws governing bilateral netting arrangements and those governing the particular settlement market involved as well as the bankruptcy provisions and other related laws of the country of the failed institution (or the laws of the resident country if the transaction is recorded on the books of a branch of the failed bank). For example, netted transactions may or may not be regarded as discharged. The bankruptcy court with jurisdiction over the transactions may decide to unbundle netted transactions, demanding payment for debts owed and disavowing liabilities to creditors. In addition, country bankruptcy law may give creditors the right to offset their liabilities to a failed entity against their claims on that entity. Thus, debts owed on foreign exchange may be discharged with debts on securities, loans, or any other assets. Not only do the bankruptcy laws affect the size of the losses but also the way in which the losses may be apportioned across various creditors.

Final disposition of the liability under netting systems depends critically on the legal rules governing the disposition of debts and transactions in the event of a default or bankruptcy.

The legal situation in multilateral netting arrangements introduces complexities several orders of magnitude greater than those affecting bilateral arrangements. There is considerable variation across countries in treatment of transactions, and thus uncertainty exists about how particular bankruptcies will be treated. The key point is that this legal uncertainty often can undermine the efficiency of bilateral and multilateral netting arrangements and creates the very real possibility that systemic risks could be heightened rather than reduced when the laws governing netting are not uniform across countries. Because these legal uncertainties complicate

3. Notice, however, that it may be a misnomer to call this type of event risk, at least in the Frank Knight ([1921] 1971) tradition; see also Hu (1994). The incidence of loss resulted from the German governmental action, which seems almost impossible to assign a probability to, and hence may be better characterized as regulatory uncertainty. See BIS (1996) for a comprehensive discussion of risks in foreign exchange markets and efforts that both private- and public-sector entities have made to identify, monitor, and control these risks.

4. Bankruptcy statutes can clearly affect the distribution of claims as well. For example, some countries have what is known as a zero-hour rule, which means that transactions taking place after the time the institution is legally closed are regarded as invalid and will be unwound.

5. See, for example, General Accounting Office (1994). An exception is Edwards (1996), who describes the possible paths of a breakdown in derivatives markets. He describes a scenario in which an end-user fails to meet its obligations as a counterparty. This failure in turn brings down a major dealer, thereby spilling over to both other counterparties and dealers. These disruptions are then transmitted to other markets as uncertainty both raises contract prices and leads to reluctance to enter into contracts. There are then price breaks, credit disruptions, falling asset prices, and, ultimately, real effects. Edwards analyzes the likelihood that such a scenario would be realized and concludes that true dealer credit exposures are small and substantially smaller than their exposure on loans and other assets.

assessment of the likely outcome of a default scenario for many transactions, authorities have paid great attention to putting transactions on a common legal basis and, as discussed in the next section, some nations have moved to establish real-time gross settlement as the basis for clearing and settlement.

Responses to Uncertainty

Both private- and public-sector entities have responded to the increased uncertainties, market risks, and evolving market technologies in many interesting ways. The responses affect contract design and the micromarket structure of exchanges and their rules governing transactions. They have given rise to proposals to change laws governing transactions and suggestions to increase governmental cross-border cooperation in financial rules, regulation, and supervision as well as changes in the structural design of transfer systems.

Systems are evolving toward real-time gross settlement, which contains inherent incentives for institutions engaged in offering payments services to price and monitor their exposures.

Given the complexity of financial transactions and their interrelationships, measuring, monitoring, and pricing what institutions' true risk exposures to each other are and how these risks flow directly and indirectly through relationships with related customer groups is difficult. For example, Customer X may have several relationships with its primary bank (Bank A). These might include a loan, a swap, a deposit account, and several foreign exchange transactions. Customer X may also have similar relationships and transactions outstanding with Bank B. In addition, Bank A may also have made loans in the form of advancing federal funds to Bank B. If Customer X fails, the entirety of its net position with Bank A across all the relationships and transactions represents its net direct risk exposure. Bank A may also be indirectly exposed through Bank B if the customer's default causes Bank B to default on its federal funds obligations to A's primary bank.

Measuring and monitoring these interrelated exposures across the world, across different markets and time zones, is a truly daunting modeling and monitoring problem. It is made even more so by the dynamic and continual evolution of new instruments and markets.

Central bank and market responses to these challenges have been to substitute rules and other mechanisms to control customer risk-taking incentives. A number of control mechanisms have been designed to limit uncertainty and to provide incentives for member institutions to control their own risk exposures. These include maintenance of adequate capitalization, reliance upon contract design to allocate risk and losses, collateralization of transactions, use of outside guarantees and bonding, pricing, imposition of system membership requirements, and self-imposed (and system-mandated) caps and other limits on risk exposure to individual and related parties. For example, in the United States, the Federal Reserve imposed limits in 1986 on participating banks' net exposures across Fedwire and CHIPS as well as bilateral limits on exposures to individual participants. Collateralization of certain positions is also required, and the system charges for intraday credit that is extended.

Contracting activities also have focused on apportioning risks, defining performance, and allocating losses among participants in a payments system or exchange in the event that a default occurs. Because of the difficulties in continuously measuring and monitoring total risk exposure to individual system members, caps on the amount of exposure with any member have been imposed, and the system imposes a similar total cap across all system members. In the case of the U.S. CHIPS system (which is not a real-time gross settlement system), participants require same-day settlement, engage in real-time monitoring, have established limits on exposures, have required collateral to cover the largest two exposures, and have instituted a loss-sharing arrangement.[6] System members also impose various types of membership and participation requirements, such as the maintenance of minimum capital requirements.

It has also been recognized that accounting rules—such as mark-to-market requirements—can affect the ease of information transfer and reduce monitoring costs. Such rules have been especially widely used in the case of futures, options, and commodities exchanges.

Finally, systems are evolving toward real-time gross settlement despite the supposed efficiency advantages of netting arrangements. Real-time gross settlement systems require those engaging in payments activities to collateralize payments fully as they are initiated. The benefits of doing so are weighed against the costs of uncertainty and credit risks. Such systems contain inherent incentives for institutions engaged in offering payments services to price and monitor their exposures. Furthermore, real-time gross settlement reduces risk exposure by limiting the duration of both credit and liquidity risk.

The first real-time gross settlement system was the Federal Reserve's Fedwire (see BIS 1997b). By the end of the 1980s, six of the Group of 10 countries had instituted RTGS systems. As the European Union proceeds, Lamfalussy Standards (BIS 1993b) specify that RTGS systems must be in place, and the union's umbrella set-

TABLE 1
Features of Selected Funds Transfer Systems

Country	System (Planned)	Type	Date	Central Bank Daylight Credit
Belgium	ELLIPS	RTGS	1996	Yes
Canada	IIPS	Net	1976	
	(LVTS)	Net	1997	
France	SAGITTAIRE	Net	1984	
	(TBF)	RTGS	1997	Yes
Germany	EIL-ZV	RTGS	1987	Yes
	EAF2	Net	1996	
Italy	BISS	RTGS	1989	
	(BI-REAL)	RTGS	1997	Yes
	ME	Net	1989	
	SIPS	Net	1989	
Japan	BOJ-NET	Net+RTGS	1988	No
	FEYCS	Net	1989	
Netherlands	FA	RTGS+Net	1985	
	(TOP)	(RTGS)	1997	Yes
Sweden	RIX	RTGS	1986	Yes
Switzerland	SIC	RTGS	1987	No
United Kingdom	CHAPS	RTGS	1984	Yes
United States	CHIPS	Net	1970	No
	Fedwire	RTGS	1918	Yes

Source: BIS (1997b).

tlement system, Target, which will link settlement systems within the union, is also designed as a real-time gross settlement system. The progress of the European Union and the European Monetary Union have also contributed to the conversion of netting systems such as the U.K. CHAPS system to real-time gross settlement even though the United Kingdom is not projected to join the European Monetary Union initially. Table 1 briefly summarizes some of the salient characteristics of settlement systems in selected developed countries and illustrates the extent to which they are evolving toward real-time gross settlement.

Conclusions and Implications

The present path on which payments systems are moving involves a seeming contradiction. On the one hand, markets are becoming more integrated and global in scope. At the same time they are becoming more segmented in the sense that there is a growing separation evolving between the clearing and settlement of transactions. This increasing separation raises the prospect that there may be a need to invoke the safety net and introduces a possible distortion into the international payments system. As a consequence, both public-sector and private markets have given great attention to attempting to identify and control risk exposures. Perhaps one of the more interesting developments in this evolution of regional and globalized payments markets in both the public and private sectors has been the push toward real-time gross settlement systems with collateralization. Nowhere are these efforts more apparent than in Europe, where the struggle to create a single financial marketplace has focused attention and generated analyses of the underlying issues, with the Bank for

6. Real-time gross settlement may also improve risk management. In the case of derivatives clearinghouses, real-time gross settlement facilitates the use of intraday margin calls and the receipt of final funds before the end of the day.

International Settlements, the Group of Ten, and central banks spearheading much of this work.

Casual empiricism suggests several reasons why the systems are evolving in this direction despite considerable analysis suggesting that netting arrangements are more operationally efficient. The first reason is that systems, instruments, and markets are evolving faster than the political entities can bring their various rules and regulations into harmony despite the many initiatives that have been undertaken. Second, harmonizing systems to control effectively the systemic risks (such as Herstatt risk) inherent in nonsynchronized clearing and settlement systems, such as foreign exchange markets, even if all the legal rules are in place requires extensive international coordination and cooperation. Third, central banks realize that, regardless of the explicit rules governing exchanges and settlement arrangements, they still may be thrust into the role of the lender of last resort should major participants get into financial difficulties that threaten to bring down settlement and clearing systems. In the United States, the decline in member bank reserve balances reduces payments system participants' liquidity positions and increases the likelihood that intraday credit may have to be extended. Finally, the movement toward expanding the overlapping hours that exchanges are open will increasingly make the operation of net settlement systems more difficult.

REFERENCES

Bank of England and APACS. 1994. "The Development of Real-Time Gross Settlement (RTGS) in the United Kingdom." Bank for International Settlements information release, April.

Bank for International Settlements (BIS). 1989. "Report on Netting Schemes." Prepared by the Group of Experts on Payments Systems of Central Banks of the Group of Ten Countries, February.

———. 1990. "Report of the Committee on Interbank Netting Schemes of the Central Banks of the Group of Ten Countries." November.

———. 1993a. "Central Bank Payment and Settlement Services with Respect to Cross-Border and Multi-Currency Transactions." Report prepared by the Committee on Payment and Settlement Systems of the Central Banks of the Group of Ten Countries, September.

———. 1993b. "Minimum Common Features for Domestic Payments Systems." Report of the Committee of Governors of the Central Banks of the Member States of the European Economic Community. Action 2 of the report on issues of common concern to EC central banks in the field of payments systems, by the Working Group on EC Payment Systems, November.

———. 1993c. *Payment Systems in the Group of Ten Countries*. Report prepared by the Committee on Payment and Settlement Systems of the Central Banks of the Group of Ten Countries. Basle, December.

———. 1996. *Settlement Risk in Foreign Exchange Transactions*. Report prepared by the Committee on Payment and Settlement Systems of the Central Banks of the Group of Ten Countries. Basle, March.

———. 1997a. *Clearing Arrangements for Exchange-Traded Derivatives*. Report prepared by the Committee on Payment and Settlement Systems of the Central Banks of the Group of Ten Countries. Basle, March.

———. 1997b. *Real-Time Gross Settlement Systems*. Report prepared by the Committee on Payment and Settlement Systems of the Central Banks of the Group of Ten Countries. Basle, March.

Benston, George J., and George G. Kaufman. 1995. "Is the Banking and Payments System Fragile?" *Journal of Financial Services Research* 9 (December): 209-40.

Corrigan, E. Gerald. 1990. "Perspectives on Payments System Risk Reduction." In *The U.S. Payments System: Efficiency, Risk, and the Role of the Federal Reserve*, edited by David B. Humphrey. Proceedings of a Symposium on the U.S. Payments System sponsored by the Federal Reserve Bank of Richmond. Boston, Mass.: Kluwer Academic Publishers.

Edwards, Franklin R. 1996. *The New Finance: Regulation and Financial Stability*. Washington, D.C.: AEI Press.

Eisenbeis, Robert A. 1995. "Private-Sector Solutions to Payments System Fragility." *Journal of Financial Services Research* 9 (December): 327-49.

General Accounting Office. 1994. *Financial Derivatives: Actions Needed to Protect the Financial System*. Report to Congressional Requestors, GAO/GGD-94-133. Washington, D.C.: Government Printing Office, May.

Hu, Jie. 1994. "Information Ambiguity: Recognizing Its Role in Financial Markets." Federal Reserve Bank of Atlanta *Economic Review* 79 (July/August): 11-21.

Knight, Frank H. [1921] 1971. *Risk, Uncertainty, and Profit*. Reprint, Chicago: University of Chicago Press.