REVIEW

Vol. 73, No. 6 November/December 1991

THE FEDERAL RESERVE BANK of ST. LOUIS

Federal Reserve Bank of St. Louis
Review
November/December 1991

In This Issue . . .

When oil prices doubled in 1990 as a result of Iraq's invasion of Kuwait, the economic effects of oil price shocks became central to discussions of the economic outlook and economic policy. In the first article in this *Review*, "The 1990 Oil Price Hike in Perspective," John A. Tatom discusses several popular but misleading perceptions about the changing effects of oil price shocks on the U.S. economy and policymakers' response to such changes in the past.

The author explains that the principal channel of influence of higher energy prices is a loss in economic capacity and productivity. Other channels of influence arise through the oil import bill or changing energy efficiency. Tatom compares the recent experience with the two previous oil price shocks and finds that there was little reason to expect the effects of the 1990 oil price shock to be much smaller.

The most significant difference in the latest shock, says the author, is that it was more clearly temporary, so that the doubling of oil prices from July to October 1990 was virtually eliminated by offsetting movements from October to March 1991. As a result, Tatom concludes, the adverse effects of the oil price shock should be reversed almost as quickly as they occurred.

* * *

In the second article in this issue, "Alternative Measures of the Monetary Base: What Are the Differences and Are They Important?" Michelle R. Garfinkel and Daniel L. Thornton explore the differences in two measures of the adjusted monetary base, one constructed by this Bank, the other constructed by the Federal Reserve Board. Noting that these two indicators of monetary policy can and frequently do give conflicting impressions of monetary policy, the authors briefly review the basic idea behind the adjustment for reserve requirement changes. They then discuss differences in the construction of the two measures and go on to explore the importance of these differences empirically. Because the differences in their construction are arbitrary and technical in nature, there is little reason to prefer one measure over the other. The authors suggest that, when the difference is important, the measure that appears to perform best for the problem at hand should be used.

* * *

In the third article in this *Review*, "A Microeconometric Approach to Estimating Money Demand: The Asymptotically Ideal Model," Piyu Yue presents an advanced approach to estimating money demand called the "Asymptotically Ideal Model" or AIM. She briefly reviews alternative microeconometric approaches to money demand, then estimates the AIM using U.S. quarterly monetary aggregate data. Dynamic simulations of the growth rate of the various aggregates and consumption suggest that the model performs well. Among other things, she concludes that the failure of conventional money demand equations may result from the inability of linear equations to approximate the behavior of non-linear demand functions.

* * *

In recent years, a number of articles discussing the theory of securities market microstructure have appeared in economics and finance journals. The study of market microstructure deals with the behavior of participants in securities markets and the effects of information and institutional rules on the economic performance of securities markets. Although a large and growing number of such articles have appeared, relatively few have attempted to model the foreign exchange market.

In the fourth article in this issue, "Microstructure Theory and the Foreign Exchange Market," Mark D. Flood reviews the theoretical literature on market microstructure to see what lessons it holds for the foreign exchange market. Microstructure theory is of interest to students of the foreign exchange market, says the author, because it can yield insights into dealers' behavior and the impact of institutional arrangements. Conversely, the foreign exchange market is of interest to students of microstructure, because it combines two very different methods of matching buyers and sellers—bank dealers trade both directly and through brokers.

* * *

John A. Tatom

John A. Tatom is an assistant vice president at the Federal Reserve Bank of St. Louis. James P. Kelley provided research assistance.

The 1990 Oil Price Hike in Perspective

THE ECONOMIC EFFECTS of the sharp rise in oil prices in 1990 were, for a while, the central issue in discussions of the economic outlook for 1990 and 1991. Iraq's maneuvers to raise the world price of oil late in July 1990 and their invasion of Kuwait less than a week later led to a doubling of oil prices. As a result, oil price shocks and the appropriate economic policy response to such shocks became subjects of renewed speculation.

One of the most popular hypotheses to emerge at the time was that, since the economy was different in 1990 than it had been when previous large oil price increases occurred, the 1990 price rise should not affect the economy to the same extent.[1] It still was widely believed, however, that the principal and most immediate effect would be the onset of a recession. In response, many analysts believed that the Federal Reserve would ease monetary policy because they thought it had done so at the outset of previous oil shocks.

This article outlines the potential channels of influence of a rise in the price of oil and the extent to which the purported differences in economic conditions in 1990 could account for differences between the economic effects of the 1990 oil price surge and those in earlier, comparable episodes.

WHY DO OIL PRICES MATTER?

One usually encounters two principal arguments in assessing how oil and energy price changes affect the economy. First, since energy resources are used to produce other goods and services, a change in their price affects how much of the goods are produced as well as the mix of resources that will be used to produce them. This argument focuses on the supply side of the markets for goods and services. It suggests that the output losses associated with higher energy prices are permanent, so that changing economic policies or shifting market prices cannot replace the loss.

A second argument focuses on the effects on the demand for a country's output. It suggests that output losses are cyclical or transitory, so that adjustments in wages and prices, or in economic policy, can reverse the loss in output.

[1]Fieleke (1990) was one of the first to develop this argument. Among the reasons he cites are differences in the size of the shock, the sensitivity of oil consumers to oil price changes, the state of the economy before the oil shock and differences in available policy options. The Council of Economic Advisers (1991) provides a more extensive discussion consistent with this view.

Each argument suggests which characteristics of the economy determine the effects of an energy price shock, as well as how changes in these characteristics would alter these effects. Each also provides a different conclusion about the potential for economic policy to ameliorate the adverse influences of energy price shocks.

Energy Prices and Economic Capacity: The Permanent Effects of an Energy Price Shock

Energy resources are used to produce most goods or services. As such, a rise in their price will (1) raise the total cost of an efficient producer's output, (2) alter the most efficient means for producing output, (3) lower the profit-maximizing level of output, (4) raise the long-run equilibrium price of output and (5) reduce the capacity output of each firm's existing stock of capital.[2] Capacity output declines when energy prices rise because firms reduce their use of energy and energy-using capital, some capital becomes obsolete, and firms use labor and capital to economize on energy costs—that is, they generally switch to less energy-intensive production methods. The shaded insert on pages 6 and 7 briefly explains the microeconomic foundations of this capacity effect.

The economy's aggregate supply is the sum of the supply decisions of the nation's firms. Thus, the effect of energy prices on the typical firm's economic capacity determines the effect on the economy's natural output and its aggregate supply. The influence of a rise in the price of energy on aggregate supply is shown in figure 1. The aggregate supply curve indicates the output that producers will supply at various levels of the aggregate price level, given other factors influencing this decision. The supply curve typically is derived from a given production function, which relates output to the employment of resources such as labor and capital. An initial level of nominal wages, the supplies of labor and capital goods and the relative price of energy resources are assumed to be given in deriving a particular aggregate supply curve.

Suppose that the price level, P_0 in figure 1, results in a real wage (nominal wage deflated by the price level) at which a given supply of labor resources is fully employed. At this level of employment, which often is referred to as natural employment, the economy produces its capacity or natural output level, X_n^0. Given the nominal wage level, the real wage is lower when prices are higher than P_0, so firms would desire to produce more output and demand more employment. Workers would be unwilling to work more at a lower real wage, however, so neither output nor employment could rise. Indeed, to maintain output and employment, the nominal wage must rise proportionately with the price level to keep the real wage unchanged. Thus, the aggregate supply curve is vertical at X_n^0 for prices above P_0. At a lower price level than P_0, the real wage is too high for firms to employ as much labor or produce as much output as at X_n^0; output and employment are below their natural counterparts along this upward-sloping portion of the aggregate supply curve.

A rise in the relative price of energy, given the short-run supply of capital and labor resources, will reduce capacity output from X_n^0 to X_n^1 and raise the aggregate level of prices associated with this output from P^0 to P^1. The percentage decline in capacity output and the rise in price level associated with each 1 percent rise in the relative price of energy generally are equal and proportional to the share of energy in the cost of output.[3] In this case, although real output has fallen, the level of nominal spending on output at point B in figure 1 will be the same as at point A. Thus, if output is measured by the nation's real GNP, then real GNP is lower at point B than at point A, but nominal GNP is the same.

Aggregate output and the price level are determined by the interaction of aggregate supply

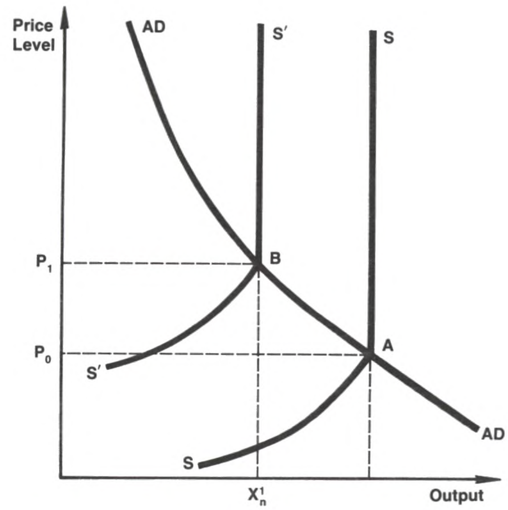
---

[2]This discussion draws upon Rasche and Tatom (1977a) and (1981); Karnosky (1976) was one of the first to argue that a rise in the price of energy reduces capacity and raises the price level. Hickman, Huntington and Sweeney (1987) summarize the similarities and differences of empirical estimates of the effects of energy price shocks in 14 prominent macroeconomic models. All of these models show a permanent output loss due to an oil price increase; in six of these models, this loss is explictly cited as a decline in potential output.

The Council of Economic Advisers (1991) suggests that any effect on capacity is transitory. Others who have been critical of the significance of the capacity effect include Berndt (1980), Berndt and Wood (1987), Denison (1979) and (1985), Darby (1984) and Olson (1988).

[3]The conditions required to obtain the equality of these outcomes are discussed in Rasche and Tatom (1977a) and derived in Rasche and Tatom (1981). The shaded insert to this article provides a summary of the analysis.

Figure 1
The Effect of a Higher Price of Energy on Output and the Price Level



and demand. Aggregate demand indicates the quantity of output demanded at various price levels and is inversely related to the general price level. The aggregate demand curve in figure 1 passes through both points A and B. At these points, nominal GNP (the product of the price level and output) is the same, indicating that a rise in the price level is associated with an equal proportionate decline in real output. Thus, the nominal value of aggregate demand is unaffected by the price level.

This assumption simplifies the analysis without reducing its generality. The higher price level reflects the permanent decline in natural output, with no cyclical loss of output or employment; the smaller natural output level is produced by an unchanged level of natural employment. Only a further reduction in output would fit the notion of a cyclical loss associated with cyclical unemployment.

For cyclical output and employment losses to arise from an energy price increase, either (1) aggregate demand must be more responsive to a rise in the price level (flatter than that drawn in figure 1), (2) an increase in the relative price of energy must cause the aggregate demand to shift to the left, or (3) there is some short-run dynamics of price and output adjustment not shown in the movement from A to B. For example, if the price level adjusts upward slowly because of temporary rigidities in the prices of goods and services, then a rise in energy prices will lead producers to reduce employment temporarily, reducing output by more than the decline in natural output. When output prices rise sufficiently to reduce real wages by the extent of the permanent decline in labor productivity, employment will be restored to its natural level and output will have fallen only to the extent of the capacity loss.[4] Thus, even if the principal effects of an energy price rise are a permanent decline in capacity and a rise in the price level, some transitory recessionary declines in output and employment are likely to occur.

Energy Prices and Aggregate Demand

The second channel of influence above indicates that a rise in the relative price of energy would shift aggregate demand to the left, reducing output and/or the level of prices. These effects are transitory, or cyclical, however, in contrast to the permanent output loss arising from reduced capacity. When output is less than its natural level, employment is as well. Thus, wages and rental prices of capital goods will tend to fall, shifting the upward-sloping portion of the aggregate supply curve and the price level down until output is restored to its natural level.

Aggregate demand will fall if a rise in oil prices raises expenditures on oil and total imports and thereby lowers net exports. In effect, the rise in the oil import bill acts like a tax on domestic income, reducing aggregate demand. For such a shift in aggregate demand, the decline in output and employment are propor-

[4]Tatom (1981) indicates that temporary cyclical effects occur for the third reason above; that is, they are short-run dynamic variations as the economy moves from point A to point B. In this analysis, sticky prices keep the price level from rising instantaneously. Inventories and increased employment initially are used to meet unchanged sales and partially offset the productivity loss. Within a short time, however, firms begin to reduce output because sales fall more as prices begin to rise; cyclical losses in output and employment occur. Empirical evidence indicates that, after about a year, the price level has adjusted fully (to P_1 in figure 1), so producers step up production and employment to their natural levels (point B).

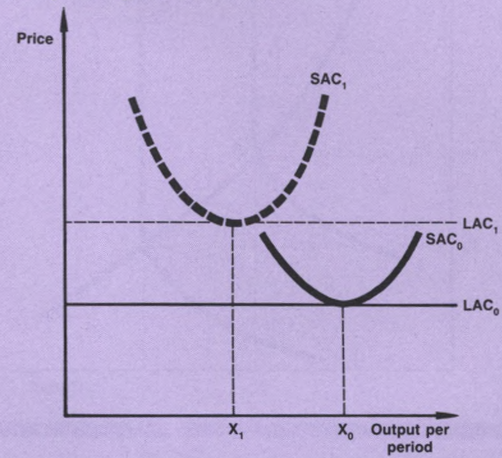
The Effect of a Higher Price of Energy on Economic Capacity

The effect of a rise in the price of energy on a firm's cost structure is illustrated in the accompanying figure, which shows the long- and short-run average cost of output and how they are affected by a rise in the price of energy. The long-run average cost curve (LAC_0) indicates the minimum cost per unit of output for the firm. This curve is derived from the least-cost combination of resources that produces the indicated quantity of output, given available technology and the prices of the resources used to produce the firm's output. The long run refers to a period over which the firm is free to vary the quantity of all resources used in production.

In the figure, the long-run average cost curve is horizontal, or unaffected by the level of output. The long-run average cost could decline over some range of output, indicating what are called "economies of scale," or it could rise over some range indicating "diseconomies of scale." When the curve is horizontal, as in the figure, the firm's production exhibits constant returns to scale so that, for any initial output level, proportional increases or decreases in output can be obtained from equiproportional changes in the employment of each resource. In this case, long-run cost varies equiproportionately with output, so that the long-run average cost is unaffected by the output level the firm chooses to produce. With constant returns to scale, the long-run average cost also indicates the long-run marginal cost, the minimum additional total cost of producing an additional unit of output.[1]

The short run is characterized by the inability to vary the use of some resources. In particular, firms have difficulty in varying their capital stock—their plant and equipment—to produce more or less output in the short run. Thus, a given size of capital stock would be freely chosen to allow least-cost production at only one level of output. At a larger (smaller) output, more (less) capital would be used to minimize the cost of production. The output level at which the existing stock of capital would be selected is called the economic capacity of the firm's capital stock. At this output, the long-run and short-run total and average cost of output are the same.

The Effect of a Higher Price of Energy on a Firm's Cost, Capacity and Price



Should the firm desire to produce more or less output, it could not do so as cheaply in the short run as it could in the long run because the capital stock cannot be varied in the short run. Higher-cost methods of production, which use relatively more labor or other variable resources, must be used until the capital stock can be altered. Since the total cost of producing any level of output other than the capacity level (X_0) is higher in the short run than the long run, the short-run average cost (SAC_0) is also higher.

When the price of energy rises, the long-run and the short-run average cost of output rise (LAC_1 and SAC_1, respectively). The size of

[1]The most general case is often illustrated with a U-shaped long-run average cost curve, which exhibits increasing returns to scale over the range of relatively low output levels and decreasing returns to scale at relatively high levels of output. At the minimum long-run average cost, there are constant returns to scale.

the rise depends, in part, on the size of the increase in energy prices and the share of energy resources in total cost. The effect on the firm's economic capacity depends on how the optimal long-run mix of resource employment changes. The higher price of energy will cause the firm to reduce its use of energy to produce a given level of output (say X_0) and increase the use of some resources whose prices have not changed. The use of some other resources could also be reduced along with energy.[2] As drawn in the figure, the capacity output of the firm falls to X_1, since the short-run average cost rises more than the long-run average cost at output X_0.

Whether capacity output falls, rises or is unaffected by a rise in the price of energy depends on the relationship between capital and energy resources in production. When these resources are substitutes, a rise in energy prices would lead the firms to substitute capital for energy in producing X_0. Thus, the existing capital stock would be chosen to produce a smaller output level (X_1).[3] If capital and energy are independent or complements, the higher short-run and long-run average cost level will be identical at the same or higher level of output, respectively. Energy and capital are substitutes, however, which implies that economic capacity will fall when energy prices rise.[4]

[2]The effect of a rise in the price of one resource on the efficient level of employment of another is assessed by the "elasticity of substitution" between the two resources, which measures the percentage change in one resource associated with each 1 percentage-point rise in the price of the other, holding output constant. If the elasticity of substitution between energy and some other resource is positive, a rise in the price of energy raises the employment of the other resource. In this case, energy and the other resource are said to be substitutes. When this elasticity of substitution is negative, energy and the other resource are "complements."

[3]Some analysts describe production as "putty-clay," meaning that, in the short run, the capital stock requires fixed proportions of other resources. In this case, the elasticity of substitution between energy and capital is zero in the short run, so that capacity is unaffected. Over time, a capacity loss would be realized as obsolete capital is replaced with capital that uses less energy. See Corcoran (1990), for example. Evidence cited in the text suggests that putty-clay considerations are not dominant in the short run.

[4]The size of the fall in economic capacity is proportional to the share of energy in factor cost and the elasticity of substitution and inversely proportional to the expenditure elasticity of the capital stock (or fixed resources generally). The elasticity of capacity equals $(-\sigma_{ke}/\eta_k)$ s_e, which reduces to $-s_e$, the share of energy in total cost, if the expenditure elasticity of capital, η_k, and the elasticity of substitution, σ_{ke}, are one; these elasticities are each one for a broad class of functions used in empirical analysis. The former elasticity is the response of desired capital use to an increase in expenditures on all resources. For increases in the relative price of energy, instead of the nominal price, the share of energy in the expressions here and above is replaced by $(s_e/1-s_e)$.

tional to the rise in the oil price and the share of oil imports in GNP. In this case, the net oil import status of a country determines the effects of an energy price shock.[5] Countries that export oil face larger aggregate demand when oil prices rise; world aggregate demand and output are redistributed from oil importers to exporters when oil prices rise, and conversely when oil prices fall.[6]

Monetary Policy and Oil Price Shocks

The appropriate monetary policy response to a rise in the relative price of energy depends on its dominant channel of influence. If the higher energy price only lowered aggregate demand, policymakers could take offsetting actions to neutralize this shift by increasing the money

[5]Feldstein (1990) and the Council of Economic Advisers (1991) provide recent restatements of this shift in aggregate demand and the price-level-induced movement along the aggregate demand curve as the central channels of influence of an oil price hike. The Council of Economic Advisers also emphasizes a decline in real consumption expenditures as a result of an oil price hike. Perry (1991) argues that the oil price hike had little effect on the economy in 1990, because it did not reduce real income much (operating through the aggregate demand channel above), nor did it induce the Fed "to raise interest rates to fight inflation" as, he argues, it had in the past.

[6]A change in the relative price of energy could also affect aggregate demand by altering investment in plant, equipment and housing. Such an effect can account for a decline in the real interest rate, which is incompatible with a conventional model of aggregate demand. Reinhardt (1991) discusses the effects of energy price shocks on interest rates.

supply, which would shift the aggregate demand curve back to the right.

If an energy price increase affects aggregate supply, however, both raising the price level *and* reducing natural output, policymakers could attempt to offset the price level rise by reducing the money stock to reduce aggregate demand. This would result in a cyclical loss in output and employment as the economy's output fell short of its lower natural output level until the price level declined sufficiently.

Alternatively, policymakers could attempt to offset the reduction in output by raising aggregate demand. Raising demand could not restore the economy's natural output, however; it would not replace the energy and capital resources that firms can no longer afford to purchase or use. Instead, it would further raise the aggregate level of prices associated with the smaller level of capacity output.[7]

Thus, there is no real policy dilemma posed by oil price increases. Raising the money stock cannot offset a loss in natural output, while reducing the money stock can only offset a price level increase at the cost of a further loss in output and a cyclical rise in unemployment. Moreover, it is virtually impossible to alter monetary policy enough to fully offset the price level surge because of the time it takes for a change in the money stock to affect the price level and because of the relatively small size of the initial price response to changes in monetary policy.[8] An unchanged growth rate for the money stock is a policy that accepts the permanent output and price level consequences described above without compounding one or the other loss.

HAVE THE ECONOMIC EFFECTS OF OIL PRICE SHOCKS CHANGED?

Many analysts argued that the rise of oil prices in 1990 would have substantially less impact on the U.S. economy than earlier oil price hikes. There were two versions of this argument. The first was that the adverse effects of an oil price rise are proportional to the share of oil imports in the economy and that this share had fallen since the earlier oil price shocks. The second argument was that the effects of an oil price rise are proportional to the use of energy per unit of output and that this dependence on energy also had fallen.[9]

Does a Smaller Import Share Reduce the Adverse Effects of an Oil Price Hike?

If the share of oil imports in GNP has fallen, then the first argument above implies that the economy's aggregate demand and output have become less sensitive to a rise in oil prices. Figure 2 shows expenditures on petroleum imports as a percent of nominal GNP since 1970. In mid-1990, this share was about 1 percent, less than half its level in early 1979, but above its 0.6 percent share in 1973. Thus, the share had fallen below its level preceding only one of the previous two oil price shocks.

The import share argument has other shortcomings. First, it suggests that oil-exporting countries, including Canada in 1974 or the United Kingdom in 1979, should gain when oil prices rise, because net exports and aggregate demand should rise. In each instance, however, output did not rise nor was there other evidence of a cyclical expansion following the previous oil price shocks. The argument also suggests that countries that import a relatively small share of their oil, like the United States, will be less affected than countries that import relatively more of their oil, like Germany or Japan. The earlier experience with oil price shocks indicates that, especially in 1973-74, both the temporary rise in inflation and the permanent loss in output were larger in Japan than in

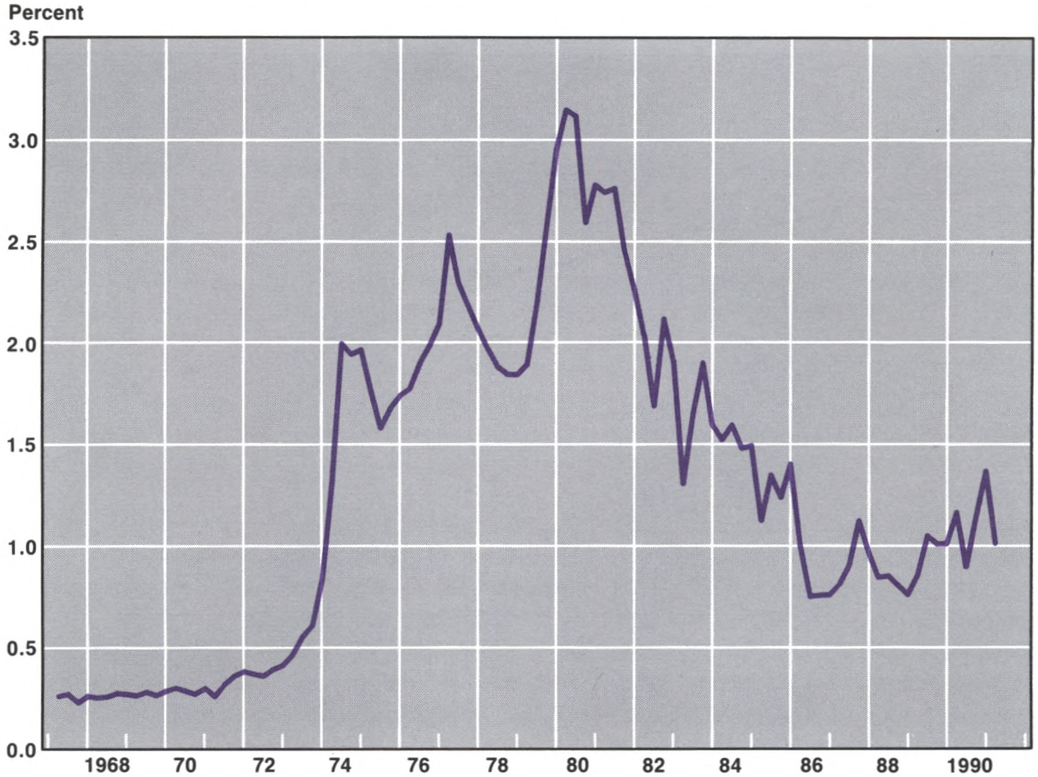
---

[7]Kahn and Hampton (1990) contrast three monetary policy options, which include tightening to offset the price level effect, easing to offset the cyclical effects and a neutral policy which "maintains constant monetary or nominal GNP growth." Feldstein (1990) endorses the third option, nominal GNP targeting, and he also equates this with unchanged money stock growth.

[8]See Tatom (1981) and (1988a), for example, for evidence on the relative size and lag lengths for energy price and monetary policy effects on prices and output.

[9]See Council of Economic Advisers (1991), Kahn and Hampton (1990), Anderson, Bryan and Pike (1990), Brinner (1990), "How Big An Oil Shock" (1990), "Shocked Again" (1990), May (1990), Yanchar (1990) and Fieleke (1990) for analyses that emphasize one or both of these arguments. Fieleke, Kahn and Hampton, May and Yanchar emphasize, to varying degrees, that the expected effects also are smaller because of a smaller expected rise in the price of oil.

Figure 2
Petroleum Imports as a Percent of GNP



the United States, but that these effects were smallest in Germany.[10]

There are three other major difficulties with the import share argument. First, it is difficult to reconcile the relatively large economic effects of oil price hikes with the relatively small size of the petroleum import share. Second, an aggregate demand reduction in the face of an oil price hike implies only a cyclical decline in output, not a permanent one. The failure of real GNP per worker and real wages to return to their previous growth trends in virtually all nations after the two previous OPEC price hikes is not consistent with the pattern expected for a purely cyclical loss. Third, the trade-based aggregate demand story predicts a decline in net exports and the currency value of a large oil importer after an oil price shock. At least for the United States, however, exports rose relative to imports so that both net exports and the exchange rate rose after each earlier oil price shock. Indeed, the only periods of positive net exports since 1970 occurred in 1974-75 and 1979-82, following the earlier oil price hikes.[11]

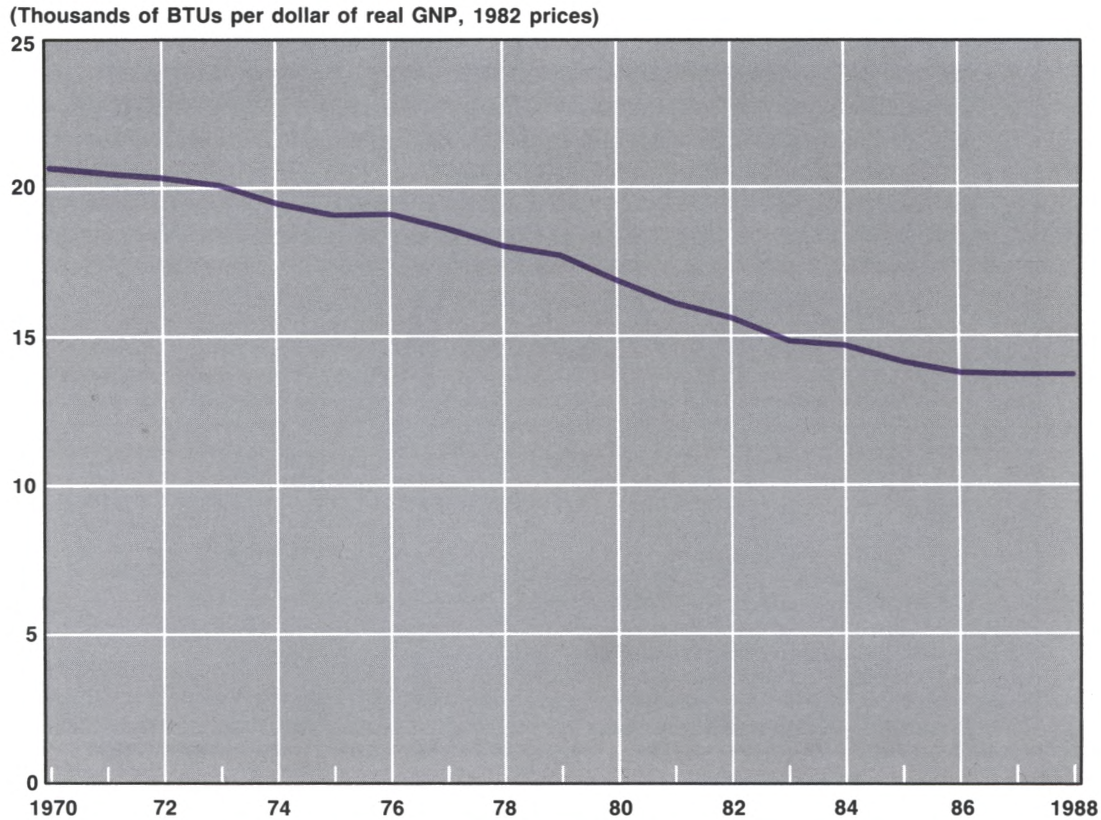
Does Increased Energy Efficiency Reduce the Adverse Effects of an Oil Price Hike?

The second argument for less adverse effects of the 1990 price hike is based on a decline in

[10]See Rasche and Tatom (1981), Tatom (1987) and Tatom (1988a) for reviews of this international evidence. In 1973, the share of petroleum imports in GNP equaled 1.7 percent in Germany and 1.6 percent in Japan, much more than the 0.6 percent in the United States. Similarly, in 1978, this share was 2.7 percent in Japan, 2.5 percent in Germany and 1.9 percent in the United States.

[11]See Tatom (1988b) for a discussion of the theory and evidence supporting such contrary effects. Consistent with this rise in net demand for U.S. goods, the trade-weighted value of the dollar rose in IV/1973 and I/1974, and was higher over the rest of 1974 than it had been in the two quarters preceding the oil price rise. In the second quarter of 1979, the value of the dollar also rose slightly. Over the next four quarters, the value of the dollar was only 0.6 percent lower than in the two quarters before the oil price hike.

Figure 3
Energy Use per Unit of Real GNP



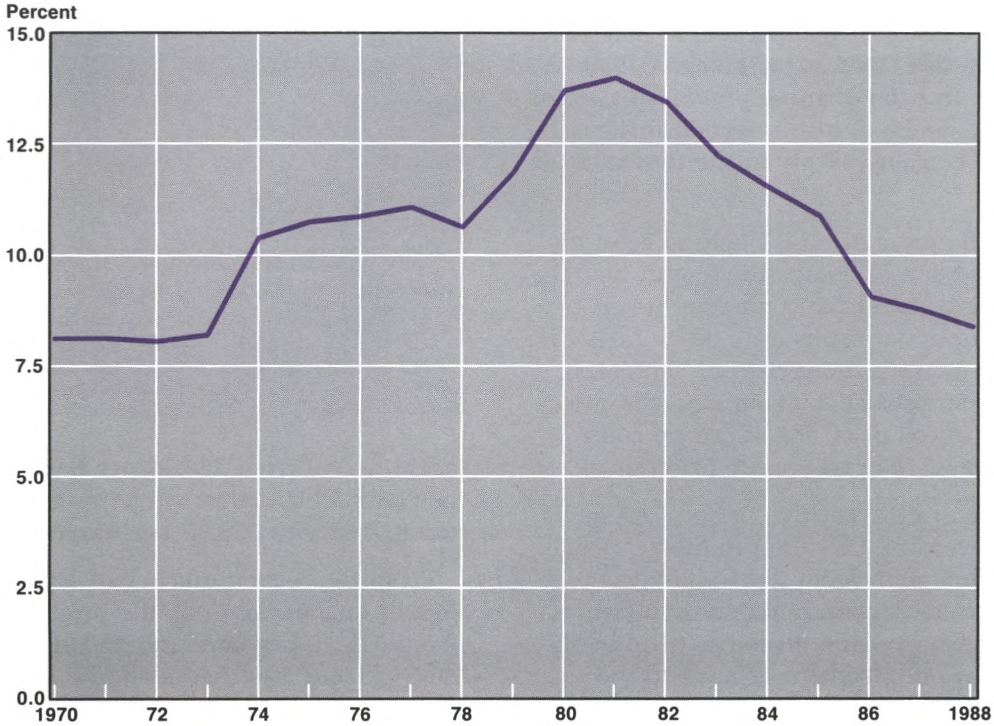
energy use per unit of output. According to this argument, energy is less important to a firm's production than in the past, so a rise in oil prices is expected to have a smaller effect on prices and production today than in the past.

Figure 3 shows total U.S. energy use per unit of output (measured in BTUs per unit of real GNP) from 1970 to 1988, the latest year available on this basis.[12] Energy use per unit of output has fallen sharply since 1973: BTUs used per unit of real GNP were about 31 percent lower in 1988 than in 1973 and about 22 percent lower than in 1979. This rise in output per unit of energy is not surprising given the rise in the relative price of energy since 1973, but it is not relevant in assessing the importance of energy as a resource or in assessing whether the effects of an energy price boost have declined in magnitude.

While energy use per unit of output is lower than earlier, the responsiveness of prices or output to a change in a resource's price are proportional to the share of the resource's cost in total cost, not to the share of its quantity in output. Consider the familiar case of labor productivity. Labor employment per unit of output in the business sector declined by nearly one-third from 1955 to 1973, as output per worker rose from $21,084 to $31,142 (1982 prices). Thus, the economy became less dependent on labor over these 18 years—in exactly the same sense and to nearly the same extent as some have suggested about energy resources over the past 18 years. Nevertheless, the share of labor in total

[12]The energy expenditures and quantity data used for figures 2 and 3 are from the Energy Information Administration, *State Energy Price and Expenditure Report, 1988* (September 1990).

Figure 4
Energy Expenditures as a Percent of GNP



cost was about the same: 65.3 percent in 1973 and 64.8 percent in 1955. For a given share of labor in cost, a percentage point rise in the wage rate will raise the cost of an additional unit of output and price in proportion to this share.[13]

Analysts who emphasized the increased productivity of energy are unlikely to espouse the equivalent view that a 10 percent rise in wages has a smaller effect on unit costs or product prices today than in 1973 or 1955. As discussed previously, the response of capacity and price to changes in a resource's price depends on the share of the resource in cost, not on its productivity or output per unit.

Figure 4 shows how the share of energy expenditures as a percent of GNP has changed from 1970 to 1988. Following each energy price hike, expenditures rose sharply relative to GNP; as energy prices fell beginning in 1982 (on an annual basis), the share fell. By 1988, the share nearly had returned to its 1970-73 level. These data suggest that the share of energy in the cost of the economy's output has not fallen below its level before the earlier oil price changes, especially the 1973-74 rise. Thus, these data do not support the view that a doubling of the price of oil should be expected to have smaller effects in 1990 than it had earlier, especially in 1973-74, because the share of energy in total cost has not declined.

RECENT OIL AND ENERGY PRICE DEVELOPMENTS

The economic effects of an energy price shock depend on the size of the price change as much as they depend on the responsiveness of measures of economic performance to a given

[13]A typical discussion of the relationship between wages, productivity and prices can be found in Fischer, Dornbusch and Schmalensee (1988), pp. 566-67. Shin (1991) discusses other shortcomings of using the energy-output ratio for analytical or policy purposes.

change in energy prices. Table 1 shows the monthly average price of oil purchased by refiners since June 1990. Following the Iraqi invasion of Kuwait and the subsequent U.N. embargo of crude oil exports from both countries, the price of oil doubled within three months. The 1990 oil price rise was comparable in magnitude to the two earlier OPEC price hikes in 1973-74 and 1979-80. In each of these previous cases, oil prices nearly doubled. In the second instance, oil prices rose again sharply in the first quarter of 1981.

A rise in the price of oil is likely to raise the cost of production of competing energy sources and raise the demand for competing forms of energy, as consumers substitute other fuels for oil. For both reasons, the prices of competing sources of energy change along with the price of oil. Thus, an oil price shock can be considered more generally an energy price shock.

Figure 5 shows the relative price of crude petroleum—measured by the producer price index for crude petroleum deflated by the business sector implicit price deflator—and the relative price of energy—the producer price of fuel, power and related products relative to the same deflator.[14] The relative price affects economic performance because producers of goods and services assess the cost of energy relative to the goods and services produced using it. From the third quarter of 1973 to the third quarter of 1974, the relative price of crude oil nearly doubled. Measured in 1990 prices, the composite refiner acquisition cost of crude oil rose from $10.67 per barrel in 1973 to $21.28 in 1974, or 99.4 percent.[15] In the second OPEC oil price shock, from early 1979 to the second quarter of 1980, this relative price of oil nearly doubled again, rising from $22.35 per barrel to $41.82 per barrel. A further surge in early 1981 put the price up to $50.75 per barrel.

Table 1
The Composite Refiners Acquisition Cost of Crude Oil (dollars per barrel)

Date	Price	Date	Price
June 1990	$14.98	January 1991	$22.90
July	16.15	February	19.02
August	23.57	March	17.89
September	30.01	April	18.43
October	33.18	May	18.60
November	30.61	June	17.98
December	26.21		

From the second quarter of 1990 to the fourth quarter of 1990, the price of oil rose from $16.10 per barrel to $30.00 per barrel, an 86.3 percent rise that is almost as large as the near-doubling in the previous two oil shocks.[16] If the effects of oil price hikes are proportional to their size, then the effects of the 1990 increase should be about the same as in the two previous instances. The relative price of energy rose about 50 percent during the previous two energy price shocks. From the second quarter of 1990 to the fourth quarter of 1990, however, the relative price of energy rose 29.6 percent, about 60 percent of the earlier magnitudes.[17] Thus, on this basis, the recent energy price shock is somewhat smaller.

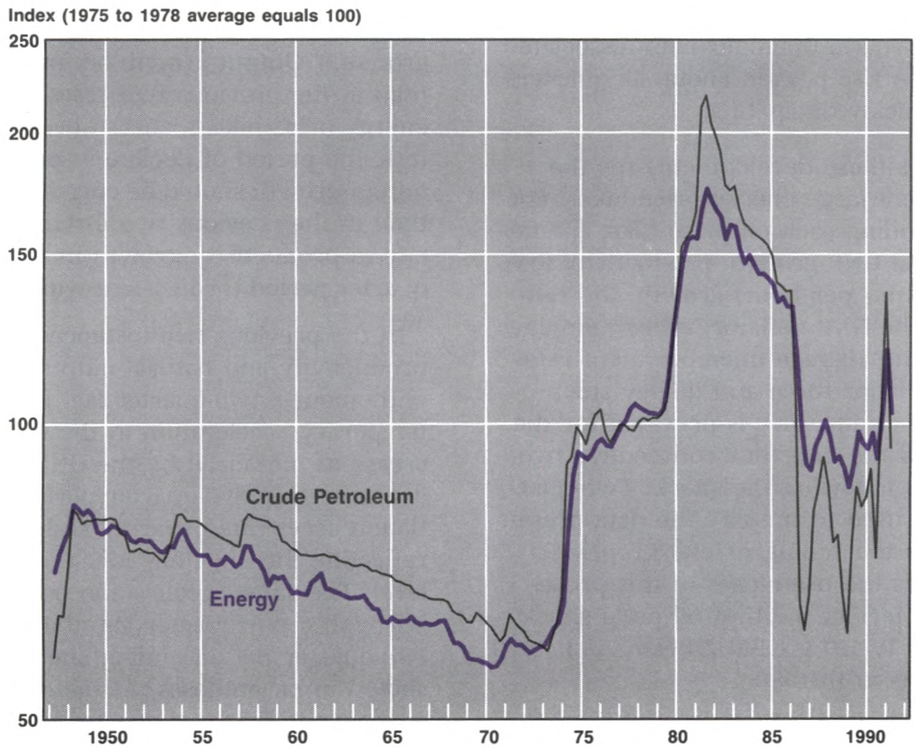
[14]A logarithmic scale is used because differences in logarithms show percentage changes; an equal-sized increase or decrease in figure 5 reflects equal *percentage* changes. For example, a rise from 50 to 100, or 100 to 200 represents a doubling of the relative price and the respective distance in each case is the same in figure 5.

[15]The rise in the relative price of oil shown in the figure actually begins in early 1973, but this earlier increase largely reflects partial and temporary relaxation of U.S. price controls on domestic crude oil prices. The much larger OPEC price increases followed the Yom Kippur War in October 1973. The 1947 oil price shock is not discussed here. The producer price for crude petroleum measures prices paid to domestic producers, which were controlled from 1971 to early 1981. Over most of this period, the composite refiner acquisition cost was higher, but was representative of oil prices paid by domestic purchasers.

[16]The total output of Kuwait and Iraq fell about 4 million barrels per day in August 1990 from its May-July 1990 average; by November and December 1990, it was down 4.6 million from the earlier average. The latter reduction equaled 7.6 percent of world production and 19 percent of OPEC output. In comparison, the reductions in the total of Iran and Iraq production from 1978 to its lowest annual average level in 1981 was somewhat larger, 5.4 million barrels per day, but this was 18.2 percent of OPEC's 1978 production.

[17]Empirical estimates suggest that the relative price of energy adjusts contemporaneously and with a one-quarter lag to changes in the relative price of oil; thus, one reason for the relatively smaller rise in the energy price is the fact that the relative price of crude oil fell 20.6 percent in the second quarter of 1990. When expressed in logarithms, each 1 percentage-point rise in the relative price of crude oil is estimated to result in about a one-half percent rise in the relative price of energy. See Tatom (1987b).

Figure 5
Relative Price of Energy and Crude Petroleum



There were two other important differences between the recent rise and the previous two. First, the recent rise occurred much more quickly—in two quarters instead of four or six. Second, the recent increase did not persist. Nevertheless, producers did not know at the time whether, or by how much, oil prices might decline in the future. This article assumes that producers treat price changes as permanent, in the sense that the expected price they use for economic decisions is the current price. It also focuses only on the effects of the recent price increase. To the extent that producers did not anticipate having to face the price increase, the effects of the price shock should be smaller.

THE EXPERIENCE IN PREVIOUS OIL PRICE SHOCKS

The previous discussion of energy price effects indicates that the 1990 oil price hike should be associated with a lower level of natural output and productivity and a rise in the price level. These changes were likely to be revealed as a temporary acceleration in inflation and a temporary reduction in output growth. Moreover, temporary rigidities in nominal prices and lags in the adjustments that firms and consumers make in response to large price changes were likely to give rise to temporary movements in employment, including a recessionary decline in employment, although past experience suggests that such a change occurs with a delay of about one year. These effects should be expected to have been somewhat smaller than those following previous oil shocks, because the rise in the relative price of energy in the 1990 episode was only about 60 percent as large as the previous increases.

Following the sharp rise in the relative price of energy in 1973-74 and 1979-80, the loss in capacity and adjustment to a higher price level, as discussed earlier in reference to figure 1, were reflected in a temporary acceleration in the inflation rate. In each case, output growth slowed, reflecting both the permanent decline in natural output and a transitory loss in output. Productivity (and real wages) fell.[18] Generally, the permanent loss in output and productivity and the rise in prices were experienced first, with the temporary surge in inflation (as measured by the GNP deflator) delayed about two quarters.[19] Employment declined much later and for only a few quarters. Cyclical unemployment associated with an oil price rise peaked about six quarters later, before quickly dissipating.

Table 2 shows these developments for the three most recent large energy price hikes. For periods surrounding each oil price hike, the table provides real GNP growth, productivity (business sector output per hour) growth, the rate of increase in the GNP deflator, civilian employment growth, the average unemployment rate for the civilian labor force and money stock (M1) growth. Each measure is provided for the year before and the first four consecutive two-quarter periods following the shock. Two-quarter periods are used to simplify the data presentation, although the timing of energy price effects facilitates the usefulness of this procedure. OPEC1 refers to the first oil price shock which began in IV/1973. OPEC2 begins in II/1979 and IRAQ begins in III/1990.

As table 2 indicates, real GNP growth slowed following the two previous oil price hikes, but did not become negative on a two-quarter basis until after the first two quarters (OPEC1) or after a year (OPEC2). The slowing in output growth reflects both the decline in natural output and, principally later, a temporary cyclical loss in output. Table 2 also shows that the expected productivity decline (negative growth) occurred more quickly than the decline in real GNP in the previous two cases; it began in the first two quarters of the energy price shock in each case. Both productivity and output growth show a sharp cyclical acceleration in the last two-quarter period.

The most recent energy price shock, like the earlier two, was accompanied by an immediate decline in productivity and a slowing in output growth.[20] Output growth became negative earlier than in the previous two cases. Since the recent energy price hike occurred over only two quarters, the period of decline in productivity and output growth should be correspondingly shorter than in the previous two instances. The slight rise in productivity growth in the second two-quarter period is consistent with this expectation.

In the previous two instances, the decline in productivity and natural output was reflected, with about a two-quarter lag, in a sharp and temporary acceleration in the rate of price increase as measured by the GNP deflator. Thus, in the second two-quarter period in OPEC1, inflation accelerated sharply and only temporarily, reflecting the one-time adjustment in the price level. The same acceleration occurs in OPEC2, but with a one-quarter lag; the data for the two-quarter period ending one quarter later are shown in parentheses. As table 2 shows, however, in the first two-quarter period, the rate of increase in the GNP deflator rose (OPEC1) or was unchanged (OPEC2); in the latest instance, it declined.[21]

In the previous two cases, the delayed acceleration in the rate of price increase persisted for about four quarters (five quarters for OPEC2), about as long as the period of sharp increases in energy prices. There is also an acceleration in the recent second two-quarter period (I/1991 and II/1991). Since the latest price hike occurred over half as many quarters as in the previous

[18]These developments were observed in nearly all countries. The notable exception was that income policies impeded the reductions in real wages (and, therefore, in labor productivity) in some countries, especially in 1973-74, so that the effective supply of natural employment fell, further reducing natural output. See Rasche and Tatom (1977a), (1977b) and (1981), Tatom (1988a) and (1987). Hamilton (1983) also provides empirical evidence supporting the permanent effect on U.S. real GNP. Helliwell, Sturm, Jarrett and Salou (1986) provide international evidence on the effect on natural output.

[19]See, for example, Tatom (1981) and (1988a). The lag for the PCE deflator and CPI is shorter (one quarter) and the magnitude is larger for these consumer price series, because the share of energy cost in expenditures is larger for consumer expenditures than for GNP as a whole. Thus, the effect of a given rise in oil prices is larger for consumer price inflation measures. The effects on producer prices occur even faster and are even larger.

[20]Productivity growth had declined more rapidly in the year before the recent oil price shock than it did in the initial two-quarter period, so productivity growth did not actually slow in the second half of 1990.

[21]The initial decline in the rate of price increase in the first two-quarter period is not out of line. In each of the previous initial two-quarter periods, this rate was much lower in at least one of the two quarters. In particular, in the first quarter of 1974, the rate of increase of the deflator fell to a 5.6 percent rate; in 1979, it fell from 9.5 percent in the first quarter to a 9.2 percent rate in the second quarter and to 8.5 percent in the third quarter of 1979.

Table 2
Economic Performance Surrounding Three Energy Price Shocks

	Previous Year	First Two-Quarter Period	Second Two-Quarter Period	Third Two-Quarter Period	Fourth Two-Quarter Period
Real GNP growth rate					
OPEC1	4.4%	0.7%	−2.0%	−5.6%	5.5%
OPEC2	5.3	1.6	1.6	−4.5	6.6
IRAQ	1.0	−0.1	−1.7		
Productivity growth rate					
OPEC1	1.5	−1.3	−2.1	0.9	6.8
OPEC2	0.6	−2.6	0.4	− 0.4	3.7
IRAQ	−0.6	−0.2	0.2		
Rate of price increase					
OPEC1	7.1	7.7	11.5	10.9	7.8
OPEC2[1]	8.9	8.9 (8.4)	8.5 (9.1)	9.4 (10.7)	11.4 (8.7)
IRAQ	4.0	3.2	4.9		
Civilian employment growth rate					
OPEC1	3.5	3.3	0.9	−3.9	1.8
OPEC2	3.9	1.6	1.7	−1.9	2.7
IRAQ	0.9	−1.1	−1.0		
Average unemployment rate					
OPEC1	5.0	5.0	5.4	7.4	8.7
OPEC2	5.9	5.8	6.1	7.5	7.4
IRAQ	5.3	5.7	6.7		
M1 growth rate					
OPEC1	7.0	6.0	3.7	3.7	6.9
OPEC2	7.4	10.3	5.1 (1.5)[1]	5.8	8.0
IRAQ	4.0	3.6	6.8		
Period					
OPEC1	III/1972 to III/1973	IV/1973; I/1974	II and III/1974	IV/1974; I/1975	II and III/1975
OPEC2	I/1978 to I/1979	II and III/1979	IV/1979; I/1980	II and III/1980	IV/1980; I/1981
IRAQ	II/1989 to II/1990	III and IV/1990	I and II/1991		

[1]Data in parentheses are for the two quarters ending one quarter later.

two, the acceleration would be expected to be reversed in the third two-quarter period, even without any effect from the decline in energy prices in I/1991 and II/1991. It remains to be seen whether inflation will decline as abruptly as it did following earlier oil price shocks.[22]

The delayed cyclical response to an energy price hike is seen most clearly by looking at the growth of civilian employment. In the two previous instances, employment growth slowed, but did not become negative until a year after the energy price shock began. Moreover, this decline occurred in only one two-quarter period (the third one), when employment fell at a relatively rapid pace. Thus, the typical recessionary characteristic of falling employment did not oc-

[22]The rate of increase in the CPI rose from a 3.8 percent rate in the second quarter of 1990 to about a 7 percent rate in the third and fourth quarters of 1990. Similarly, the rate of increase of the producer price index rose from a 0.3 percent rate in the second quarter of 1990 to a 6.6 percent rate and a 10.8 percent rate in the third quarter and fourth quarters of 1990, respectively. The rate of increase of the latter two price measures fell sharply in the first half of 1991, reflecting the quicker response of these measures to a rise in energy prices as well as to their subsequent decline.

cur until a year after the onset of the two previous energy price hikes.

The unemployment rate also did not rise immediately after the two previous adverse energy price shocks. In 1973-74, it fell slightly in the fourth quarter of 1973, rose only 0.8 percentage points by the third quarter of 1974, then peaked 3.3 percentage points higher three quarters later.[23] The unemployment rate peaked six quarters after the initial surge in energy prices, in the last period shown in the table. In the second quarter of 1979, the initial quarter of OPEC2, the unemployment rate also fell slightly, then rose gradually for the next three quarters so that it was only 0.4 percentage points higher in I/1980 than it was before the energy price shock. The unemployment rate then rose 1.4 percentage points to a peak in III/1980, six quarters after the initial energy price surge.[24]

In the most recent case, the unemployment rate rose immediately, climbing from 5.5 percent in July 1990 to 7 percent in June 1991. Such a rise is substantially different from the pattern in the initial stages of the previous energy price shocks.

Its behavior might better be understood in the context of the slowing in U.S. economic activity that began in 1988. For example, civilian employment actually began declining sharply in March 1990, five months before the energy price hike; civilian employment fell at a 0.9 percent rate from March to July 1990 and declined further at a 0.5 percent rate from July to October 1990, when energy prices peaked; from October 1990 to August 1991, such employment fell at a 1.3 percent rate. Thus, the path of economic activity downward into recession had begun well before energy prices rose.[25]

A Comparison of Changes in Monetary Policy Actions

Each of the two previous oil shocks were followed by changes in monetary policy actions. There is no clear initial pattern, as money growth slowed in the initial two quarters in 1973-74 but accelerated in 1979. As shown at the bottom of table 2, however, in each case, M1 growth then slowed sharply during the second two-quarter period, at the same time that the rate of price increase temporarily accelerated.[26] Then, in each instance, M1 growth accelerated sharply in the fourth two-quarter period following the sharp rise in the unemployment rate.

The expectation that the economy would quickly experience a recessionary rise in unemployment because of the 1990 oil price rise was widespread. There were equally widespread warnings against repeating the "typical" policy response of easing monetary policy to combat this unemployment.[27] While there is evidence of rising unemployment and subsequent accelerations in M1 growth following previous oil price surges, these changes came more than a year after the initial oil price rise. These changes also occurred after the substantial slowing of M1 growth and the transitory inflation rate hike that are more closely associated with the oil price increases.

In the most recent case, money (M1) growth slowed from a 4.8 percent rate from IV/1989 to II/1990 to a 3.7 percent rate in III/1990 and to a 3.5 percent rate in IV/1990. Money growth quickly reversed course, however, accelerating to a 6.8 percent rate, as the unemployment rate continued to rise in the first half of 1991. This

---

[23]One explanation for the initial decline in the unemployment rate when oil prices rise relies on the capacity loss and "sticky" prices. The initial fall in productivity and initial absence of a price-related decline in aggregate demand when oil prices rise require that producers raise employment to offset some of the output loss and avoid larger-than-desired depletion of inventory. See Tatom (1981) and Ott and Tatom (1986) for discussions of this effect. Rasche and Tatom (1977a) show that employment rose during the first three quarters of the 1973-74 oil shock and did not fall until five quarters later.

[24]In this second instance, a further rise in energy prices late in 1980 and early in 1981 contributed to a further rise in the unemployment rate about a year later, from IV/1981 to II/1982.

[25]Other analysts have emphasized this point. See Weidenbaum (1990) and Erceg and Leovic (1990), for example.

[26]After late 1982, monetary policymakers placed relatively more emphasis on M2 instead of M1. Another measure, the adjusted monetary base, is often a convenient summary measure of monetary policy actions. Higher energy prices significantly raise relative currency demand one quarter later, reducing monetary aggregates relative to the adjusted monetary base; see Tatom (1990). Thus, monetary base growth is less useful as an indicator of monetary policy during energy price shocks. Bullard (1991) discusses these and other indicators of monetary policy and the potentially conflicting signals they offer.

[27]For example, according to Trehan (1990), "Researchers have generally concluded that the Fed eased policy to overcome the reduction in output caused by the oil embargo" and "... the Fed's initial response to the second oil shock also was similar to its response to the first oil shock." See also, Council of Economic Advisers (1991), which indicates that policy was excessively stimulative prior to the previous oil shocks so that it lacked credibility, making efforts to ease ineffective. The Council of Economic Advisers (p. 80) suggests such temporary actions would be appropriate and effective today.

acceleration in M1 growth occurred earlier than it had following the previous oil price hikes, although it did follow both a previous slowing in M1 growth and a recessionary rise in the unemployment rate, just as had similar accelerations in M1 following the two previous energy price increases.[28]

CONCLUSION

The rise in oil prices from August to October 1990 set in motion renewed concern and confusion over both the effects of oil price hikes and the appropriate monetary policy response. Three views achieved widespread acceptance. First, the economy was believed to be less sensitive to oil price hikes than it had been earlier. Second, it was widely believed that the principal and most immediate effect would be a cyclical decline in output and employment. Third, analysts believed that the Fed would ease policy, as it had when faced with this problem in the past.

These views are at odds with previous experience. In 1990, the share of oil imports in GNP and energy per unit of GNP had not fallen to the level before the first oil price shock in 1973. Moreover, the relevant parameter, the share of energy in cost, had not fallen below its 1973 level either. Thus, U.S. economic performance should not have become less sensitive to oil price shocks than it was before. In addition, negative employment growth and an acceleration in money growth had not characterized the initial year of previous energy price shocks.

Earlier evidence suggests that the principal cost of an energy price hike is the loss in capacity output and productivity. A counterpart of this loss is a one-time surge in the general level of prices, which follows the energy price hike relatively closely. The adverse cyclical consequences of past shocks occurred later. The principal policy response following previous oil price hikes was a slowing in money growth. Later, when inflation declined and the unemployment rate rose sharply, money growth accelerated.

The 1990 oil price rise occurred against the backdrop of a slowing in money and output growth that had been under way since late in 1988. Thus, the expected productivity decline and temporary surge in inflation were accompanied by a continuing decline in employment and cyclical output loss. While these developments were uncharacteristic of the initial effects of previous oil price hikes, monetary growth slowed in the second half of 1990 anyway.

There were other distinguishing features associated with the 1990 oil price hike. Foremost among them was its brevity: it occurred over a three-month period and was nearly reversed in another five months. Thus, while the response of output, productivity and prices appears consistent with the capacity-loss-induced effects associated with previous oil price doublings, the subsequent decline in oil prices from October 1990 to March 1991 can be expected to result in offsetting price, output and employment movements.

REFERENCES

Anderson, Gerald H., Michael F. Bryan, and Christopher J. Pike. "Oil, the Economy and Monetary Policy," Federal Reserve Bank of Cleveland *Economic Commentary* (November 1, 1990).

Berndt, Ernst. "Energy Price Increases and the Productivity Slowdown in United States Manufacturing," *The Decline in Productivity Growth*, Federal Reserve Bank of Boston Conference Series, no. 22 (1980), pp. 60-89.

Berndt, Ernst R., and David O. Wood. "Energy Price Shocks and Productivity Growth: A Survey," in Robert C. Gordan, et. al., eds., *Energy: Markets and Regulation, Essays in Honor of M.A. Adelman* (MIT Press, 1987), pp. 305-67.

Brinner, Roger E. "Finding Keys to Turn," *U.S. Forecast Summary* (DRI/McGraw Hill, October 1990).

Bullard, James B. "The FOMC in 1990: Onset of Recession," this *Review* (May/June 1991), pp. 31-53.

Corcoran, Patrick J. "'Putty' and 'Clay'—A Parable for Potential GNP," *Prudential Economics*, The Prudential (February 15, 1990), pp. 6-9.

Council of Economic Advisers. "Oil Price Shocks and Economic Policy," chapter 3 in *Economic Report of the President* (1991), pp. 79-110.

Darby, Michael R. "The U.S. Productivity Slowdown: A Case of Statistical Myopia," *American Economic Review* (June 1984), pp. 301-22.

Denison, Edward F. *Trends in American Economic Growth*, 1929-1982 (The Brookings Institution, 1985).

_______ . "Explanations of Declining Productivity Growth," *Survey of Current Business* (August 1979, Part II), pp. 1-24.

Erceg, John J., and Lydia K. Leovic. "The Outlook After the Oil Shock: Between Iraq and a Soft Place," Federal Reserve Bank of Cleveland *Economic Commentary* (October 15, 1990).

[28]M2 shows the same pattern. It grew at a 2.5 percent rate from II/1990 to IV/1990, down from a 4 percent rate in II/1990 or the 5.1 percent rise in the two-quarter period ending in II/1990. In the first half of 1991, M2 growth also rose, but only to a 4.2 percent rate. Bullard (1991) indicates that Fed decisionmakers were keenly aware of the policy dilemma and chose to pursue a course of neither easing nor tightening. He indicates that there was a concern for *actual* inflationary pressures late in 1990, but concern for the cyclical consequences of the oil price hike was framed only in terms of the potential risk.

Feldstein, Martin. "The Fed Must Not Accommodate Iraq," *Wall Street Journal*, August 13, 1990.

Fieleke, Norman S. "Oil Shock III?" *New England Economic Review* (September/October 1990), pp. 3-10.

Fischer, Stanley, Rudiger Dornbusch, and Richard Schmalensee. *Economics*, 2nd ed. (McGraw Hill Book Company, 1988).

Hamilton, James D. "Oil and the Macroeconomy since World War II," *Journal of Political Economy* (April 1983), pp. 228-48.

Helliwell, John, Peter Sturm, Peter Jarrett, and Gerard Salou. "The Supply Side in the OECD's Macroeconomic Model," *OECD Economic Studies* (Spring 1986), pp. 75-131.

Hickman, Bert, Hillard G. Huntington, and James L. Sweeney. *Macroeconomic Impacts of Energy Shocks* (North Holland Press, 1987).

"How Big An Oil Shock," *The Economist* (August 11, 1990), pp. 12-13.

Kahn, George A., and Robert Hampton, Jr. "Possible Monetary Policy Responses to the Iraqi Oil Shock," Federal Reserve Bank of Kansas City *Economic Review* (November/December 1990), pp. 19-32.

Karnosky, Denis S. "The Link Between Money and Prices—1971-76," this *Review* (June 1976), pp. 17-23.

May, Todd Jr. "Oil Prices Won't Bring a Recession," Fortune Forecast/Special Report, *Fortune* (September 10, 1990), pp. 51-52.

Olson, Mancur. "The Productivity Slowdown, The Oil Shocks and The Real Cycle," *Journal of Economic Perspectives* (Fall 1988), pp. 43-69.

Ott, Mack, and John A. Tatom. "Are Energy Prices Cyclical?" *Energy Economics* (October 1986), pp. 227-36.

Perry, George L. "The Anomalous Recession," *The Brookings Review* (Spring 1991), p. 51.

Rasche, Robert H., and John A. Tatom. "Energy Price Shocks, Aggregate Supply and Monetary Policy: The Theory and the International Evidence," in Karl Brunner and Allan H. Meltzer, eds., *Supply Shocks, Incentives and National Wealth*, Carnegie-Rochester Conference Series on Public Policy (North Holland, 1981), pp. 9-93.

_______ . "The Effects of the New Energy Price Regime on Economic Capacity, Production and Prices," this *Review* (May 1977a), pp. 2-12.

_______ . "Energy Resources and Potential GNP," this *Review* (June 1977b), pp.10-24.

Reinhart, Vincent. "Reading the Effects of an Energy Shock in Financial Markets," *Journal of Economics and Business* (1991), pp. 115-32.

Shin, David. "International Comparisons of Energy-Gross National Product Ratios," American Petroleum Institute Discussion Paper 68 (June 1991).

"Shocked Again," *The Economist* (August 11, 1990), p. 23.

Tatom, John A. "The Effects of Financial Innovations on Checkable Deposits, M1 and M2," this *Review* (July/August 1990), pp. 37-57.

_______ . "Are The Macroeconomic Effects of Oil Price Changes Symmetric?" in Karl Brunner and Allan H. Meltzer, eds., *Stabilization Policies and Labor Markets*, Carnegie-Rochester Conference Series in Public Policy (North Holland, 1988a), pp. 325-68.

_______ . "Macroeconomic Effects of the 1986 Oil Price Decline," *Contemporary Policy Issues* (July 1988b), pp. 69-82.

_______ . "The Macroeconomic Effects of the Recent Fall in Oil Prices," this *Review* (June/July 1987), pp. 34-45.

_______ . "Energy Prices and Short-Run Economic Performance," this *Review* (January 1981), pp. 3-17.

Trehan, Bharat. "Lessons from the Oil Shocks of the 1970s," Federal Reserve Bank of San Francisco *Weekly Letter* (November 9, 1990).

Weidenbaum, Murray. "Recessionary Signs Were Evident Even Before Oil Crisis," *Christian Science Monitor*, October 16, 1990.

Yanchar, Joyce. "A Look at Oil Market Fundamentals," Special Study, *DRI/McGraw-Hill U.S. Review* (August 1990), pp. 25-31.

Michelle R. Garfinkel and
Daniel L. Thornton

Michelle R. Garfinkel, an assistant professor at the University of California at Irvine, was a senior economist at the Federal Reserve Bank of St. Louis while this paper was written. Daniel L. Thornton is an assistant vice president at the Federal Reserve Bank of St. Louis. *Richard I. Jako provided research assistance.*

Alternative Measures of the Monetary Base: What Are the Differences and Are They Important?

THE MONETARY BASE, adjusted for changes in reserve requirements, is a measure intended to summarize the net effect of all Federal Reserve actions on the money stock.[1] As such, it serves as an indicator of the effects of monetary policy actions on the money stock.

Because of the Federal Reserve Bank of St. Louis' long-standing interest in monetary policy, it began publishing a series on the adjusted monetary base in August 1968.[2] Nearly 11 years later, the Federal Reserve Board began publishing an alternative series. While the objective of the Board's series is the same, it has always differed from the series constructed by the St. Louis Fed in a number of respects. These differences have changed over time with changes in the structure of reserve requirements and, thus, changes in the methods of calculating the respective series.[3]

The two series have been used separately and, on occasion, jointly to address a number of issues of importance in monetary theory and policy. Occasionally, they have yielded significantly different results.[4] Moreover, at times

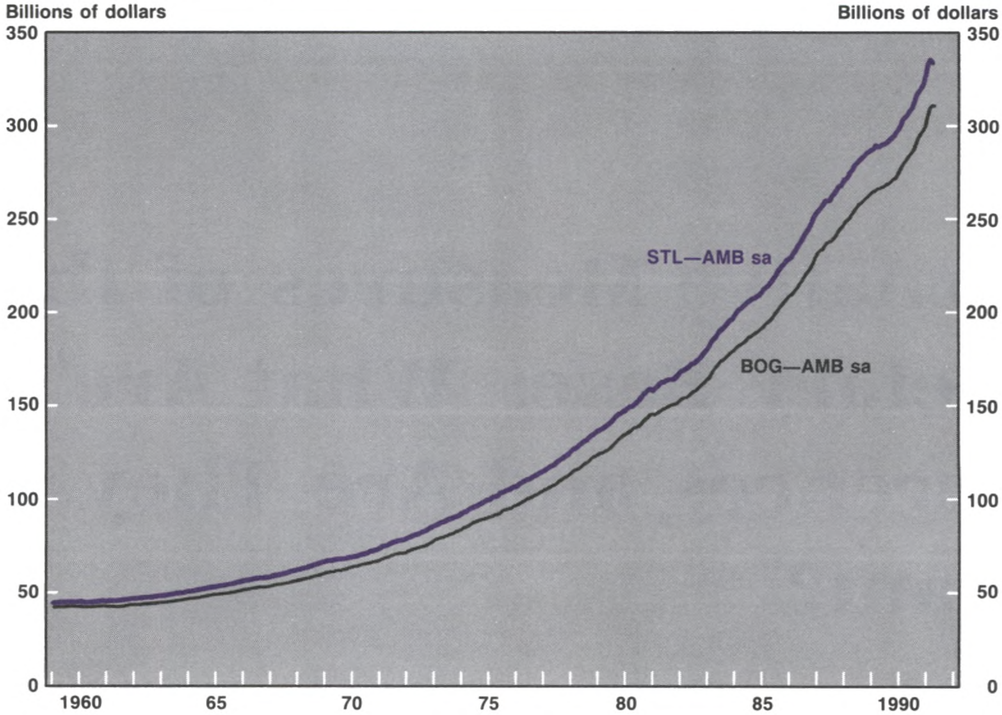
[1]The idea of adjusting the monetary base to reflect changes in reserve requirements was proposed initially by Karl Brunner (1961) in an effort to formulate an "empirically significant theory" of the money supply process. Brunner called this adjustment "liberated reserves." He was the first to compile data on the adjusted monetary base and empirically investigate the relationship between his measure and the money supply. This research agenda was pursued by both him and Allan Meltzer in a number of articles dealing with the money supply process and monetary policy.

[2]See Andersen and Jordan (1968).

[3]One of the most dramatic changes in the structure of reserve requirements occurred with the implementation of the Monetary Control Act of 1980 (MCA). Garfinkel and Thornton (1989) discuss the changes in the structure of reserve requirements brought about by the phase-in of the MCA. Garfinkel and Thornton (1991) discuss the effects of the MCA on the money supply process and the usefulness of the monetary base as an indicator of the effect of monetary policy on the money stock. See Burger and Rasche (1977) and Gilbert (1987) for two significant changes in the calculation of the St. Louis adjusted monetary base.

[4]See Friedman (1988), McCallum (1988a,b), Haslag and Hein (1990) and Muelendyke (1990) for some examples of differential performance.

Figure 1
Levels of the STL-AMB and BOG-AMB, Seasonally Adjusted



Data plotted from January 1959 thru April 1991.

they have presented conflicting pictures of monetary policy.

Because of changes in their calculation and the recent conflicting results, now seems to be an appropriate time to re-examine these series to see how and why they differ, both historically and currently. We can also investigate whether they are likely to continue to behave differently in the years ahead and whether their differential performance is attributable to fundamental differences or merely to arbitrary differences in their construction. Finally, we will provide preliminary evidence on whether the existing difference is potentially important for money stock control.

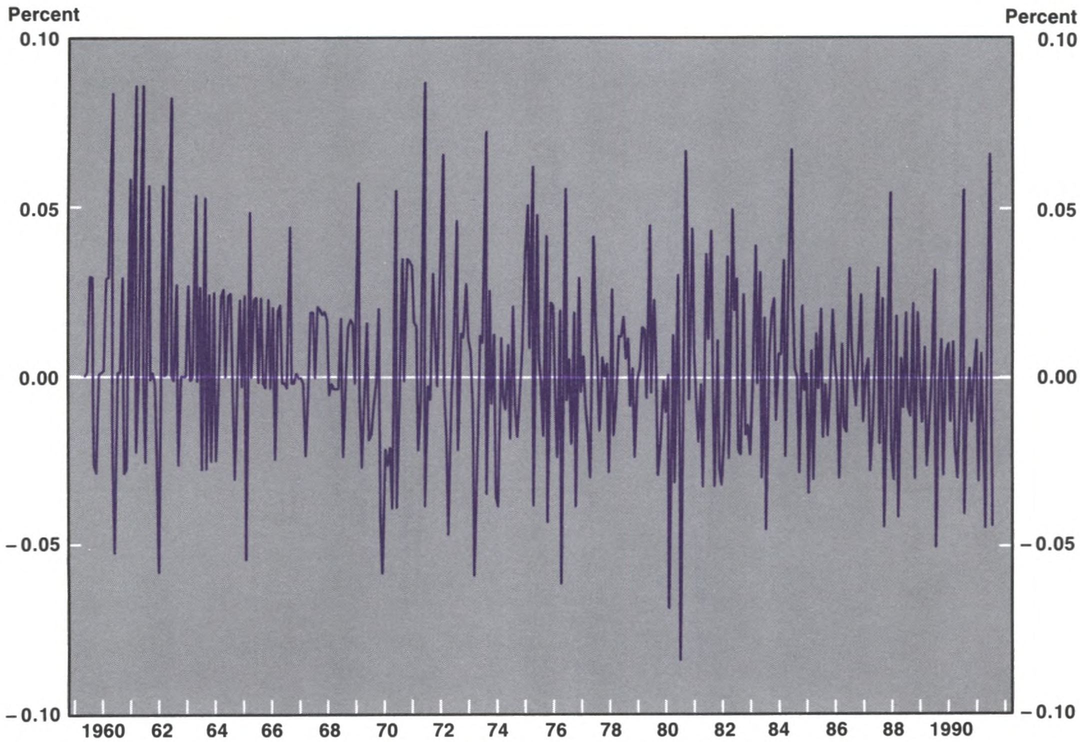
THE DIFFERENCE BETWEEN THE TWO ADJUSTED MONETARY BASE MEASURES

The adjusted monetary base series constructed by the St. Louis Fed (hereafter labeled STL-AMB) and that constructed by the Board of Governors of the Federal Reserve System (hereafter labeled BOG-AMB) were designed for the same purpose. Each is intended to isolate the effects of monetary policy actions—that is, changes in the supply of reserves and changes in reserve requirements—in a single measure. Nevertheless, the two series are quite different.

How Do They Differ?

The difference in their behavior can be seen in figure 1, which shows both adjusted monetary base measures, seasonally adjusted, from January 1959 to April 1991. Although the two measures behave similarly throughout the sample period, STL-AMB is always larger than BOG-AMB and the spread between them increases over time. As will be discussed later, much of this spread can be attributed to differences in the reserve requirement measure used to calculate the two series. Currently, this difference is

Figure 2
Difference Between the Growth Rates of STL-AMB and BOG-AMB



Data plotted from February 1959 thru April 1991.

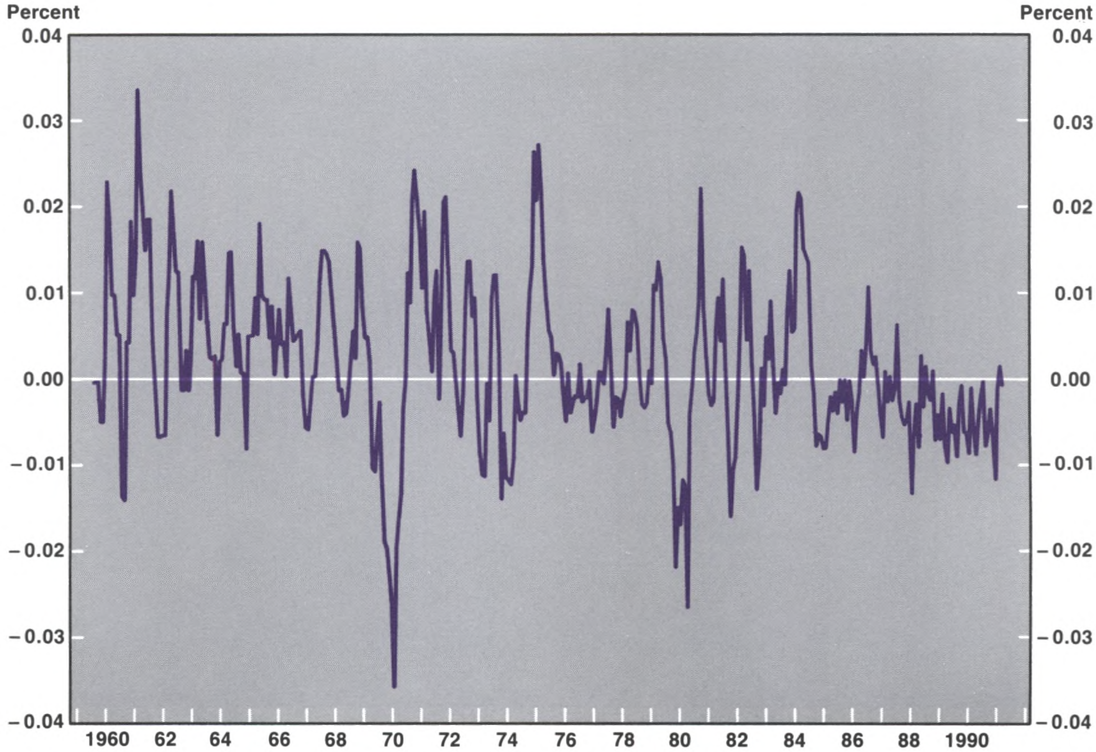
due to the weight each assigns to the level of transaction deposits. As a consequence, the recent widening in the spread between the two AMB series is driven by the growth of deposits.[5]

Although the difference in the level of the two series gets larger over time, figure 2 shows that the difference in the monthly compounded annual growth rates of the two adjusted monetary base measures does not exhibit a significant trend. Indeed, although the monthly difference in the growth rates ranges from −8.4 percent to 8.6 percent, the average difference in their growth, 0.18 percent, is not significantly different from zero at the 5 percent significance level (the t-statistic is 1.35). Thus, over a sufficiently long period, the growth rates of the two series are nearly identical.

Over shorter periods, however, the differences in the growth rates of these aggregates persist, as illustrated by a six-month moving average of the difference between the growth rates of the two series—presented in figure 3. The six-month moving average, which ranges from −3.6 percent to 3.4 percent, shows that,

[5]Thus, the series should be strongly cointegrated. Haslag and Hein (1990), using data from 1959 through 1989, find that the two adjusted monetary base measures are cointegrated. Their results, based on a procedure suggested by Engle and Granger (1987), indicate that the null hypothesis of no cointegration can be rejected at the 5 percent level. This hypothesis, however, cannot be rejected at the 1 percent level. Nevertheless, alternative tests for cointegration, using a procedure developed by Johansen (1988), add further support to the notion that the two series are cointegrated [This procedure and others are discussed in Dickey, Jansen and Thornton (1991).] The chi-square statistic for the null hypothesis that there is one cointegrating vector is 74.5, compared with a critical value at the 1 percent significance level of 22, which provides evidence that the two monetary base series are cointegrated, as expected.

Figure 3
A Six-Month Moving Average of the Difference Between the Growth Rates of STL-AMB and BOG-AMB



Data plotted from July 1959 thru April 1991.

for periods as long as six months, the two adjusted monetary base measures can give considerably different pictures of monetary policy. In fact, as an examination of figure 3 shows, the growth rates of the two series can differ significantly even for fairly long periods of time. For example, the six-month moving average of the difference in the growth rates was strictly positive before April 1969 and nearly always negative after August 1984. The difference in the monthly growth rates averaged 0.60 percent during the former period and −0.32 percent during the latter, and both differences are statistically significant (the t-statistic is 7.96 in the former period, 7.04 in the latter).

Why Do They Differ?

Much of the difference in the two series is attributable to the method of adjusting for reserve requirement changes. Each adjustment creates an index of reserves that would have been held during some "base period." The magnitude of this adjustment, hereafter referred to as RAM, reflects the reserves that are absorbed (released) if the reserve requirements were higher (lower) than those in effect during the base period.[6]

When the STL series was first created, a bank's reserve requirements depended on its location, the type of deposits it held (transaction vs. non-transaction) and whether it was a mem-

[6]Only the STL refers to this adjustment as RAM—short for reserve adjustment magnitude. For expository convenience, this article also will refer to the BOG's adjustment for reserve requirements as RAM.

ber of the Federal Reserve System.[7] The adjustment for reserve requirement changes was made by multiplying the deposits in a given reserve category by the difference between that category's reserve requirement in the base period and its corresponding reserve requirement in the current period. The initial base period used was from August 1935 to July 1936.[8]

In contrast, the BOG has always used the current period as the base period in calculating its reserve adjustment. Thus, each time reserve requirements change, the BOG revises the historical data to reflect the "current" system of reserve requirements. For example, when reserve requirements are reduced, the BOG calculates the amount of reserves that would have been required had the lower reserve requirement been in effect previously. The actual level of required reserves in past periods are multiplied by the ratio of the new average reserve requirement to the old, thereby creating a new, counterfactual "adjusted reserve" series. In this example, the new historical series for adjusted reserves would be lower than the previous historical series. The lower level of the "new" historical series relative to the current period's levels reflects a hypothetical release of reserves in the past brought about by the decrease in requirements in the current period.

In essence, both the STL and BOG methods create counterfactual series for adjusted reserves. For STL, the counterfactual series is the reserves that would have been held if the historical reserve requirements were in effect today. Hence, if reserve requirements have been reduced (raised), actual reserves would be lower (higher) than adjusted reserves. For the BOG, the counterfactual series is the reserves that would have been held in the past if the current reserve requirements were in effect. Actual reserves held in the past would be higher (lower) than adjusted reserves if reserve requirements were reduced (raised) in the current period.

With the STL approach, it is not necessary to revise the historical monetary base series each time reserve requirements are changed. As noted above, the BOG series, in contrast, requires that all historical data be revised, which means there is a delay in the availability of data. Consequently, the STL approach has a comparative advantage for empirical research over the BOG approach because it produces a series that is readily available at the time it is most needed—when there is a change in reserve requirements.[9] Beyond this advantage, the base-period distinction between these approaches appears to be unimportant.

A Recent Change in the Construction of the STL Series

In 1987 the STL adjusted monetary base series was revised in response to fundamental changes in the structure of reserve requirements associated with the Monetary Control Act (MCA) of 1980. Currently, the series is obtained by splicing two adjusted monetary base series with RAMs based on different systems of reserve requirements. Before November 1980, RAM is based on *average* ratio of reserves to deposits, for transaction and non-transaction deposits, during the period from January 1976 to August 1980. After November 1980, RAM is based on the structure of *marginal* reserve requirements on transaction deposits in effect under the MCA. These series are spliced together at the first reserve maintenance period (Novem-

[7]Prior to MCA, non-member banks did not have to maintain reserves with the Federal Reserve System. At the time the Federal Reserve Act was passed, reserve requirements were different for "central reserve city," "reserve city" and "country" banks. This distinction was not based explicitly on the size of the institution, but on the location of the bank at the time the Federal Reserve Act was passed. Over time, this classification system became less meaningful, with many large banks classified as country banks.

[8]See Burger and Rasche (1977), Gilbert (1980) and Tatom (1980).

[9]Although reserves absorbed or released by changes in reserve requirements typically are offset through open market operations so that there is no marked change in the adjusted monetary base, RAM does change significantly. It should be noted, however, that the base period for the St. Louis series has changed when fundamental changes in the structure of reserve requirements made it difficult or impossible to continue using the original base period. At these times, the historical data were revised. The first of these changes occurred in 1977 when Burger and Rasche (1977) altered the series by both changing the method used to calculate RAM and adjusting the series to account for the significant change in the structure of reserve requirements in November 1972. The next occurred in December 1980 when the base period was changed to the period from January 1976 to August 1980. The most recent, discussed next, was in 1987.

ber 19, 1980) under the new reserve requirements imposed by the MCA.[10]

At the splice date, the second part of the series is calculated under the assumption that the marginal reserve requirements on reservable categories of time and savings deposits are zero.[11] Reserve requirements on all non-transaction deposits were eliminated in December 1990. Consequently, after November 1980, the STL series utilizes the present structure of reserve requirements for its base period. Because the base period for the BOG's series is always the current one, the "base-period" distinction between the two series has been virtually eliminated for the period since November 1980.[12] This distinction remains relevant for the pre-November 1980 data, however.

Another distinction remains, however. STL calculates its RAM using the *marginal* reserve requirement on transaction deposits, 12 percent, while the BOG uses the *average* reserve requirement ratio at the time of the last change in reserve requirements. Currently, this ratio is about 8 percent for transaction deposits.[13] As a result, the level of STL-AMB is larger than that of BOG-AMB.

DETERMINING THE SOURCE OF THE DIFFERENCES BETWEEN THE TWO SERIES

Although the difference in adjustment methods for reserve requirement changes accounts for much of the difference exhibited by the two monetary base series, it is not the entire source of their differential behavior. There are two other potentially significant sources: the treatment of vault cash and the seasonal adjustment methods used. Each of these three sources and their empirical importance is discussed below. (A more detailed discussion of the current construction of the two series and their differences appears in the appendix.)

The Treatment of Vault Cash

Currently, the two adjusted monetary base series start with slightly different "raw" data. Both unadjusted monetary base measures, roughly speaking, are the sum of reserve balances held by depository institutions at Federal Reserve Banks and currency in circulation—in other words, currency held by the public, including depository institutions. The differences in these raw data lies primarily in the treatment of vault cash—that is, cash held by depository institutions in their vaults. The BOG, in contrast to STL, adjusts its series for the timing of reserve requirements as satisfied by depository institutions with vault cash.

To get a better understanding of this difference, it is helpful to review briefly the Federal Reserve's system of reserve requirements. Under the current system, depository institutions are required to hold reserves in the form of vault cash and/or reserve balances at the Federal Reserve equal to a fixed percentage of their reservable deposit liabilities—specifically, transaction deposits held by the public, government and foreigners.[14] A depository institution's required reserves are calculated on the basis of the transaction deposits it holds during a two-week period ending every other Monday. An institution can satisfy its requirements with deposit balances at the Federal Reserve during the two-week reserve-maintenance period ending

[10]The procedure scales the "older" part of the series down to the level of the "newer" part of the series to reach the level consistent with the post-MCA base period. The *growth rates* of the data before the splice date are unaffected by the change in the base-period for the level data. See Gilbert (1987), p. 26, for a detailed discussion of the procedure.

[11]This was done because the data necessary for making a RAM adjustment for non-transaction deposits were not available. See Gilbert (1987).

[12]The only base-period distinction that remains is due to the fact that the Garn-St. Germain Depository Institutions Act of 1982 requires that a certain level of transaction deposits at each depository institution be subject to a reserve requirement of zero and that this rate be adjusted upward with the rise in total reservable liabilities.

[13]Because the average ratio of reserves held against transaction deposits to transaction deposits for the period from January 1976 to August 1980 was 0.12664, the use of the marginal reserve requirement, rather than the average ratio of reserves to deposits previously used, minimizes the difference between the "older" and "newer" series at the splice date.

[14]There are lower reserve requirements on a tranche of deposits for each depository institution. Also, depository institutions are required to maintain reserves against their net checkable deposits with other institutions. Aggregated over all institutions, however, the net deposits are zero. Consequently, in the aggregate, no reserves are held against checkable deposits with other depository institutions. Because reserve balances held against such deposits do not net out to zero for individual institutions, however, reserves held against net "inter-bank" checkable deposit liabilities affects the distribution of required reserves among these institutions.

two days after the period used in computing its reserves, and with vault cash held during a two-week period ending 30 days before the end of the maintenance period. Vault cash used to satisfy statutory reserve requirements is called "applied vault cash."

The BOG's adjustment for vault cash involves a distinction between "bound" institutions, whose statutory reserve requirements exceed their holdings of vault cash, and "non-bound" institutions, whose vault cash exceeds their statutory reserve requirements. Another important distinction is between weekly reporting institutions, called EDDS, and quarterly reporting institutions, called QEDS. In the BOG's adjustment for vault cash, the difference between current vault cash and lagged (applied) vault cash of bound EDDS (see appendix) is excluded from the BOG's raw monetary base series. Hence, the STL and BOG unadjusted, not-seasonally-adjusted series effectively differ by the change in vault cash of bound EDDS.[15]

What is the empirical importance of this difference? Because of limited data availability, an exact measure of the magnitude of the BOG's adjustment to vault cash cannot be obtained over the entire sample. As a proxy, the difference between the STL source base and the BOG's not-break-adjusted monetary base (with no seasonal adjustments) is used.[16] This measure, denoted here by ΔTVC, is depicted in figure 4 for the full sample period from February 1959 to April 1991. As the figure shows, the difference in the treatment of vault cash fluctuates between −$2.3 billion to $1.7 billion, but is positive for most of the sample period. Indeed, ΔTVC averages $.15 billion over the period. While this average is small, both in absolute terms and relative to the difference in the two base measures, it is statistically significant from zero at the 5 percent level (the t-statistic is 7.82).

As discussed in more detail below, the difference in the treatment of vault cash has a pronounced seasonal pattern, which has become more amplified over time. Thus, although the difference in the treatment of vault cash contributes relatively little to the "low-frequency" (quarterly or annual) variation in the difference between the two series, it contributes somewhat to the "high-frequency" (monthly) variation.

The Reserve Adjustment

Before November 1980, there are two basic differences between the adjustments that STL and the BOG use on the raw data for reserve requirement changes. First, as noted previously, STL uses a fixed, historical period, while the BOG uses the current period as the base period. Because the average ratio of required reserves to deposits was substantially higher during this period than it is today, the STL adjustment before November 1980 is significantly larger than the BOG's. Second, the BOG makes an additional adjustment for changes in reserve requirements on applied vault cash (see appendix for details).

These two differences continue to be relevant after November 1980. Because STL uses the marginal reserve requirement of 12 percent, which is larger than the average ratio of reserves to deposits used by the BOG to calculate its RAM, the levels of the series are quite different even after this date. As in the period before November 1980, the BOG makes, as part of its break-adjustment procedure, a separate adjustment to applied vault cash. In addition, since February 1984, with the switch to contemporaneous reserve accounting, the BOG makes a separate break adjustment to its adjustment for lagged vault cash of bound EDDS.

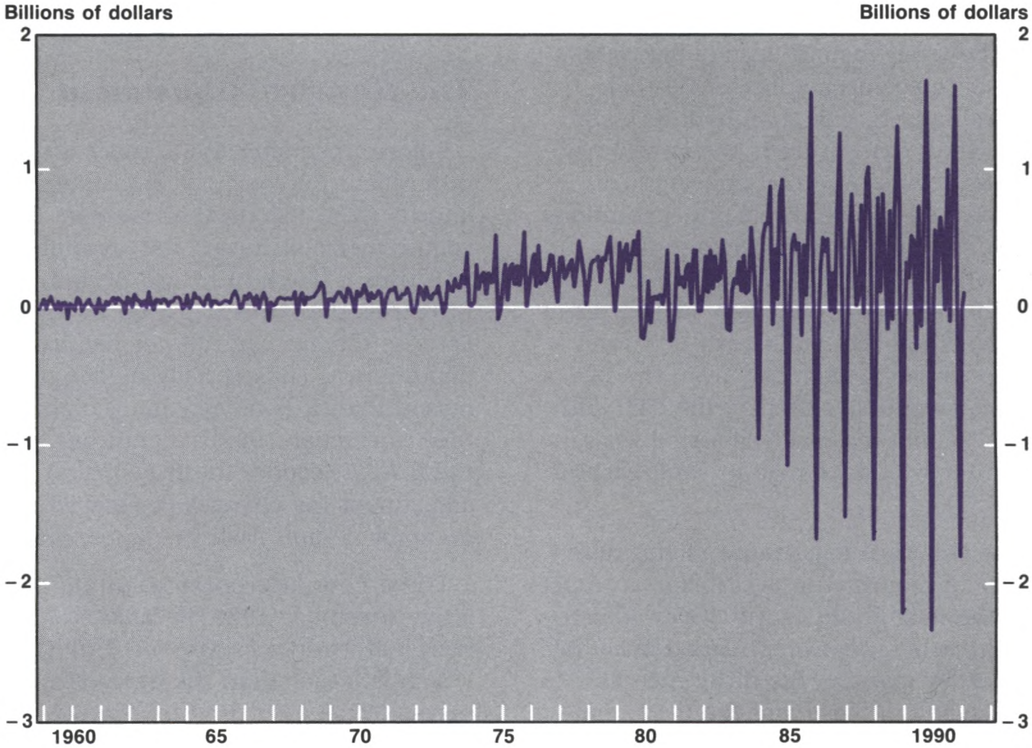
The difference between the STL adjusted monetary base and the BOG's break-adjusted monetary base can be used to measure the empirical magnitude of differences in the method of adjusting for reserve requirement changes. ΔTVC is added to this measure to isolate the ef-

[15]They also differ because of "as-of" adjustments, which the BOG makes but STL does not, and because the BOG includes "required clearing balances" in its not-break-adjusted series.

The rationale for the BOG's adjustment to vault cash is not entirely clear. A recent Board memo states that the adjustment is made on the belief that current vault cash constrains the lending activities of non-bound institutions, while lagged vault cash is the relevant constraint for bound institutions.

[16]As noted in the appendix, required clearing balances and "as-of" adjustments are included in the Board's not-break-adjusted monetary base series, but not in the source base. (See table A1 of the appendix for details). Required clearing balances and these "as-of" adjustments, however, are not included in the Board's break-adjusted series and, thus, play no role in explaining the difference between the two adjusted monetary base measures. To isolate the effect of the difference in the treatment of vault cash on the difference between the two series, required clearing balances and the as-of adjustments (when these data are available) are removed from the difference between the two unadjusted base series.

Figure 4
Difference in the Treatment of Vault Cash ΔTVC



Data plotted from January 1959 thru April 1991.

fect of the difference in RAM from that of the BOG's different treatment of vault cash. The resulting series, denoted here by ΔRAM, and the difference between the two seasonally adjusted, adjusted monetary base measures, denoted by ΔAMBSA, are presented in figure 5. The figure shows that most of the difference between the levels of the two seasonally adjusted bases is, in fact, explained by differences in the method used to adjust for reserve requirement changes.

Seasonal Adjustment

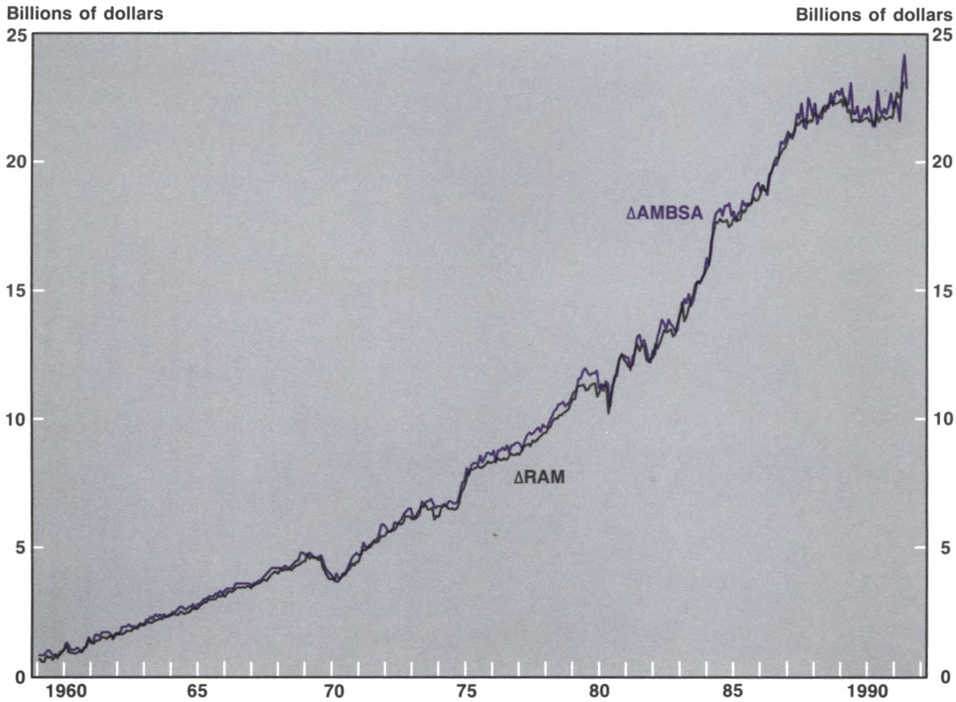
To remove regular variations in the AMB series due to seasonal factors, both series are seasonally adjusted. For monthly and quarterly data, STL adjusts its monetary base series by simply applying the standard X-11 seasonal adjustment program to its not-seasonally-adjusted series. Weekly data are seasonally adjusted with a separate program that inputs unadjusted weekly data and seasonally adjusted monthly data.[17]

The BOG seasonally adjusts weekly data using a model-based approach; it then obtains seasonally adjusted monthly and quarterly data from the seasonally adjusted weekly data.[18] In contrast to STL, the components of the base are seasonally adjusted separately: break-adjusted required reserves against transaction deposits, the break-adjusted measure of surplus vault

[17]See Zeller (1972). Also, see Gilbert (1985) for a discussion of the change in seasonal adjustment associated with the switch to contemporaneous reserve accounting.

[18]See Pierce, Grupe and Cleveland (1984) and Farley and O'Brien (1987) for a discussion of the seasonal adjustment procedures used by the Board.

Figure 5
Difference Between STL-AMB and BOG-AMB and the Difference Between Their Reserve Adjustments, ΔAMBSA and ΔRAM



Data plotted from January 1959 thru April 1991.

cash used by the BOG, and currency held by the nonbank public.[19]

The difference in the two series due to different seasonal adjustment procedures is presented in figure 6. This difference, denoted by ΔSAM, is measured by subtracting the BOG's seasonal adjustment (the difference between the seasonally adjusted and the not-seasonally-adjusted, break-adjusted monetary base) from STL's seasonal adjustment (the difference between the seasonally adjusted and the not-seasonally-adjusted, adjusted monetary base). Not surprisingly, the average difference in the series due to differences in seasonal adjustment is essentially zero over the sample period.[20] Nevertheless, ΔSAM ranges from –$1.6 billion to $2.6 billion, suggesting that differences in the seasonal adjustment are a source of high-frequency variation in the difference between the two series.

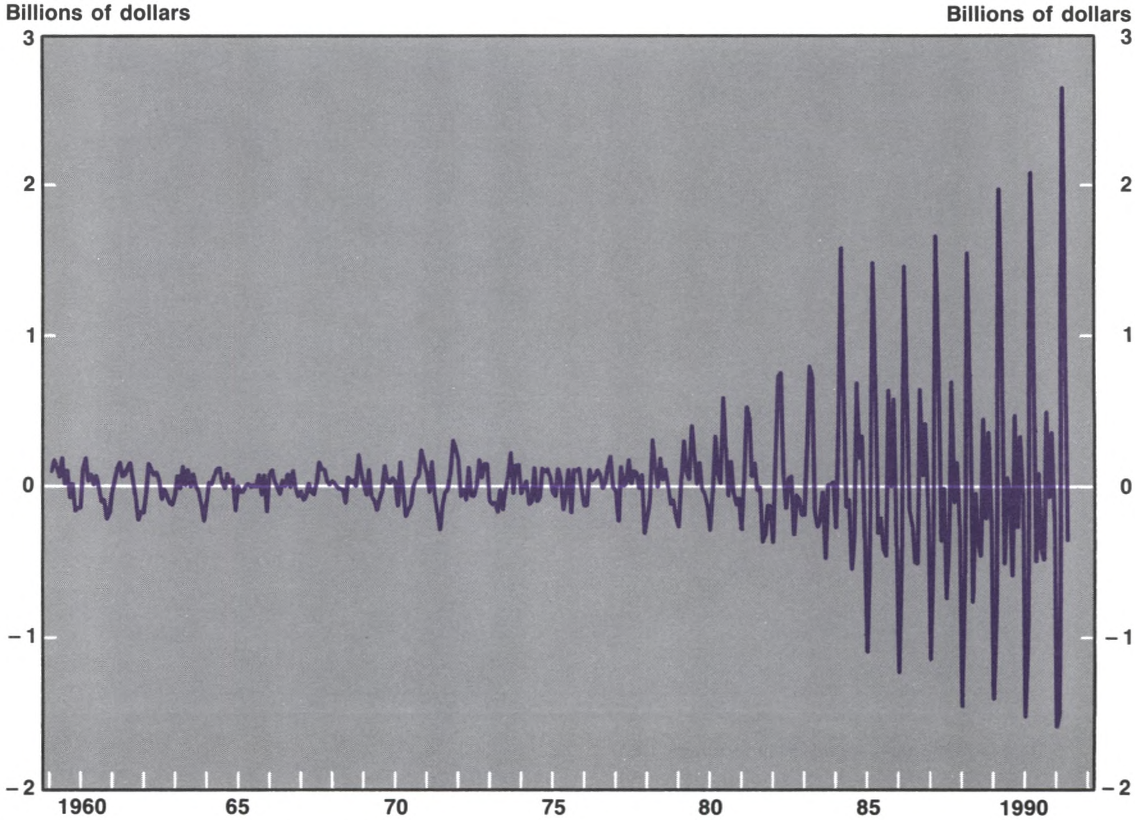
The Relative Importance of These Differences Over Time

As indicated by figures 4 and 6, the contribution of the differences in the treatment of vault cash and the seasonal adjustment have become more variable starting shortly after the MCA, especially around 1984. Moreover, after 1984, both differences have large seasonal components. ΔTVC becomes larger and more variable, perhaps because the BOG's adjustment for lagged

[19]All components of the monetary base, excluding excess reserves, are adjusted as a whole before the switch to contemporaneous reserve accounting. For the series after the switch, the weekly series is adjusted by component using a model-based procedure and then is modified to be made consistent with the monthly series.

[20]The average difference is $.025 billion, with a t-statistic of 1.14 for the test of the null hypothesis that the difference is zero.

Figure 6
Difference in the Seasonal Adjustment ΔSAM



Data plotted from January 1959 thru April 1991.

vault cash at bound EDDS does not change with the Fed's move from lagged to contemporaneous reserve accounting. Separate break adjustments are made to this series and to lagged vault cash at QEDS starting in February 1984, however.

The difference due to seasonal adjustments becomes larger with higher frequency variation about this same time. This may be the result of the Board's changing its seasonal adjustment method in February 1984.[21]

The seasonal variations in these two series tend to offset each other so that the difference between STL-AMB and BOG-AMB does not have a large seasonal component. Indeed, ΔTVC and ΔSAM are highly negatively correlated after 1984—the simple correlation between changes in the two series after January 1984 is −.84.

As shown in table 1, which reports the variances of ΔAMBSA, ΔRAM, ΔTVC, ΔSAM and (ΔTVC + ΔSAM) for various subperiods, the variability in ΔAMBSA has also increased over the entire sample period. The subperiods correspond to various reserve requirement regimes. The full sample is broken at the time of the implementation of the MCA and the time of its effective completion, which coincides with the switch to contemporaneous reserve accounting.

[21]As noted above, the BOG seasonally adjusts its break-adjusted monetary base as a whole for data before the switch to contemporaneous reserve accounting in February 1984. Thereafter, it has used a model-based approach to seasonalize the components of the base separately for weekly data. In addition, as discussed by Gilbert (1985), the seasonal factors used by the St. Louis Fed also changed around that time.

Table 1

The Variances of Changes in the Difference between STL-AMB and BOG-AMB and Changes in its Sources

Period	ΔAMBSA	ΔRAM	ΔTVC	ΔSAM	ΔTVC + ΔSAM
1959.02-1991.04	.093	.034	.245	.290	.104
1959.02-1980.11	.033	.023	.016	.019	.025
1980.12-1991.04	.216	.058	.725	.858	.269
1980.12-1984.01	.120	.072	.069	.119	.094
1984.02-1991.04	.258	.051	1.011	1.181	.346

Table 2

Simple Correlations between Changes in the Difference between STL-AMB and BOG-AMB and Changes in its Sources

Period	ΔRAM	ΔTVC	ΔSAM	ΔTVC + ΔSAM
1959.02-1991.04	.206	.175	.334	.826
1959.02-1980.11	.563	.233	.488	.614
1980.12-1991.04	.022	.177	.333	.887
1980.12-1984.01	.528	.227	.421	.669
1984.02-1991.04	−.156	.180	.334	.925

As the table shows, the variance of changes in the differences of the two seasonally adjusted AMB series has increased throughout the sample period, especially since February 1984, when the variances of both ΔTVC and ΔSAM sharply increase. Because these series are negatively correlated, however, this increased variability is not reflected in a similar rise in the variance of these combined series (ΔTVC + ΔSAM). Nevertheless, the simple correlations between changes in ΔAMBSA and both changes in ΔRAM and changes in (ΔTVC + ΔSAM) presented in table 2 show that more of the month-to-month changes in ΔAMBSA is attributable to changes in ΔTVC + ΔSAM since January 1984.

ALTERNATIVE MEASURES OF THE MONETARY BASE: ARE THE DIFFERENCES IMPORTANT?

Since the adjusted monetary base is intended to be a summary measure of the policy actions of the Federal Reserve, an important question arises: are the differences in the STL and BOG measures of the adjusted monetary base important?[22] Over a sufficiently long period of time, the answer to this question is an emphatic, "No!" As noted earlier, over sufficiently long periods of time, the average difference in the growth rates of the two series is negligible. Nevertheless, over shorter

[22]Garfinkel and Thornton (1991) have shown that the relationship between the money supply and the adjusted monetary base has weakened since the MCA. More important, they argue that the usual linear relationship between the money supply and the monetary base, as a model of the money supply process, fails to perform well since then.

Table 3

Estimate of Changes in M1 on Changes in Alternative Measures of the Adjusted Monetary Base Monthly Results

	St. Louis				Board			
Period	Constant	ΔAMB	$\bar{R}^2$	D.W.	Constant	ΔAMB	$\bar{R}^2$	D.W.
1959.02-1980.10	.080	2.210*	.424	1.69	.169	2.237*	.348	1.53
	(0.89)	(13.86)			(1.75)	(11.81)		
1980.11-1984.02	.730	2.025*	.210	2.11	.444	2.594*	.195	2.15
	(0.96)	(3.37)			(0.51)	(3.23)		
1984.03-1991.04	.344	2.135*	.299	1.12	−.039	2.480*	.257	1.03
	(0.53)	(6.11)			(0.05)	(5.51)		
1984.03-1990.11	−1.156	3.244*	.396	1.27	−2.239*	4.139*	.414	.81
	(1.56)	(7.32)			(2.64)	(7.58)		

* Indicates statistical significance at the 5 percent level.

periods, the two measures could lead to different interpretations of monetary policy.

The adjusted monetary base has been used frequently in theoretical and empirical models of money stock control. Hence, one way to assess whether the difference in the two AMB measures is important is to see if either measure explains more of the movements in the money stock, M1.[23]

Some preliminary evidence on the relative performance of these alternative measures in controlling the money stock can be obtained by regressing changes in M1 on changes in the alternative base measures. Estimates of these equations using monthly data are reported in table 3.[24] The results are reported for three subperiods from February 1959 to April 1991 based on important changes in the construction of the two series. The results indicate that, in all cases, there there is a statistically significant relationship between changes in M1 and changes in each of the base measures.[25] The relationship between M1 and either adjusted monetary base appears to have deteriorated since the effective implementation of the MCA in February 1984. This result appears to be due, however, to the sharp rise in depository institutions' holdings of excess reserves following the December 1990 elimination of reserve requirements on non-transaction deposits and the sensitivity of ordinary-least-squares regression analysis to outliers. When the last five months of data are deleted, the adjusted R-squares are similar to those obtained over the first period—indeed, the adjusted R-square for the BOG's measure rises by nearly 20 percent.[26]

The relationship between M1 and the adjusted monetary base is somewhat tighter when the STL measure is used for all three periods; this is not the case when the last five months are deleted

[23]M2 is not included in the analysis because, since December 1990, there is no direct relationship between the non-M1 components of M2 and policy actions. M2 can only be controlled directly through its M1 component. Equations similar to those reported here, using M2, produce results broadly similar to those here, except where noted below.

[24]These equations are simple and are not intended to be *the* specification of the money supply process over the time period. Moreover, as Garfinkel and Thornton (1991) have shown, such a simple specification of the money supply function is inappropriate since February 1984. The strong serial correlation of the residual since then is evidence of this misspecification using either base measure. For these reasons, the results here are intended to be illustrative.

[25]The equations were also estimated using the growth rates of M1 and the two AMB measures. In all instances, the adjusted R-squares were smaller when growth rates were used. Qualitatively, the results are the same as those in the text when quarterly and annual data are used.

[26]As expected, after February 1984, there is no statistically significant relationship between M2 and either base measure, at the monthly or quarterly frequency.

Table 4

Estimates of Changes in M1 on Changes in Alternative Measures of the Adjusted Monetary Base Quarterly and Annual Results

	St. Louis				Board			
Period	Constant	ΔAMB	$\bar{R}^2$	D.W.	Constant	ΔAMB	$\bar{R}^2$	D.W.
QUARTERLY RESULTS								
II/59-IV/79	.151	2.254*	.774	1.52	.428*	2.266*	.732	1.41
	(0.75)	(16.77)			(2.06)	(14.99)		
I/80-I/84	−1.250	3.086*	.507	2.32	−1.858	3.678*	.406	2.54
	(0.50)	(4.18)			(0.58)	(3.45)		
II/84-I/91	.850	2.148*	.231	.37	1.892	2.014*	.156	.41
	(0.23)	(3.02)			(0.47)	(2.44)		
II/84-III/90	−8.096*	4.406*	.524	.42	−7.089	4.360*	.394	.46
	(2.14)	(5.34)			(1.55)	(4.15)		
ANNUAL RESULTS								
1959-90	−2.321	2.794*	.757	.82	−.829	2.85*	.716	.77
	(0.77)	(9.72)			(0.26)	(8.75)		

* Indicates statistical significance at the 5 percent level.

from the last period.[27] Overall, the two measures appear to differ little in their relationship to monthly M1 growth.

Because the difference in the growth rates of the two AMB series declines at lower frequencies, the equations also were estimated using quarterly data and annual data for the period from 1959 to 1990. These results are summarized in table 4. The quarterly estimates are similar to the monthly ones; the overall performance is better, however, using quarterly data during the first period and somewhat worse during the last. Again, much of the deterioration in the last period is due to including the quarters during and immediately after the elimination of reserve requirements on non-transaction deposits.

For quarterly data, the St. Louis measure always explains somewhat more of the variation in M1 growth. Estimates using annual data provide a similar result, with the STL series explaining about 4 percent more than the BOG series of the annual variation in M1.

CONCLUDING REMARKS

While no formal tests of the difference in the performance of the two base measures for money stock control have been made here, the differences might be statistically significant. Indeed, this has been the case in other applications. For example, Haslag and Hein (1990) have reported statistically significant differences in the explanatory power of the two monetary bases for nominal GNP. More importantly, Friedman (1988) and McCallum (1988a,b) report substantive differences in the performance of the two measures for monetary policy analysis.

The problem is that the two measures differ by their reserve adjustments, their treatment of vault cash and their methods of seasonal adjustment. While differences in the reserve adjustment procedures account for the bulk of the discrepancies, differences arising from the treatment of vault cash and seasonal adjustments have become more important in explaining short-run variations between the two series in the 1980s. Because there is little objective rea-

[27]Also, the BOG's measure produces a higher adjusted R-square than does the STL measure in the last period when monthly growth rates are used.

son to prefer one method of technical adjustment over the other, there is little basis for choosing one measure over the other in empirical studies, when the two measures produce substantially different results.

In instances where the results are qualitatively the same but quantitatively different, such as the results reported here or by Haslag and Hein, the researcher must be content to choose the measure that performs "best" for the problem at hand. If the problem is money stock control, the preliminary evidence presented here suggests that the St. Louis measure holds an edge.

REFERENCES

Andersen, Leonall C. and Jerry L. Jordan. "The Monetary Base-Explanation and Analytical Use," this *Review* (August 1968), pp. 7-11.

Brunner, Karl. "A Schema for the Supply Theory of Money," *International Economic Review* (January 1961), pp. 79-109.

Burger, Albert E. "Alternative Measures of the Monetary Base," this *Review* (June 1979), pp. 3-8.

Burger, Albert E., and Robert H. Rasche. "Revision of the Monetary Base," this *Review* (July 1977), pp. 13-28.

Dickey, David A., Dennis W. Jansen and Daniel L. Thornton. "A Primer on Cointegration with an Application to Money and Income," this *Review* (March/April 1991), pp. 58-78.

Engle, Robert T., and C. W. J. Granger. "Co-Integration and Error Correction: Representation, Estimation and Testing," *Econometrica* (March 1987), pp. 251-76.

Farley, Dennis E., and Yueh-Yun C. O'Brien. "Seasonal Adjustment of the Money Stock in the United States," *Journal of Official Statistics* (1987), pp. 223-33.

Friedman, Benjamin M. "Conducting Monetary Policy by Controlling Currency Plus Noise: A Comment," *Carnegie-Rochester Conference Series on Public Policy* (Autumn 1988), pp. 205-12.

Garfinkel, Michelle R. and Daniel L. Thornton. "The Multiplier Approach to The Money Supply Process: A Precautionary Note," this *Review* (July/August 1991), pp. 47-64.

_______ ."The Link Between M1 and the Monetary Base in the 1980s," this *Review* (September/October 1989), pp. 35-52.

Gilbert, R. Alton. "A Revision in the Monetary Base," this *Review* (August/September 1987), pp. 24-29.

_______ . "New Seasonal Factors for the Adjusted Monetary Base," this *Review* (December 1985), pp. 29-33.

_______ . "Calculating the Adjusted Monetary Base under Contemporaneous Reserve Requirements," this *Review* (February 1984), pp. 27-32.

_______ . "Two Measures of Reserves: Why Are They Different?" this *Review* (June/July 1983), pp. 16-25.

_______ . "Revision of the St. Louis Federal Reserve's Adjusted Monetary Base," this *Review* (December 1980), pp. 3-10.

Haslag, Joseph H., and Scott E. Hein. "U.S. Monetary Policy Measures: Are They Roughly Equivalent?" mimeo, Federal Reserve Bank of Dallas (1990).

Johansen, Soren. "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control* (June/September 1988), pp. 231-54.

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

_______ . "Reply to Comments by Benjamin Friedman," *Carnegie-Rochester Conference Series on Public Policy* (Autumn 1988b), pp. 213-14.

Meulendyke, Ann-Marie. "Possible Roles for the Monetary Base," in *Intermediate Targets and Indicators for Monetary Policy: A Critical Survey* (Federal Reserve Bank of New York, July 1990) pp. 20-66.

Pierce, David A., Michael R. Grupe, and William P. Cleveland. "Seasonal Adjustment of the Weekly Monetary Aggregates: A Model-Based Approach," *Journal of Business and Economic Statistics* (July 1984), pp. 260-70.

Tatom, John A. "Issues in Measuring an Adjusted Monetary Base," this *Review* (December 1980), pp. 11-29.

Zeller, Louis. "Weekly Seasonal Adjustment Program for the IBM S/360 Computer," unpublished manuscript, Board of Governors of the Federal Reserve System (February 15, 1972).

Appendix
The Construction of the Two Adjusted Base Measures Since February 1984

This section discusses how each series, not seasonally adjusted, is constructed in two steps: the raw data series (called "source base" by STL and "not-break-adjusted monetary base" by the BOG) and the adjustment for reserve requirement changes. Table A1 shows the construction of the STL source base and the analogous BOG series and summarizes the differences between them. Similarly, table A2 shows the adjustments made to each series and their differences.

The Raw Data

The St. Louis Series. The STL source base measure is the sum of reserve balances at Federal Reserve Banks and currency in circulation—i.e., currency held by the public. As shown in panel A of table A1, reserve balances are de-

Table A1

Calculating the St. Louis Source Base and the Board's Not-Seasonally-Adjusted, Not-Break-Adjusted Monetary Base (since February 1984)

A. STL:	$290.759		
= RESERVE BALANCES		34.090	
= total reserve balances			35.714
− required clearing balances			(−) 1.624
+ CURRENCY IN CIRCULATION		256.669	
= currency held by the nonbank public			223.000
+ current total vault cash			33.669
B. BOG:	$292.131		
= TOTAL RESERVES		62.931	
= total reserve balances			35.714
− required clearing balances			(−) 1.624
+ applied vault cash			28.841
+ CURRENCY HELD BY THE NONBANK PUBLIC		223.542	
+ SURPLUS VAULT CASH (BOG)		3.550	
= current total vault cash			33.619
− applied vault cash at non-bound banks			(−) 3.331
− applied vault cash at bound QEDS			(−) .196
− current vault cash at bound EDDS			(−) 26.543
+ REQUIRED CLEARING BALANCES		1.624	
+ FLOAT-PRICING AS-OFs		.484	
C. STL − BOG	−$1.372[1]	−.879	
= − ADJUSTMENT FOR LAGGED VAULT CASH AT BOUND EDDS		(−) −1.229	
= current vault cash at bound EDDS			26.543
− applied vault cash (lagged) bound EDDS			(−) 25.314
− REQUIRED CLEARING BALANCES		(−) 1.624	
− FLOAT-PRICING AS-OFs		(−) .484	

NOTE: Example data are for January 1990. Columns might not add due to rounding errors.

[1]This difference does not equal the sum of column 2, section C, because the BOG's figure for currency held by the nonbank public and vault cash used in calculating their adjusted monetary base differ from their published series. The published data are monthly averages of daily figures, while the monthly data used to calculate their monetary base are obtained by prorating data averaged over reserve maintenance periods.

fined as total reserve balances net of required clearing balances of depository institutions.[1]

The Board of Governors' Series. As shown in panel B of table A1, the BOG's not-break-adjusted monetary base is essentially the same as the STL source base. Roughly speaking, the not-break-adjusted monetary base is the sum of total reserves and currency. Total reserves used in the calculation is defined as the sum of total reserve balances, net of required clearing balances, and applied vault cash.[2] Required clearing balances plus currency held by the nonbank public and surplus vault cash are added to the measure of reserves. In constructing this monetary base series, the BOG defines "surplus vault cash" as *current* vault cash net of *lagged* (applied) vault cash held by both non-bound and bound depository institutions that report quarterly

[1]Required clearing balances are deposit balances that depository institutions are required to maintain at the Federal Reserve to ensure that the dollar volume of their check clearings and other transfers of funds through the Federal Reserve System are covered. These balances are subtracted from total reserve balances since they do not satisfy statutory reserve requirements and, hence, do not constrain a depository institution's expansion of deposit liabilities explicitly.

[2]As noted in the main text, applied vault cash refers to the vault cash held by depository institutions during the two-week period ending 30 days before the end of the current maintenance period.

Table A2

Calculating the St. Louis Adjusted Monetary Base and the Board's Break-Adjusted Monetary Base, Not Seasonally Adjusted (since February 1984)

A. STL in period t – s:	$299.961		
= SOURCE BASE (t – s)			290.759
+ RESERVE ADJUSTMENT MAGNITUDE (t – s)			9.202
= $r^T(b)$ × transaction deposits (t – s)			
– required reserves (t – s)			
B. BOG:	$277.636		
= NOT-BREAK-ADJUSTED MONETARY BASE (t – s)			292.131
– FLOAT-PRICING AS-OFs (t – s)		(–)	.484
– REQUIRED CLEARING BALANCES (t – s)		(–)	1.624
– REQUIRED RESERVES AGAINST EURODOLLARS (t – s)		(–)	1.347
+ ADJUSTMENT FOR THE BREAK IN REQUIRED RESERVES (t – s)		(+)	– 12.440
= r^T (t) × transaction deposits (t – s)			
– r^T (t – s) x transaction deposits (t – s)			
– ADJUSTMENT FOR BREAK IN APPLIED VAULT CASH		(–)	– 1.492
+ ADJUSTMENT FOR THE BREAK IN LAGGED VAULT CASH AT BOUND EDDS AND BOUND QEDS		(+)	– .088
C. STL – BOG:	$22.325[1]		21.467
= RAM (STL)			9.202
– ADJUSTMENT FOR THE BREAK IN REQUIRED RESERVES		(–)	– 12.440
+ ADJUSTMENT FOR THE BREAK IN APPLIED VAULT CASH		(+)	– 1.492
– ADJUSTMENT FOR THE BREAK IN LAGGED VAULT CASH AT BOUND EDDS AND BOUND QEDS		(–)	– .088
– ADJUSTMENT FOR LAGGED VAULT CASH AT BOUND EDDS		(–)	– 1.229

NOTES: Example data are for January 1990. The series are constructed for periods i = t,t – 1,t – 2,...,t – s,..., where t is the most recent period. $r^T(b)$ is the reserve requirement in the base period (12 percent). $r^T(i)$ is the average reserve requirement on transaction deposits in period i. As in table A1, the columns might not add due to rounding errors.

[1]This does not equal the sum of column 2, section C.

(QEDS) and net of current vault cash held by bound institutions that report weekly (EDDS).[3]

As shown in panel C of table A1, the key difference between the STL source base and the BOG's unadjusted base measure is the treatment of vault cash held by bound EDDS. In particular, the two measures are essentially identical except that the Board's measure excludes the surplus vault cash of bound EDDS—i.e., current vault cash net of lagged applied vault cash at bound EDDS. The two base series also differ in that the BOG's measure does not remove required clearing balances and includes float-pricing "as-ofs." As noted in the main text and below, however, required clearing balances and float-pricing as-ofs are removed when the Board calculates its break-adjusted series.

Adjustments for Reserve Requirement Changes

The St. Louis Series. Panel A of table A2 shows the computation of the STL reserve adjustment magnitude, RAM, which is simply added to the source base. RAM is the difference between actual required reserves and the reserves that would have been required, given the actual level and composition of deposits held at depository institutions, under the reserve re-

[3]The measure of surplus vault cash reported on the Board's H.3 release is not the measure used in this computation. Rather than using *current* vault cash of all depository institutions, the surplus vault cash in the H.3 release includes *lagged* vault cash of only those institutions subject to reserve requirements. In addition, rather than subtracting current vault cash of bound EDDS, the surplus vault cash reported in the H.3 release subtracts applied vault cash of these institutions.

quirements of a chosen base period. The reserve requirement ratio on transaction deposits currently used for the base period is 12 percent, the marginal reserve requirement in effect since the full implementation of the MCA.[4]

This adjusted base measure indicates what the monetary base would have been given the current level of currency held by the nonbank public and the current level and composition of deposits under the base period's system of reserve requirements. If the required reserve ratio were to be reduced in any period, the adjustment would increase.[5] In this case, the source base would not change initially, but the adjusted monetary base measure would increase as RAM increased, reflecting a release of reserves into the system available to expand deposit liabilities.[6]

The Board of Governors' Series. The BOG adjusts the series for breaks in reserve requirements historically. In particular, it treats the most current period as its base period so that the break-adjusted series and the unadjusted series are identical. Reserves are adjusted with a ratio method. For any period in which the system of required reserves differs from that in the current period, required reserves held against these deposits are multiplied by the ratio of current required reserves to what reserves would have been required under the old system, given the current composition of deposits. As shown in panel B of table A2, this adjustment simplifies to a measure similar to the STL-RAM, where the base-period reserve requirements are replaced by the most recent reserve requirements. An adjustment for breaks in applied vault cash is subtracted and an adjustment for breaks in lagged vault cash at bound EDDS and QEDS is added.[7] In addition, reserves held against Eurodollar deposits are excluded from this series after the switch to contemporaneous reserve accounting.[8]

As summarized in panel C of table A2, the main difference between these series (not seasonally adjusted) revolves primarily around the treatment of vault cash and the method of reserve adjustment.[9]

---

[4]The actual average reserve ratio is lower than 12 percent because, for the individual bank, the reserve ratio is only 3 percent for transaction deposits below the reserve tranch and 12 percent for deposits in excess of that tranch. Even before last year's elimination of reserve requirements on personal time and saving deposits, which was completed in two steps starting in the middle of December 1990 and ending after the first maintenance period in January 1991, the STL-RAM assumed that base period requirements on non-transaction deposits were zero. It should also be noted that both RAMs assume that the base period requirements on Eurodollar deposits is zero since February 1984.

[5]For example, the recent elimination of reserve requirements on non-transaction deposits released about $13.6 billion during the two (two-week) maintenance periods starting in the middle of December.

[6]Typically, however, the Fed offsets the release of reserves generated by a reduction in reserve requirements, at least partially, by removing reserves from the banking system. Such a defensive measure would prevent the money supply from accelerating in response to the reserve requirement change. Nevertheless, the change in reserve requirements would affect the adjusted monetary base provided that the resulting release of reserves were not perfectly offset by such a defensive measure.

[7]The adjustment for the lag in applied vault cash at bound EDDS is not break-adjusted before the switch to contemporaneous reserve accounting. After the switch, the BOG also started to make the break-adjustment to lagged vault cash at QEDS.

[8]Because the STL-RAM assumes that the base period reserve requirements on Eurodollar deposits are zero, the Board's exclusion of reserves held against these deposits plays no role in explaining the difference between the two AMB series.

[9]It should also be noted that the Board's series is adjusted for the annual increase in the reserve tranch as if the change were being phased in gradually over the whole year. In contrast, the St. Louis measure adjusts for the change when it occurs. See Meulendyke (1990), p. 59.

Piyu Yue

Piyu Yue, a research associate at the IC² Institute, University of Texas as Austin, was a visiting scholar at the Federal Reserve Bank of St. Louis when this article was written. *Lynn Dietrich provided research assistance. The author wishes to thank Professor Leon Lasdon for the GRG2 Fortran code and Professor Douglas Fisher for providing the data. Estimations that appear in this article were calculated at the University of Texas at Austin and are the responsibility of the author.*

A Microeconometric Approach to Estimating Money Demand: The Asymptotically Ideal Model

THE DEMAND FOR MONEY plays a critical role in macroeconomics. In conventional money demand analysis, the demand for real money balances is typically expressed as a function of such variables as real income, the expected rate of inflation and the nominal interest rate.[1] Empirical investigations using these variables have not been particularly useful in predicting the demand for money or in formulating and evaluating monetary policy.[2]

More recently, a number of researchers have attempted to estimate money demand in a manner consistent with microeconomic foundations. Even in these cases, however, the empirical results have been largely discouraging.[3]

This paper reviews the general micro-econometric approach to estimating the demand for money, culminating with an advanced micro-econometric model, called the Asymptotically Ideal Model (AIM). AIM is applied to U.S. time-series data and the results are compared briefly with those from previous studies. AIM results are consistent with microeconomic theory and provide insight into the behavior of money demand in the 1970s and 1980s.

[1]See Friedman (1956) for one of the most comprehensive discussions of the money demand function.

[2]These functions have been subject to several unexplainable shifts and often imply a larger liquidity effect than is typically experienced. Perhaps the most dramatic example of this phenomenon occurred in the early 1980s with the yet unexplained break in the income velocity of M1. For this and other examples, see Goldfeld (1976), Friedman (1984), Lucas (1988) and Rasche (1990).

[3]Frequently, the estimated own price elasticities of demand for monetary assets are positive, implying that their demand curves slope upward. For example, see Serletis (1988), Fisher (1989) and Moore, Porter and Small (1990).

MICROECONOMIC MODELING

As a result of developments in macroeconomic theory over the past two decades, "almost all macroeconomists agree that basing macroeconomics on firm microeconomic principles should be higher on the research agenda than it has been in the past."[4] Problems arise, however, when aggregate, macroeconomic data are used to estimate microeconomic-based models of money demand. Some of these problems are illustrated by a simple example that uses two approaches to microeconomic modeling: the demand function approach and the utility function approach.[5]

The Demand Function Approach

Consider an economy where the representative consumer allocates income between a composite consumer good, A, and a monetary asset, M, that yields monetary services. The consumer's objective is to maximize the utility function (subject to a budget constraint), given the price of the composite commodity and the user cost of the monetary asset. Let P, u and E denote the price level (the price of A), the nominal user cost of holding one real unit of M and total expenditures (or income), respectively.[6] The consumer's decision problem is expressed by

$$\text{Max } f(A, M) = A^r M^{1-r},$$

$$\text{subject to } PA + uM = E.$$[7]

For simplicity, the utility function, f, is Cobb-Douglas, where r, an unknown parameter, characterizes the consumer's taste or preference. The optimal solution to the consumer's decision problem yields ordinary demand functions for A and M. In this case, the demand functions are:

$$(1)\quad A = rE/P = r/(P/E) = G_1(u/E, P/E, r), \text{ and}$$

$$(2)\quad M = (1-r)E/u = (1-r)/(u/E) = G_2(u/E, P/E, r).$$

The demands for A and M are functions, G_1 and G_2, respectively, of E, P, u and the unknown parameter r. Because the budget constraint is linear in P and u, the normalized price, P/E, and user cost, u/E, can replace P, u and E. In general, demand functions can be expressed by normalized prices (including the user cost) and the unknown parameter. This parameter can be estimated by simultaneously fitting equations (1) and (2) using data on real quantities of A and M and the normalized price and user cost.

This approach is called the "demand function" approach because estimation begins after demand functions are specified. For this approach to yield meaningful estimates, however, the specified system of demand functions must correspond to the neoclassical utility function from which they were derived. Consequently, the conditions for estimating the system of demand functions are fairly restrictive. For instance, the Rotterdam model (a well-known demand system used in empirical studies) requires specific forms for demand functions and specific constraints on parameters during estimation.[8]

Even if these conditions are satisfied, however, the Rotterdam model is still highly restrictive because the assumed underlying utility function (either Cobb-Douglas or CES) is a member of a narrow class of utility functions with constant elasticities of substitution.[9]

---

[4]Mankiw (1990), page 1658.

[5]Some economists argue that aggregate data cannot be applied to microeconomic models without considering the problems of aggregation. Aggregation problems are not discussed in this paper, although the aggregation error might be one source of the unsatisfactory performance of conventional money-demand functions.

[6]The user cost of holding a unit of a real monetary asset is computed by the formula, $u = p^*(t)\,[R(t) - i(t)]/[1 + R(t)]$, where $p^*(t)$ is the "true" cost of living index defined as the geometric average of the consumer price index and the consumption goods deflator, R(t) is the benchmark interest rate or the maximum rate in the economy at each period and i(t) is the interest rate on the monetary asset. The formula is derived from a widely applicable consumer decision model.

[7]Distinct views about money have resulted in two approaches to analyzing consumer demand for money. In the first approach, money is viewed as a commodity which provides a monetary service flow to holders. Thus, real balances of the monetary assets directly enter the consumer's utility function along with real consumption. In the other approach, money is viewed as intrinsically worthless; consumers hold it only to finance current and future consumption. As a result, real money balances do not enter the consumer's utility function per se. Instead, the liquidity cost of holding real money balances is taken into account in the budget constraint. Feenstra (1986) shows that these two approaches are equivalent.

[8]For an application of the Rotterdam model to the money-demand system, see Fayyad (1986). For the theory of the Rotterdam model, see Barnett (1981).

[9]It is easy to encounter difficulties using the demand function approach. The failure to specify functions correctly or impose relevant restrictions can result in biased or inefficient parameter estimates.

The Utility Function Approach

The utility function approach to demand estimation also has been used in empirical studies. To understand this approach, reconsider the consumer's decision problem and the demand functions shown in equations 1 and 2. In the utility function approach, demand functions for A and M are substituted into the utility function, f(A, M), to obtain the indirect utility function,

$$h(v_1, v_2, r) = f[A(v_1, v_2, r), M(v_1, v_2, r)],$$

where $v_1 = P/E$, $v_2 = u/E$. Because the indirect utility function has properties that are the inverse of those for the utility function, it is more convenient to use the reciprocal of the indirect utility function,

$$F(v_1, v_2, r) = 1/h(v_1, v_2, r).$$[10]

By definition, demand functions can be expressed in terms of their expenditure shares, $s_1 = AP/E$ and $s_2 = Mu/E$. That is,

$$A \equiv s_1/v_1 \text{ and } M \equiv s_2/v_2.$$[11]

In this way, demand functions can be obtained without solving first-order conditions. Consequently, no matter how complicated the utility function might be, the derivation of share equations and demand functions is straightforward.

Of course, if the utility function is relatively simple and well-behaved (for example, when the Cobb-Douglas function is used), there is no need to use the utility function approach. However, if the utility function includes more than two goods or is sufficiently complicated, the Lagrange multiplier procedure cannot be used to derive demand functions.

THE SEMI-NONPARAMETRIC METHOD FOR ESTIMATING THE DEMAND SYSTEM

The critical step in applying the utility function approach is the specification of the proper reciprocal indirect utility function, F. To simplify the terminology, the term "utility function" will indicate "the reciprocal of the indirect utility function" in the following discussion.

Flexible Functional Form Modeling

Cobb-Douglas and CES functions have been used extensively in theoretical and applied work because of their relative simplicity. Despite their apparent successes, however, such use has been criticized. For example, if there are more than two goods, the CES utility function can only generate demand systems when each pair of goods has the same constant elasticity of substitution.[12] Unless there is prior information to the contrary, however, the elasticities of substitution should be determined by the data rather than restricted by the specification of the utility function. This limitation has motivated researchers to look for utility functions that are more flexible and allow for data-determined elasticities of substitution.

Flexible functional form models have attracted considerable attention in economics literature since the early 1970s, when it was proposed that the translog and generalized Leontief functions should replace neoclassical utility functions. It was recognized that the values of the elasticities of substitution are determined by the value of the utility function and the values of its first- and second-order derivatives are evaluated at its extreme point. Consequently, if the values of the utility function and these derivatives can be estimated, so too can the elasticities of substitution. This idea forms the basis for the flexible functional form approach.

A functional form is said to be flexible if its level and the first- and second-order derivatives at a point in its domain are allowed to reach the respective values of the "true" utility function at that point. The true utility function is assumed consistent with the properties of the data, so that, in principle, elasticities of substitution consistent with the data can be estimated.

[10] The duality theory states that if the reciprocal of the indirect utility function, F, is nondecreasing and quasi-concave with respect to normalized prices, the respective utility function, f, must be nondecreasing and quasi-concave with respect to quantity variables. In this sense, the utility function is equivalent to its reciprocal indirect utility function.

[11] Expenditures shares can be obtained by using the modified Roy's *identity from duality theory. That is,*

$s_i = (v_i\, dF/dv_i)/V'\nabla F(V,r)$, i = 1, 2,

where the vector $V' = (v_1, v_2)$ and the gradient vector $\nabla F(V,r) = (dF/dv_1, dF/dv_2)'$. See Diewert (1974).

[12] See Uzawa (1962).

One flexible functional form is derived from a Taylor series expansion where all terms greater than second-order are eliminated, that is,

$$(3)\ F = a_0 + \Sigma_i a_i x_i + \Sigma_i \Sigma_j a_{ij} x_i x_j.^{13}$$

This approximation is flexible because it has enough free coefficients, a_0, a_i, a_{ij}, to allow for any desired value of the first- and second-order derivatives of function F.

Two frequently used flexible functional forms, the translog and generalized Leontief functions, are given by simple substitution into equation 3. For the translog function,

$$F = \ln(f(x)) \text{ and } x_i = \ln(q_i),$$

where f denotes the utility function and q_i represents the quantity of good i. The generalized Leontief function is attained by letting

$$F = (f(x))^{1/2} \text{ and } x_i = (q_i)^{1/2}.$$

The coefficients in these functional forms can be estimated and, in turn, the demand system and the elasticities of substitution can be derived.

Caveats For Flexible Functional Forms

Theoretically, the second-order Taylor approximation can attain flexibility only at a single point or in an infinitesimally small region. Hence, estimates of the elasticities of substitution are valid only for the range of observations covered by the sample data. Therefore, the second-order Taylor series approximation should be viewed as "locally flexible."

Such models are also subject to another, potentially more serious, problem. Experience has demonstrated that regularity conditions are frequently violated! Therefore, the restrictions that microeconomic theory imposes on consumer behavior are not embedded in these flexible functional forms. This point is illustrated later in the empirical section of this paper.

In an attempt to solve these problems, microeconomists have developed a variety of flexible functional forms that maintain their flexibility and have larger regularity regions.[14] A family of such flexible functional form models has been proposed (for example, Barnett's (1981) minflex Laurent model).[15] To gain *global regularity*, however, additional constraints are imposed on the parameters which result in a loss of *local flexibility*. This tradeoff between flexibility and regularity is characteristic of flexible functional form modeling. None of these models is both globally regular and globally flexible.[16]

Semi-nonparametric Method

Gallant (1981) created the "semi-nonparametric method" to remove the local flexibility limitation. His method specifies a series of models that approximate the underlying utility function at every point in the function's domain. Hence, the models are globally flexible.

The "semi-nonparametric method" is built upon a well-known result in mathematics: a Fourier series expansion can converge to any continuous function.[17] In contrast to the local convergence of the flexible functional forms, the Fourier series can approximate a continuous function in the entire domain. Gallant proposed to use the Fourier series expansion to specify a series of utility functions that can converge to any neoclassical utility function. Because neoclassical functions are a subset of continuous functions, the property of the Fourier series expansion will guarantee asymptotic convergence to an underlying neoclassical utility function.

Fourier series modeling consists of a series of expansions of models, with succeeding models nested in the preceding one. When the sample size increases, higher-order models can be speci-

[13]This equation is written in general form, where a_0 denotes all the constants (that is, the function evaluated at the point of interest and the partial derivatives evaluated at the same point). The use of this general form to estimate these equations is one of the procedure's limitations because the point about which the expansion is made is estimated by the data, rather than being specified by the researcher. Hence, there is no guarantee that this point will necessarily correspond with the maximum value of the function itself.

[14]The regularity region is the subset of the domain of the utility function in which all regularity conditions are satisfied.

[15]Instead of the Taylor series expansion, Barnett used the Laurent expansion to enlarge the regularity region and maintain enough parametric freedom to satisfy requirements for flexibility. See Barnett (1983).

[16]See Diewert and Wales (1987).

[17]The function must be integrable or, more generally, it must lie in a Hilbert space. See Telser and Graves (1972).

fied by simply adding more terms of the component functions. For instance, the first-order model is defined by the utility function,

$$F = a_0 + \Sigma_i a_i q_i + \Sigma_i \Sigma_j a_{ij} q_i q_j + \Sigma_i b_i \cos(\Omega q_i) + \Sigma_i d_i \sin(\Omega q_i).$$

The j^{th}-order model is defined by the utility function,

$$F = a_0 + \Sigma_i a_i q_i + \Sigma_i \Sigma_j a_{ij} q_i q_j + \Sigma_i \Sigma_j b_{ij} \cos(j\Omega q_i) + \Sigma_i \Sigma_j d_{ij} \sin(j\Omega q_i).$$

Asymptotically, the model contains an infinite number of terms and unknown parameters. Therefore, asymptotic inference based upon the Fourier series expansion models is free from functional-form specification error. This is its principal advantage.

In empirical work, however, the number of terms must be finite. Consequently, the properties of lower-order Fourier models become decisive. The harmonic component functions, such as sines and cosines which are frequently used in engineering and physics, are not suitable in economic applications because they do not satisfy the usual regularity conditions, such as monotonically increasing and strictly quasi-concave. This means that lower-order Fourier series models can violate regularity conditions.[18] Nevertheless, Gallant's approach permitted micro-econometric models to achieve both *global regularity* and *global flexibility*.

The AIM Demand System

To solve the problems of Fourier series models, another infinite function series, called the Muntz-Szatz series, is adopted. A typical form of the series is expressed as:

$$\begin{aligned} f = a_0 &+ \Sigma_i a_i^1 q_i^{1/2} + \Sigma_i a_i{}^2 q_i^{1/4} + \Sigma_i a_i^3 q_i^{1/8} + \ldots \\ &+ \Sigma_i \Sigma_j a_{ij}^1 q_i^{1/2} q_j^{1/4} + \Sigma_i \Sigma_j a_{ij}^2 q_i^{1/2} q_j^{1/8} + \ldots \\ &+ \Sigma_i \Sigma_j b_{ij}^1 q_i^{1/4} q_j^{1/4} + \Sigma_i \Sigma_j b_{ij}^2 q_i^{1/4} q_j^{1/8} + \ldots \end{aligned}$$

The Muntz-Szatz series expansion converges to a continuous function, and any continuous function can be approximated by the Muntz-Szatz series.[19] Consequently, this series can be used to approximate a neoclassical utility function asymptotically.[20]

The Muntz-Szatz series is a linear combination of a set of special power functions. In contrast to the Fourier series, the component functions of the Muntz-Szatz series, $q_i^{1/2}$, $q_i^{1/4}$, ... , $q_i^{1/2} q_j^{1/4}$, ... , are neoclassical functions. In other words, they are monotonically increasing and quasi-concave with respect to variables q_i and q_j. The Muntz-Szatz series is necessarily neoclassical, however, only if all of the coefficients, a_i, a_{ij}^k, b_{ij}^k, ... , are non-negative, because only positive linear combinations of the neoclassical component functions are necessarily neoclassical. As a result, the coefficient-restricted Muntz-Szatz series can approach a neoclassical function but may not approach *any* continuous function. Imposing these restrictions guarantees that the estimated function will not violate regularity conditions.

The Muntz-Szatz series is used in place of the Fourier series in Gallant's semi-nonparametric method. A series of models can be defined by increasing degrees of the Muntz-Szatz approximations. Under the parameter constraint, these models are globally regular; the respective utility functions are neoclassical everywhere in their domain. When the sample size increases, higher-order models can be specified with more free parameters to best fit the data and derive the elasticities of substitution that the data suggest. Hence, the Muntz-Szatz series gives rise to a model that is asymptotically globally flexible. Even a low-order approximation requires a fairly large number of parameters to be estimated, however. Hence, while the model is asymptotically globally flexible, finite samples will limit the researcher's ability to fully utilize this property.

The model has two additional features that make it particularly attractive for applied work. First, although there are a relatively large number of free parameters to be estimated, it is impossible to overfit the noise in the data. Because movements due to measurement errors are irregular and cannot be expressed by the neoclassical component functions, the model simply

[18]Moreover, the Fourier series models can easily overfit the noise of the data. Usually, the measurement errors of economic variables can be decomposed into a pure white noise plus some high-frequency periodic functions. These latter functions might be mistaken for useful information if their frequencies are close to that of the sine and cosine functions in the Fourier series models.

[19]Once again, the function must be integrable or, more generally, lie in a Hilbert space. See Telser and Graves (1972).

[20]See Barnett and Jonas (1983).

ignores them. Also, because the component functions are not periodic, the high-frequency, periodic movements in the data are likewise ignored.

Consequently, models based on the Muntz-Szatz series expansion are globally regular and are asymptotically globally flexible. This is why they are called Asymptotically Ideal Models (AIMs).[21]

ESTIMATION OF THE AIM MONEY DEMAND SYSTEM

Four subaggregated goods are included in the empirical work presented here: a consumer good, A_4, and three monetary assets, A_1, A_2 and A_3. A_4 is an aggregate good of consumer durables, nondurables and services and its respective aggregate price is denoted by p^*. A_1 consists of currency, demand deposits of households and other checkable deposits; A_2 is composed of savings deposits at commercial banks and thrifts, super NOW accounts and money market deposit accounts; and A_3 is small time deposits at commercial banks and thrifts.[22] For each subset, an aggregate quantity is defined as a sum of *per capita* real balances of the component monetary assets.[23] The opportunity cost of holding a unit of a real monetary asset is measured by its user cost, a quantity-share-weighted sum of the individual user costs that compose it. The user cost of A_i is denoted by u_i.

The representative consumer solves an optimal allocation problem by selecting real consumption, A_4, and real balances of the monetary assets, A_1, A_2 and A_3, to maximize the utility function $f(A_1, A_2, A_3, A_4)$, subject to the budget constraint, given p^*, u_1, u_2, u_3 and the total expenditure, E. Following the utility function approach and using the first-order AIM model, we specify the reciprocal indirect utility function for the four goods case as:

$$\begin{aligned}F(v,a) = {} & a_1v_1^{1/2} + a_2v_2^{1/2} + a_3v_3^{1/2} + a_4v_4^{1/2} \\ & + a_5v_1^{1/2}v_2^{1/2} + a_6v_1^{1/2}v_3^{1/2} + a_7v_1^{1/2}v_4^{1/2} \\ & + a_8v_2^{1/2}v_3^{1/2} + a_9v_2^{1/2}v_4^{1/2} \\ & + a_{10}v_3^{1/2}v_4^{1/2} + a_{11}v_1^{1/2}v_2^{1/2}v_3^{1/2} \\ & + a_{12}v_1^{1/2}v_2^{1/2}v_4^{1/2} + a_{13}v_1^{1/2}v_3^{1/2}v_4^{1/2} \\ & + a_{14}v_2^{1/2}v_3^{1/2}v_4^{1/2} + a_{15}v_1^{1/2}v_2^{1/2}v_3^{1/2}v_4^{1/2},\end{aligned}$$

where v_1, v_2, v_3 and v_4 are the normalized prices, and a_1, a_2, ..., and a_{15} are the parameters of the indirect utility function.

The share equation for each good is derived from equations 3 using the modified Roy's identity. These are

$s_1 = S_1/S$, $s_2 = S_2/S$, $s_3 = S_3/S$, and $s_4 = (1 - s_1 - s_2 - s_3)$, where

$$\begin{aligned}S_1 = {} & a_1v_1^{1/2} + a_5v_1^{1/2}v_2^{1/2} + a_6v_1^{1/2}v_3^{1/2} + a_7v_1^{1/2}v_4^{1/2} \\ & + a_{11}v_1^{1/2}v_2^{1/2}v_3^{1/2} + a_{12}v_1^{1/2}v_2^{1/2}v_4^{1/2} \\ & + a_{13}v_1^{1/2}v_3^{1/2}v_4^{1/2} + a_{15}v_1^{1/2}v_2^{1/2}v_3^{1/2}v_4^{1/2}, \\ S_2 = {} & a_2v_2^{1/2} + a_5v_1^{1/2}v_2^{1/2} + a_8v_2^{1/2}v_3^{1/2} + a_9v_2^{1/2}v_4^{1/2} \\ & + a_{11}v_1^{1/2}v_2^{1/2}v_3^{1/2} + a_{12}v_1^{1/2}v_2^{1/2}v_4^{1/2} \\ & + a_{14}v_2^{1/2}v_3^{1/2}v_4^{1/2} + a_{15}v_1^{1/2}v_2^{1/2}v_3^{1/2}v_4^{1/2}, \\ S_3 = {} & a_3v_3^{1/2} + a_6v_1^{1/2}v_3^{1/2} + a_8v_2^{1/2}v_3^{1/2} + a_{10}v_3^{1/2}v_4^{1/2} \\ & + a_{11}v_1^{1/2}v_2^{1/2}v_3^{1/2} + a_{13}v_1^{1/2}v_3^{1/2}v_4^{1/2} \\ & + a_{14}v_2^{1/2}v_3^{1/2}v_4^{1/2} + a_{15}v_1^{1/2}v_2^{1/2}v_3^{1/2}v_4^{1/2} \text{ and} \\ S = {} & a_1v_1^{1/2} + a_2v_2^{1/2} + a_3v_3^{1/2} + a_4v_4^{1/2} + 2a_5v_1^{1/2}v_2^{1/2} \\ & + 2a_6v_1^{1/2}v_3^{1/2} + 2a_7v_1^{1/2}v_4^{1/2} + 2a_8v_2^{1/2}v_3^{1/2} \\ & + 2a_9v_2^{1/2}v_4^{1/2} + 2a_{10}v_3^{1/2}v_4^{1/2} + 3a_{11}v_1^{1/2}v_2^{1/2}v_3^{1/2} \\ & + 3a_{12}v_1^{1/2}v_2^{1/2}v_4^{1/2} + 3a_{13}v_1^{1/2}v_3^{1/2}v_4^{1/2} \\ & + 3a_{14}v_2^{1/2}v_3^{1/2}v_4^{1/2} + 4a_{15}v_1^{1/2}v_2^{1/2}v_3^{1/2}v_4^{1/2}.\end{aligned}$$

Only the first three share equations are independent and can be written generally as:

(4) $s_i = S_i/S = g_i(v,a)$ for $i = 1, 2, 3$.

When the additional parameter normalization $a_1 + a_2 + a_3 + a_4 = 1$ is imposed, one parameter, for example a_4, can be eliminated by substitu-

[21]See Barnett and Yue (1988) and Barnett, Geweke and Yue (1991).

[22]This division of monetary assets is proposed in Fisher (1989), who has performed separability tests over a variety of deposit categories included in M2. He found that this division (A_1, A_2, A_3, A_4) is weakly separable in terms of the General Axiom of Revealed Preference (GARP) test. The weak separability condition offers theoretical assurance that A_1, A_2, A_3 and A_4 are meaningful aggregate goods by the definition of economic aggregation theory. However, the partition, $[(A_1, A_2, A_3), A_4]$ did not pass the GARP test for weak separability. This raises the question of the existence of the monetary aggregate, M2 ($M2 \cong A_1 + A_2 + A_3$). If this is the case, the demand system of consumption and the subaggregate M2 monetary assets are the appropriate way to study demand for money in terms of the economic aggregation theory.

[23]The reader is cautioned that these results are not directly comparable with those using conventional money demand specifications because the latter employ conventional monetary aggregate data that differ in several respects from those used here. First, business demand deposits are excluded from the data used here, so that M1 is not equal to A_1 and M2 is not equal to $A_1 + A_2 + A_3$. Moreover, unlike the conventional monetary aggregates, each of the sub-aggregates is seasonally adjusted separately.

tion. Hence, in the case of four goods, the first-order AIM system contains 14 free parameters.[24]

The share equations are nonlinear with respect to the normalized prices and hence, to income and prices as well. By the definition of expenditure shares, demand functions can be expressed as $A_i = s_i/v_i$. The complicated nonlinearity of the share equations, however, makes it impossible to derive a closed-form expression for the demand functions, such as the conventional linear or log-linear functions of income, prices and interest rates. Fortunately, the estimated parameters and share equations can be used to compute the income and price elasticities for consumer goods and monetary assets.

Estimation of the AIM Demand System

The AIM model is estimated by a maximum likelihood procedure under the assumption that each share equation in (4) has an additive error term, ε_{it}. That is,

(5) $s_i = g_i(v,a) + \varepsilon_{it}, \ i = 1,2,3.$

The disturbances are assumed to be independent and identically distributed multivariate normal random variables with zero mean and covariance matrix Σ. The sample disturbance covariance matrix, $\hat{\Sigma}$, is defined as

$$\hat{\Sigma} = (\sum_{t=1}^{N} \hat{\varepsilon}_t \hat{\varepsilon}'_t)/N,$$

where N is the sample size, and the sample disturbance, $\hat{\varepsilon}_t$, is computed by

$\hat{\varepsilon}_{ti} = s_i - g_i\,(v,a)$, for $i = 1, 2, 3$.

Maximizing the likelihood function for the system is equivalent to minimizing the generalized variance, $|\hat{\Sigma}|$.[25]

The estimation was accomplished using a nonlinear program (GRG2). To find a global optima, an extensive search over a large range of initial conditions was conducted. Because of the complex nonlinearity of the AIM demand system, the true maximum likelihood estimates are difficult to obtain. The possibility of missing the global optima was reduced, however, by an extensive search of the parameter space.[26]

All parameters are subject to non-negativity constraints to guarantee that global regularity conditions are satisfied. Because inequality constraints limit the applicability of the existing theoretical sampling distribution theory, the usual methods for testing hypotheses cannot be used.[27]

Income Elasticities and Price Elasticities

Because the share equations are so complex, AIM does not yield explicit functional forms for demand functions. This is the consequence for correctly embedding utility maximization into an econometrically estimable demand system that can be used to compute economically meaningful income and price elasticities.

The Allen Partial elasticities of substitution and income elasticities are defined and expressed by the following formulas:[28]

for $i \neq j$,

$$\sigma_{ij} = \frac{\partial A_i^c p_j}{\partial p_j A_i^c} = \left[\frac{1}{v_i}\frac{\partial s_i}{\partial p_j} + \frac{s_j}{v_j}\left(\frac{1}{v_i}\frac{\partial s_i}{\partial E} + \frac{s_i}{p_i}\right)\right]\frac{Ev_iv_j}{s_js_i}$$

for $i = j$,

$$\sigma_{ii} = \frac{\partial A_i^c p_i}{\partial p_i A_i^c} = \left[\frac{1}{v_i}\frac{\partial s_i}{\partial p_i} - \frac{s_i}{v_i p_i} + \frac{s_i}{v_i}\left(\frac{1}{v_i}\frac{\partial s_i}{\partial E} + \frac{s_i}{p_i}\right)\right]\frac{Ev_i^2}{s^2_i}$$

where p_i are the prices (and user costs), A_i^c denotes the income-compensated demand functions for the i^{th} asset, s_i denotes the expenditure shares and E denotes total expenditures. The elements, σ_{ij}, constitute a symmetrical matrix called the Allen Partial matrix.

The income elasticities are defined by

$$\eta_{i0} = \frac{\partial A_i E}{\partial E\, A_i} = \frac{\partial s_i}{\partial E}\frac{E}{s_i} + 1,$$

and the uncompensated price elasticities are denoted by

[24]The higher-order AIM system contains many more unknown parameters; see Barnett and Yue (1988).

[25]See Barnett (1981).

[26]To estimate parameters, a_i $i = 1, 2, 3, 5, \ldots, 15$, initial values of the unknown parameters were assigned. A number of different initial values of a_i, 0.01, 0.03, 0.05, 0.08, 0.1, 0.2,..., 1.0, 2.0,..., 10.0 were used.

[27]See Dufour (1989), Kodde and Palm (1986), Wolak (1989) and Barnett, Geweke and Yue (1991).

[28]See Diewert (1974) and Barnett and Yue (1988).

$$\eta_{ij} = \frac{\partial A_i p_j}{\partial p_j A_i} = s_j\sigma_{ij} - s_j\eta_{i0},$$

where A_i are the ordinary or uncompensated demand functions. The connection between compensated and uncompensated demand functions is stated by the usual Slutsky equation.

Gross substitutability and complementarity is provided by the off-diagonal terms of the uncompensated price elasticity matrix. If η_{ij} is positive, good i and good j are substitutes; in other words, when the price of good i rises, demand for good j increases to replace a cutback in demand for good i. If it is negative, they are complements—an increase in the price of good i (or j) causes the demand to fall for both goods.

Similarly, the pure substitution effects are defined by the Allen Partial matrix. If the utility function obeys regularity conditions, the own compensated price elasticities ($s_i\sigma_{ii}$) and (σ_{ii}), must be negative. Hence, the compensated price elasticity matrix represents potential movements along the consumer's indifference curves and can be used to examine whether the estimated underlying utility function satisfies regularity conditions.

The computed elasticities of the AIM demand system are compared with other money demand systems in the next section. Because of the complexity of share equations, a numerical method is used to compute the partial derivatives of the expenditure shares with respect to prices and income that occur in the elasticity formula. The computation of elasticities is calculated using the estimated share equations. Time series of the elasticities are produced by substituting time series of normalized prices and respective partial derivatives into the elasticity formula.

EMPIRICAL RESULTS OF THE AIM MONEY DEMAND SYSTEM

In this section, the AIM demand system is estimated and the income and substitution elasticities are compared with those for the translog and Fourier demand systems. In addition, characteristics of monetary assets relative to consumer goods are analyzed.

Table 1

Estimates of Coefficients of the AIM: 1970.1 to 1985.2

a_1	a_2	a_3	a_4	a_5	a_6
.0000	.0000	.0122	.9878	.2225	.2857
a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
.2299	.4466	.3568	.0515	.4446	.3961
a_{13}	a_{14}	a_{15}			
.4056	.4535	.4489			

Table 2

Income and Substitution Elasticities

		Allen Partial Matrix			
	Income	A_1	A_2	A_3	A_4
A_1	.502 (.005)	–16.564 (1.880)			
A_2	.512 (.008)	2.693 (.213)	–15.450 (2.649)		
A_3	.521 (.003)	5.553 (.721)	5.312 (.609)	–31.375 (6.944)	
A_4	1.045 (.007)	.364 (.009)	.378 (.008)	.170 (.009)	–.030 (.005)

NOTE: Standard deviations in parentheses.

Estimates of Parameters and Income and Price Elasticities

Table 1 displays the coefficient estimates from the AIM demand system derived by U.S. quarterly data from 1970.1 through 1985.2.[29] These parameters represent the consumer's taste or preference and determine the utility function that underlies the estimated AIM demand system. Because the taste parameters are assumed to be constant, the consumer's utility function and preference did not change over time. The estimates of a_1 and a_2 were zero due to the non-negativity constraint.[30]

The estimated Allen Partial elasticities of substitution and income elasticities are reported in table 2. The numbers represent the averages

[29]Douglas Fisher provided all of the data used to estimate the AIM demand system. He had previously used these data to estimate the translog and the Fourier demand systems. Hence, the empirical results presented here can be compared directly with his. See Fisher (1989).

[30]To this extent, the AIM model is at odds with the data because the estimated parameters would have been negative if unconstrained.

Table 3
Income and Substitution Elasticities

	Translog					Fourier				
	Income	A_1	A_2	A_3	A_4	Income	A_1	A_2	A_3	A_4
A_1	0.542	0.106				0.879	−3.851			
A_2	3.669	−0.490	−0.967			2.714	1.258	−8.104		
A_3	2.741	0.323	2.192	−0.262		2.803	1.119	2.876	−1.590	
A_4	0.737	−0.150	0.805	0.428	−0.672	0.898	0.171	0.412	−0.224	−0.042

and their standard deviations (in parentheses) over the sample period. Table 3 displays the estimated substitution and income elasticities from the translog and the Fourier series models previously reported by Fisher (1989, page 103).

Table 4 presents the average uncompensated price elasticities and their standard deviations over the sample period for AIM. The corresponding elasticities for the other two models are not available.

What's Wrong with the Translog and Fourier Demand Systems?

In the translog demand system results (shown in table 3), the positive sign of σ_{11} indicates that the regularity condition is violated. This result suggests that the higher the opportunity cost of holding currency and demand deposits, the greater their demand. Given this violation of the "law of demand," the results from the translog demand system must be considered suspicious at best and, at worst, unreliable.

Problems in the Fourier series demand system cannot be seen in table 3 because the numbers reported there are the average values of these coefficients. According to Fisher, however, except for η_{40}, the income elasticities and Allen Partial elasticities of substitution changed signs frequently over the period.[31] For example, in 1970, σ_{12} was significantly negative, implying complementarity; in 1971-1972, it was significantly positive, implying substitutability; and then in 1974-1975, it became negative again. Figure 1 displays a σ_{12}-comparison of the Fourier and AIM money demand systems. It is inexplicable that currency and demand deposits, A_1, and savings deposits and money market deposit accounts, A_2, should be complements during some periods and substitutes during others.

Table 4
Uncompensated Price Elasticities

	A_1	A_2	A_3	A_4
A_1	−.534 (.004)	.075 (.010)	.084 (.011)	−.126 (.009)
A_2	.069 (.005)	−.537 (.007)	.080 (.010)	−.123 (.007)
A_3	.157 (.005)	.162 (.009)	−.518 (.004)	−.322 (.004)
A_4	−.022 (.003)	−.023 (.005)	−.015 (.003)	−.985 (.003)

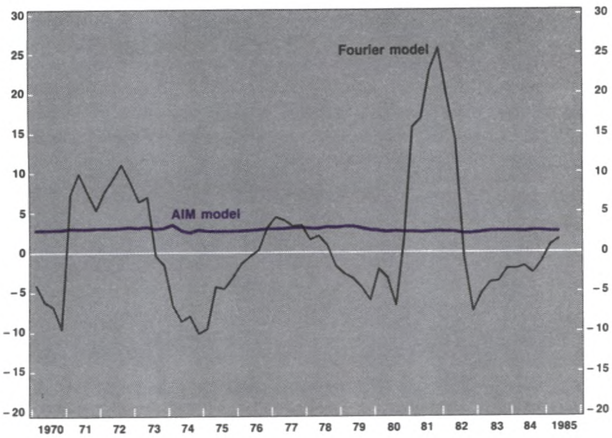
NOTE: Standard deviations in parentheses.

Empirical Inference of Characteristics of Monetary Assets by the AIM Demand System

The anomalies observed with the translog and the Fourier series demand systems do not occur in the AIM demand system. The own-price elasticities are negative and all estimated elasticities maintain their signs over the entire sample period. Moreover, their smaller standard deviations indicate that they are more stable; this can also be seen in figure 1. In the Allen Partial matrix, the diagonal elements are all negative

[31]See Fisher (1989), pp. 105-06.

Figure 1
The Allen Partial Elasticity of A_1 and A_2



while the off-diagonal elements are positive. This implies that the three monetary aggregates and aggregate consumption are substitutes for each other in the presence of income compensation. Moreover, the pure substitution effect between each pair of the three aggregated monetary assets is much greater than between the consumption good and monetary assets.

The income elasticities in table 2 are all positive, with the income elasticity of the consumption good about unity, and the income elasticities of the three monetary assets roughly equal to 0.5.[32] These results suggest that consumption goods and monetary assets are normal goods.

Table 4 shows that the uncompensated cross-price elasticities of (A_1, A_2), (A_1, A_3) and (A_2, A_3) are positive, implying that these monetary assets are gross substitutes. The uncompensated cross-price elasticities of (A_1, A_4), (A_2, A_4) and (A_3, A_4) are negative, indicating that consumption goods and monetary assets are gross complements.

These results show that, if the user costs of savings deposits (or money market deposit accounts, or small time deposits) rise, consumers will shift their funds to demand deposits (or to checkable deposits or currency). An opposite shift of funds will take place if the user costs of currency, demand deposits and checkable deposits should rise. If these user-cost changes are sufficiently large, ignoring the cross-price effects among monetary assets will produce large errors in their demand functions.

Monetary services and consumer goods are consumed jointly. If a consumer increases the consumption of commodities for some reason, the demand for monetary assets is also increased. This is consistent with the idea that consumers hold monetary assets to finance current and future consumption. Also, the negative η_{i4} indicates that price inflation will reduce demand for monetary assets.

Not surprisingly, the own price elasticities of monetary assets are greater than their cross price elasticities. Similarly, it is not surprising to see that the cross-price effects of a change in the price of consumption goods on the demand for monetary assets are greater than the cross-price effects of a change in the price of monetary assets on consumption. Consequently, empirical results from the AIM demand system provide a reasonable quantitative analysis of the characteristics of monetary assets—characteristics that are broadly consistent with conventional views of demand for money.

A DYNAMIC ANALYSIS BY THE AIM DEMAND SYSTEM

Constructing a dynamic model of demand for money has been difficult because the current state of economic knowledge about dynamic behavior is incomplete; the dynamics of money demand are still very much a "black box" mystery.[33]

Unlike most multivariate time series models, the AIM model is static. It does not consider specific dynamic effects among monetary assets and consumption goods. The utility function is not intertemporal and its parameters are time-invariant—the consumer's preference is not permitted to change over time.

[32] This may not be justified by the unity income elasticity in the reduced form of the aggregate money-demand function. As pointed out in the text, no reduced form demand functions are estimated in our study and the income elasticity is defined by the microeconomic approach.
Because their income elasticities are significantly less than one, the monetary assets in our study are not "luxury goods," as has been claimed in some previous research. See Serletis (1988).

[33] This situation occurs in a recent debate in economic literature; see Hendry and Ericsson (1991) and Friedman and Schwartz (1991).

Figure 2
Income Elasticities E_1, E_2, E_3 and E_4 for A_1, A_2, A_3 and A_4

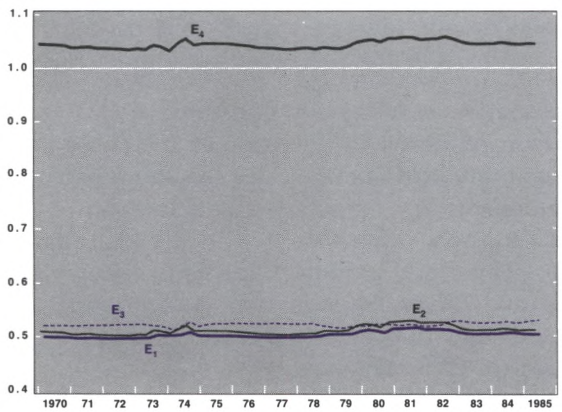


Figure 3
Uncompensated Own Price Elasticities

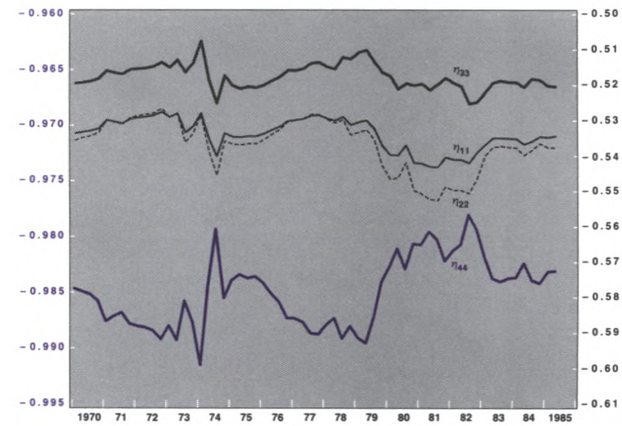
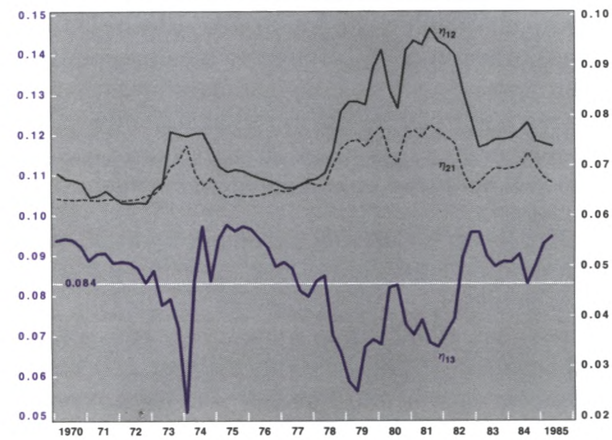


Figure 4
Uncompensated Cross Price Elasticities



Nevertheless, a simple dynamic analysis can be used to examine the AIM demand system. Time series of income and price elasticities can be computed using estimated share equations. Movements in these elasticities reflect both changes in user costs and the consumer's reactions to such changes. Both of these are reflected in the shares, s_i. In this way, the dynamics of the AIM demand system can be investigated even though demand for money is stable by assumption.

This dynamic analysis is displayed in figures 2, 3 and 4. Figure 2 shows that the income elasticities of the three monetary assets are relatively constant over the entire sample period. In contrast, the price elasticities (shown in figures 3 and 4), exhibit sizable fluctuations. Major shifts in price elasticities occurred during 1973.1-1974.4 and 1978.2-1982.2. During these periods, a number of studies have reported that demand for monetary aggregate M1 was "erratic."[34]

Price and user cost elasticities moved drastically during these periods. In the second period (1978.2-1982.2), the cross price elasticity, η_{12}, rose by 50 percent of its 1977 level, implying that demand for A_1 (currency plus demand deposits, plus checkable deposits) became much more sensitive than it was previously to changes in the opportunity cost of holding A_2 (savings deposits and money market demand accounts). Meanwhile, there was a sharp rise in the user cost of A_2. These factors would appear to account for the major shift of funds from A_2 to A_1 during the period.

The opposite price elasticity, η_{21}, also rose by 20 percent. Nevertheless, it was less than 80 percent of the value of η_{12} and the rise in the user cost of A_1 was more modest than that of A_2. It was observed that the opportunity cost of A_2 increased much faster than that of A_1. Hence, the actual flow of funds from A_1 to A_2 might not be significant.

The cross price elasticity, η_{13}, dropped 30 percent in 1979, implying that the demand for A_1

[34]See Goldfeld (1976) and Friedman (1984).

was less sensitive to changes in the user cost of A_3. Hence, the shift of funds from A_3 to A_1 should have been moderate despite a substantial increase in the user cost of A_3.

These results are roughly consistent with developments during the period. In November 1978, commercial banks were authorized to offer automatic transfer service (ATS) from savings accounts to checking accounts. Other interest ceiling-free accounts were also introduced in the early 1980s. In January 1981, NOW accounts were introduced nationwide. These financial innovations should have encouraged consumers to shift funds from savings accounts and money market deposit accounts in A_2 into NOW accounts in A_1. This may have increased the interest sensitivity of demand for A_1.[35]

Simulating the Growth Rates of Monetary Aggregates

A further investigation of the behavior of monetary aggregates can be made by a dynamic simulation. For example, suppose that demand for A_i has been derived by utility maximization and expressed by the ordinary demand functions of price and user costs and total expenditure

(6) $A_i = G_i(u_1, u_2, u_3, u_4, E)$.

The total differentiation of (6) results in

(7) $dA_i = \sum_{j=1}^{4} dG_i/du_j \; du_j + dG_i/dE \; dE$.

Dividing both sides of (7) by (6) and using definitions of the uncompensated price elasticities and the income elasticity gives

(8) $dA_i/A_i = \sum_{j=1}^{4} \eta_{ij} (du_j/u_j) + \eta_{i0} (dE/E)$.

Using time series of the elasticities and the growth rates of price and user costs and total expenditure, the right-hand sides of the equations in (8) are computed. In this way, the growth rates of demand for A_i can be "simulated."

The actual and simulated growth rates of demand for monetary aggregates and consumption are displayed in figures 5 through 8. The simulations match the actual growth rates fairly well, especially for consumption. The simulation rates of monetary assets had large fluctuations

Figure 5
Growth Rates of Aggregate A_1 Actual vs. Simulation

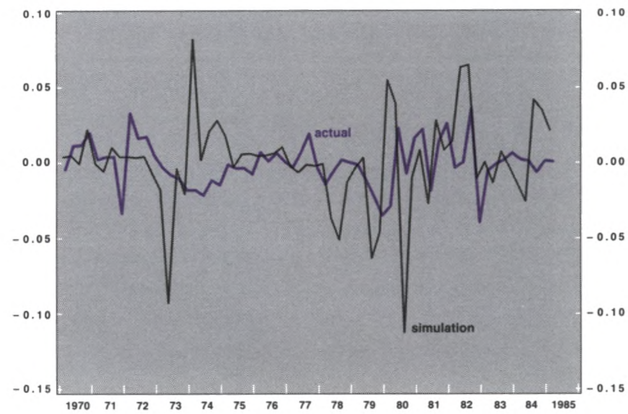


Figure 6
Growth Rates of Aggregate A_2 Actual vs. Simulation

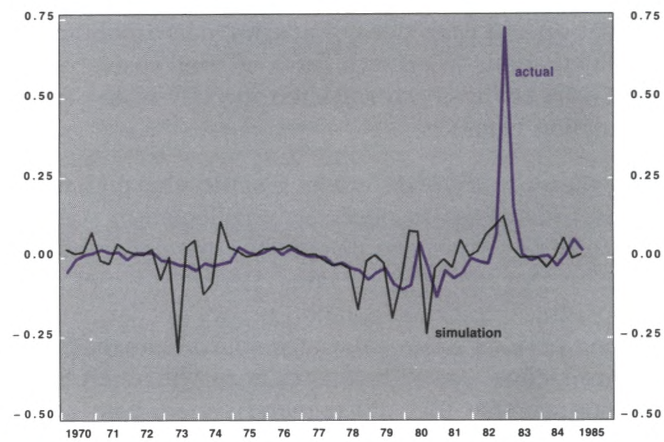
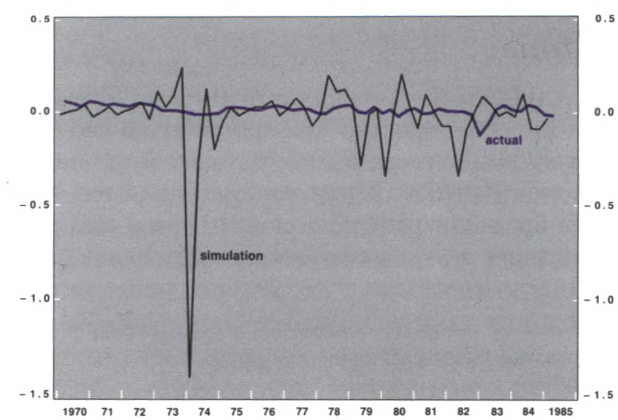
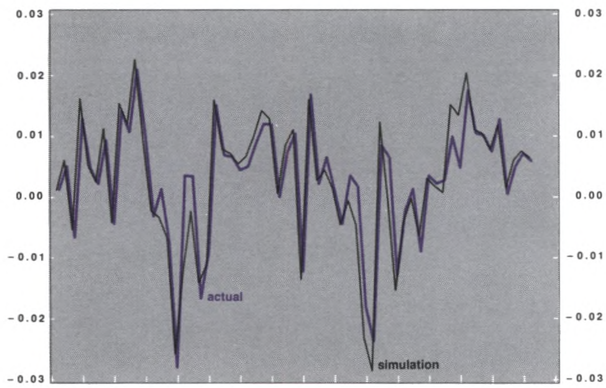


Figure 7
Growth Rates of Aggregate A_3 Actual vs. Simulation



[35]See Thornton and Stone (1991) for a discussion of this possibility.

Figure 8
Growth Rates of Aggregate A_4 Actual vs. Simulation



around the actual growth rates in the periods 1972.4-1975.1 and 1978.2-1982.2. Fluctuations in interest rates and inflation rates were substantial during each period, causing corresponding fluctuations in growth rates of user costs.[36] These changes are reflected directly in the simulation rates.

Because the AIM model is static, sharp changes in user costs are necessarily reflected in correspondingly sharp changes in the simulated growth rates of aggregates. Hence, it is not surprising that the simulation errors are large during periods when there are sharp changes in user costs. Nevertheless, figures 5 through 8 suggest that the AIM demand system has captured many of the characteristics of the U.S. monetary system during the sample period.

Can the AIM Demand System Explain the Case of the Missing Money?

Although the simulation of the growth rate of A_1 indicates that the AIM model produced relatively large errors during the period of "missing money" (1973.4-1976.2), an analysis of the AIM results might provide a clue.[37] During this period, there was a sharp decline in demand for M1; conventional money demand equations consistently overpredicted demand.

A potential explanation can be obtained by considering figures 9 and 10. Figure 9 shows the partial derivative of demand for A_1 with respect to its own user cost. Figure 10 shows the partial derivative of demand for A_1 with respect to the user costs of A_2, A_3, A_4 and E. Figure 9 shows a sharp rise in the rate of change in demand for A_1 with respect to a change in its user cost. Indeed, figure 11 shows that the user cost of A_1 increased relative to the price level. Hence, the demand for A_1 should have declined by a proportionately larger amount than the rise in its user cost. Conventional linear money demand equations with fixed regression coefficients could not accommodate this nonlinearity because, in such linear demand systems, the coefficients (derivatives) are assumed to be constant. However, figures 9 and 10 clearly suggest that this is not the case. Moreover, ordinary least squares is relatively sensitive to "outliers." Consequently, there will be substantial changes in the estimated regression coefficients when the equations are estimated over periods when these derivatives change significantly.

Conventional money demand equations may be misspecified for another reason, as well. Usually they include a single short-term interest rate intended to reflect the opportunity cost of holding money. AIM analysis indicates that the demand for money does not depend on a single "representative" interest rate, but on its user cost and the user costs of "close" substitutes (recall that the demand for A_1 was sensitive to changes in the user costs of A_2). Hence, conventional money demand equations may produce misleading results when interest rates change relative to the user costs of M1 or relative to the user costs of close substitutes for M1.

Therefore, the case of "missing money" and "unexplained" parameter shifts in conventional money demand functions may result from the fact that they are essentially linear approxima-

[36]Some terms are essentially zero and can be ignored. The following growth rate equations are accurate enough to produce the simulation:

$$dA_1/A_1 = \eta_{11}du_1/u_1 + \eta_{12}du_2/u_2 + \eta_{13}du_3/u_3 + \eta_{44}du_4/u_4 + \eta_{10}dE/E$$
$$dA_2/A_2 = \eta_{22}du_2/u_2 + \eta_{23}du_3/u_3$$
$$dA_3/A_3 = \eta_{33}du_3/u_3$$
$$dA_4/A_4 = \eta_{42}du_2/u_2 + \eta_{43}du_3/u_3 + \eta_{44}du_4/u_4 + \eta_{10}dE/E.$$

In the equation for A_1 there are more affecting elements; the own and cross-price effects of A_2, A_3 and A_4 are important in simulating the growth rate of A_1. The growth rates of demand for the other two monetary aggregates, however, are determined mainly by their own price effects and cross-price effect, η_{23}. This suggests that ignoring the substitution effects of non-M1 components of M2 might be one of the factors that discredit reliability of the conventional M1 demand function.

[37]See Goldfeld (1976).

Figure 9
The First Derivative of A_1 with Respect to Own User Cost

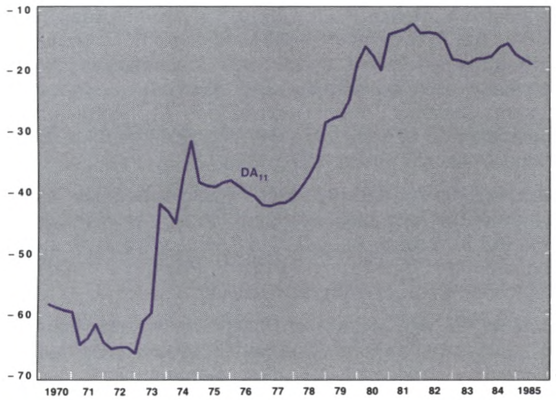


Figure 11
User Cost of A_1 vs. Price of A_4

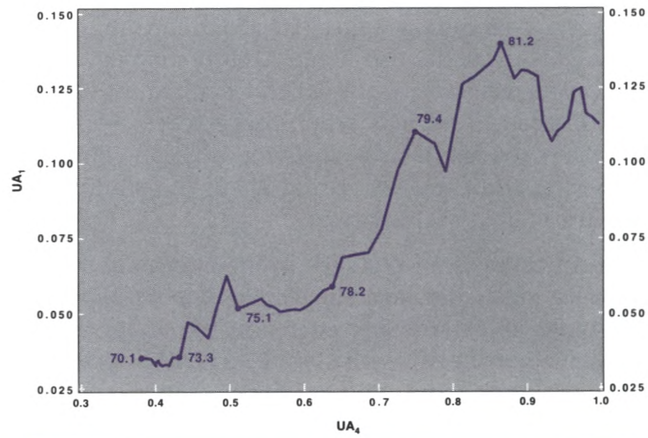
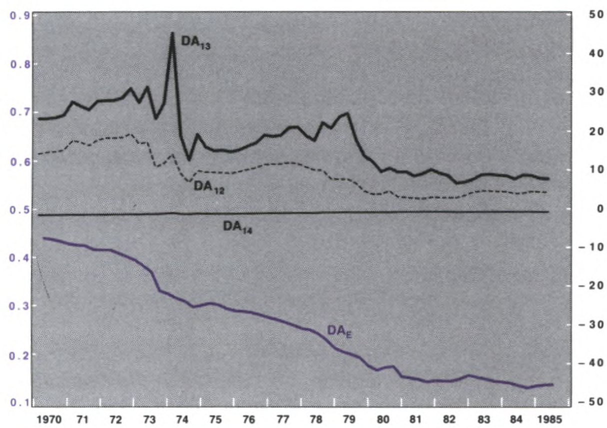


Figure 10
The First Derivatives of A_1 with Respect to User Costs, Price and Income



tions to nonlinear demand functions. If so, they intuitively will provide much poorer approximations during periods when there are dramatic changes in user costs.

Is The Money Demand Function Stable?

The erratic behavior of conventional money demand functions and, more recently, the income velocity of M1, have led many researchers to assert that the demand for money is "unstable."[38] Others have asserted that money demand is stable based on the observed stability of the consumption function.[39]

The AIM demand system integrates demand for both consumption and money and then estimates them simultaneously. These estimates suggest that, while the own price and cross price elasticities show considerable variation due to changes in the price level and user costs, they change little on average over the period (see figures 3 and 4).[40] Moreover, the estimated income elasticities for all three monetary aggregates are nearly constant (see figure 2). Of course, these results are obtained from a model where the estimated parameters are time-invariant, that is, the preference function is constant. Because of this, it is necessarily true that demand functions are "stable." Nevertheless, the relatively good performance of AIM provides some promise that, like consumption, the demand for money will ultimately be shown to be a stable function of a relatively few economic variables—in this case, income and user costs.

CONCLUSION

Two distinctly different micro-econometric demand system approaches to the demand for money were presented and discussed. An advanced AIM demand system was presented and estimated using U.S. time-series data. Unlike

[38]For a discussion of the velocity of M1 and an analysis of some of the explanations, see Stone and Thornton (1987).

[39]For example, see Friedman (1956) and Lucas (1988).

[40]In the parlance of modern time-series analysis, these elasticities are said to be stationary, that is, mean reverting. However, no formal tests of stationarity were performed in this study.

other utility function-based approaches, AIM estimates are consistent with microeconomic theory. Dynamic simulations of the growth rates of various monetary aggregates and consumption suggest that the estimated AIM model performed well; nevertheless, the largest simulation errors occurred in periods when there were relatively sharp swings in user costs or inflation. This is perhaps not too surprising given the static nature of the AIM analysis.

An analysis of changes in income and cross price elasticities are suggestive of portfolio shifts among monetary aggregates in the 1970s and 1980s consistent with the observed behavior of these aggregates. The results of AIM suggest that the reported failure of conventional linear (or log-linear) money demand equations may result from trying to fit fundamentally non-linear functions with linear ones. The results shown here suggest that this problem will be particularly acute whenever there are sharp changes in user costs. Unfortunately, these are precisely the times when AIM performance was also poor. The key to solving this problem in AIM, however, is to find a way to make AIM explicitly dynamic. It may not be necessary to assume that consumer preferences are unstable.

The sampling distribution theory for AIM has not been worked out at this time, so relevant hypothesis tests cannot be conducted yet. Also, because the time series on the relevant user costs of monetary aggregates is limited, the available data cover a relatively short sample period. These factors, coupled with the fact that even low-order (first-order) AIM systems require a relatively large number of estimated parameters, place severe limits on attempts to evaluate the performance of AIM using out-of-sample forecasts. Despite these problems, the estimated AIM system appears to have captured many of the characteristics of monetary assets and offers some useful explanations to puzzling empirical issues. Hence, these results are encouraging to those who believe that microeconomic principles, such as utility maximization, can be applied usefully to macroeconomic problems.

REFERENCES

Barnett, William A. *Consumer Demand and Labor Supply* (Amsterdam: North-Holland, 1981).

________. "New Indices of Money Supply and the Flexible Laurent Demand System," *Journal of Business and Economics Statistics* (January 1983), pp. 7-23.

Barnett, William A., and Andrew B. Jonas. "The Muntz-Szatz Demand System: An Application of a Globally Well Behaved Series Expansion," *Economics Letters* Vol. 11, No.4 (1983), pp. 337-42.

Barnett, William A., John Geweke, and Piyu Yue. "Seminonparametric Bayesian Estimation of the Asymptotically Ideal Model: the AIM Consumer Demand System," in William Barnett, James Powell, and George Tauchen, eds., *Nonparametric and Seminonparametric Methods in Econometrics and Statistics*, Proceedings of the Fifth International Symposium in Economic Theory and Econometrics (Cambridge University Press, 1991).

Barnett, William A., and Piyu Yue. "Seminonparametric Estimation of the Asymptotically Ideal Model: the AIM Demand System," in George Rhodes and Thomas Fomby, eds., *Nonparametric and Robust Inference*, Advances in Econometrics, Volume 7 (1988), pp. 229-51.

Diewert, W. E. "Applications of Duality Theory," in Michael D. Intriligator and David A. Kendrick, eds., *Frontiers of Quantitative Economics,* Vol. II (Amsterdam: North-Holland, 1974), pp. 106-71.

Diewert, W. E., and T. J. Wales. "Flexible Functional Forms and Global Curvature Conditions," *Econometrica* (January 1987), pp. 43-68.

Dufour, Jean-Marie. "Nonlinear Hypotheses, Inequality Restrictions, and Non-Nested Hypotheses: Exact Simultaneous Tests in Linear Regressions," *Econometrica* (March 1989) pp. 335-56.

Fayyad, Salam K. "A Microeconomic System-Wide Approach to the Estimation of the Demand for Money," this *Review* (August/September 1986), pp. 22-33.

Feenstra, Robert C. "Functional Equivalence Between Liquidity Costs and the Utility of Money," *Journal of Monetary Economics* (March 1986), pp. 271-91.

Fisher, Douglas. *Money Demand and Monetary Policy* (The University of Michigan Press, 1989).

Friedman, Benjamin. "Lessons From the 1979-82 Monetary Policy Experiment," *American Economic Review* (May 1984), pp. 382-87.

Friedman, Milton. "The Quantity Theory of Money—A Restatement," in Milton Friedman, ed., *Studies in the Quantity Theory of Money* (University of Chicago Press, 1956).

Friedman, Milton, and Anna J. Schwartz. "Alternative Approaches to Analyzing Economic Data," *The American Economic Review* (March 1991), pp. 39-49.

Gallant, A. Ronald. "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form," *Journal of Econometrics* (February 1981), pp. 211-45.

________. "The Fourier Flexible Form," *American Journal of Agricultural Economics* (May 1984), pp. 204-08.

Gallant, A. Ronald, and Douglas W. Nychka. "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica* (March 1987), pp. 363-90.

Goldfeld, Stephen M. "The Case of the Missing Money," *Brookings Papers on Economic Activity* (3:1976), pp. 683-739.

Hendry, David F., and Neil R. Ericsson. "An Econometric Analysis of U.K. Money Demand in Monetary Trends in the United States and the United Kingdom by Milton Friedman and Anna J. Schwartz," *American Economic Review* (March 1991), pp. 8-38.

Judd, John P., and John L. Scadding. "The Search for a Stable Money Demand Function: A Survey of the Post-1973 Literature," *Journal of Economic Literature* (September 1982), pp. 993-1023.

Kodde, David A., and Franz C. Palm. "Wald Criteria for Jointly Testing Equality and Inequality Restrictions," *Econometrica* (September 1986), pp. 1243-48.

Lasdon, Leon S., and A. D. Warren. "GRG2 User's Guide" University of Texas at Austin, 1989.

Lucas, Robert E., Jr. "Money Demand in the United States: A Quantitative Review," *Carnegie-Rochester Conference Series on Public Policy* No. 29, (Autumn 1988), pp. 137-68.

Mankiw, N. Gregory. "A Quick Refresher Course in Macroeconomics," *Journal of Economic Literature* (December 1990), pp. 1645-60.

Moore, George R., Richard D. Porter, and David H. Small. "Modeling the Disaggregated Demands for M2 and M1: The U.S. Experience in the 1980s," in Peter Hooper and others, eds., *Financial Sectors in Open Economies: Empirical Analysis and Policy Issues* (Board of Governors of the Federal Reserve System, 1990), pp. 21-97.

Pierce, James L. "Did Financial Innovation Hurt the Great Monetarist Experiment?" *American Economic Review* (May 1984), pp. 392-96.

Rasche, Robert H. "Demand Functions for Measures of U.S. Money and Debt," in Peter Hooper and others, eds., *Financial Sectors in Open Economies: Empirical Analysis and Policy Issues* (Board of Governors of the Federal Reserve System, 1990), pp. 113-61.

Serletis, Apostolos. "Translog Flexible Functional Forms and Substitutability of Monetary Assets," *Journal of Business and Economic Statistics* (January 1988), pp. 59-67.

Stone, Courtenay C., and Daniel L. Thornton. "Solving the 1980s Velocity Puzzle: A Progress Report," this *Review* (August/September 1987), pp. 5-23.

Swofford, James L., and Gerald A. Whitney. "Flexible Functional Forms and the Utility Approach to the Demand for Money: a Nonparametric Analysis," *Journal of Money, Credit and Banking* (August 1986), pp. 383-89.

Telser, Lester G., and Robert L. Graves. *Functional Analysis in Mathematical Economics* (The University of Chicago Press, 1972).

Thornton, Daniel L., and Courtenay C. Stone. "Financial Innovation: Causes and Consequences," in Kevin Dowd and Mervyn K. Lewis, eds., *Current Issues in Monetary Analysis and Policy* (MacMillan Publishers, 1991).

Uzawa, Hirofumi. "Production Functions with Constant Elasticities of Substitution," *Review of Economic Studies* (October 1962), pp. 291-99.

Wolak, Frank A. "Local and Global Testing of Linear and Nonlinear Inequality Constraints in Nonlinear Econometric Models," *Econometric Theory* (April 1989), pp. 1-35.

Mark D. Flood

Mark D. Flood is an economist at the Federal Reserve Bank of St. Louis. David H. Kelly provided research assistance.

Microstructure Theory and the Foreign Exchange Market

A GROWING BODY OF theoretical literature, known as the study of securities market microstructure, deals with the behavior of participants in securities markets and with the effects of information and institutional rules on the economic performance of those markets. These institutional factors may arise from technology, tradition or regulation. Microstructure and its impact are important, because of the vast amounts of wealth which pass through securities markets — including the foreign exchange market — every day.

Microstructure is of interest to students of the foreign exchange market: microstructural analyses of other markets have yielded insight into traders' behavior and the effect of various institutional arrangements. Conversely, the foreign exchange market is also of special interest to students of microstructure, because it combines two very different arrangements for matching buyers and sellers — bank dealers trade with one another both directly and through foreign exchange brokers.[1]

Standard models of exchange-rate determination concentrate on relatively long-run aspects, such as purchasing power parity. While microstructure theory cannot address these issues directly, it can illuminate a more narrowly focused array of institutional concerns, such as price information, the matching of buyers and sellers, and optimal dealer pricing policies. Despite the substantial literature on microstructure, little attention has been paid to the particular microstructure of the foreign exchange market.[2]

[1]Similar arrangements exist for other securities—for example, the federal funds market and the secondary market for Treasury securities—but these too have been relatively neglected in the literature.

[2]The shaded insert on the opposite page provides a context in which the microstructural approach can be compared with more traditional approaches to market efficiency.

Following some early articles by Demsetz (1968), Tinic (1972) and Tinic and West (1972), Garman (1976) performed the crucial task of defining market microstructure as an independent area of the literature, thus focusing the debate. Since then, market microstructure has burgeoned, led by Cohen, Maier, Schwartz and Whitcomb (1978a, 1978b, 1981, 1983), Amihud and Mendelson (1980, 1986, 1988), Stoll (1978, 1985, 1989) and Ho and Stoll (1980, 1981). See also Beja and Hakansson (1977), Cohen, Hawawini, Maier, Schwartz and Whitcomb (1980), Cohen, Maier, Ness, Okuda, Schwartz and Whitcomb (1977), Amihud, Ho, and Schwartz (1985), Schreiber and Schwartz (1986), Schwartz (1988) and Cohen and Schwartz (1989). Cohen, Maier, Schwartz and Whitcomb (1979, 1986) and Stoll (1985) have surveyed the microstructure literature.

In addition to the early note by Allen (1977), very recently there have appeared some microstructural studies of the foreign exchange market: Bossaerts and Hillion (1991), Lyons (1991), Rai (1991) and Flood (1991). There is also an empirical literature measuring the determinants of the bid-ask spread in the foreign exchange market. See Black (1989), Wei (1991) and Glassman (1987) as well as the references therein. Because the focus of this article is on microstructure theory, such empirical studies receive little attention here.

Finally, although a consideration of the results of laboratory experiments would expand the scope of this paper to unwieldy dimensions, their role in establishing the sensitivity of market behavior to institutional factors must at least be acknowledged; see Plott (1982, 1991) for an introduction.

Price Efficiency in a Heterogeneous Marketplace

Implicit in most microstructural models is a presumption that participants in any given market are heterogeneous, that is, that they differ in certain key determinants of economic behavior: information, beliefs, preferences and wealth. Although this assumption consumes little attention in the microstructure literature — it is taken for granted — it is valuable to discuss it in the more familiar theoretical context of market efficiency.

The standard definition of price efficiency is: $f_m(p_t|I_{mt-1}) = f^*(p_t|I_{t-1})$. In other words, the joint distribution over future prices, $f_m(p_t)$, assessed by the monolithic market (or a representative agent in that market) and made conditional on the current information, I_{mt-1}, available to the market is equal to the "true" joint distribution, $f^*(p_t)$, made conditional on all current information, I_{t-1}. Roughly speaking, the market sorts things out as accurately as possible.[1]

This approach breaks down in a microstructural analysis. First, the simplifying assumption of homogeneous participants is abandoned. Although it is widely recognized "that investors do not show the homogeneity of beliefs which characterize our theories," the benefits of realism (i.e., the heterogeneity assumption) are often outweighed by other criteria (e.g., testability, tractability, etc.).[2] Emphasizing testability, Ross offers a standard rejoinder, namely that "since a single ex post distribution of returns is observed by all, over time one would not expect to observe systematic and persistent differences." This is a rational expectations argument, which depends crucially on the stationarity of the returns distribution and which ignores the effect of differences in opinions and beliefs, which go beyond differences in information. In general then, at the level of detail involved in microstructural studies, the homogeneity assumption is not an excusable flaw; in a homogeneous market why — let alone how — would anyone trade?

More fundamentally, the notion of a "true" price must be questioned. In the context of the literature on price efficiency, the introduction of a "true" distribution as a theoretical conceit leads to joint testing problems, as the "true" distribution is *ipso facto* unobservable. More fundamentally, positing a "true" distribution confuses the chain of causality; it presumes that future prices are drawn from some exogenous probability distribution and that investor behavior is concerned with accurately estimating that distribution.

In fact, investor behavior in the marketplace determines the distribution of future prices, not the other way around. This fact in no way depends on the ultimate basis or motivation for investor behavior. In an explicit model of price discovery, the assertion of an *ex ante* exogenous equilibrium price is meaningless. As Schreiber and Schwartz put it, "the fact that security analysts assess the value of a stock for their own portfolios does not imply that they undertake a treasure hunt to find some golden number which one might call an intrinsic value."[3] In sum, the standard theory of efficient markets is ill-suited to the modeling of price discovery. In comparing observed prices to an imputed "true" distribution, studies of market efficiency ignore more immediate concerns — for example, how well the institutional structure transmits information, whether arbitrage opportunities occur, and how well the market allocates assets among investors. These concerns are the focus of microstructural analysis.

[1]See Fama (1976), chapter 5, for the definitive presentation.

[2]See Ross (1978), pp. 889-90. See Varian (1989) for a more thorough review of the theoretical issues involved in the heterogeneity assumption.

[3]See Schreiber and Schwartz (1985), p. 22.

This paper examines the extant literature on market microstructure to determine how it might be applied to the foreign exchange market.

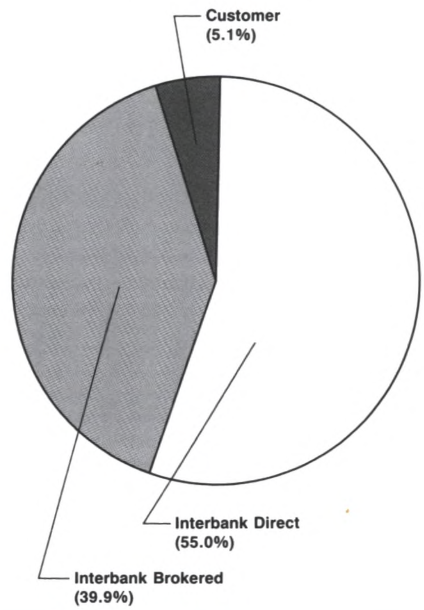
The paper begins with a brief description of the foreign exchange market. Aspects of the literature concerned with institutional details are addressed second, noting how such details can affect the performance of the market. Next, the literature dealing with behavioral details, especially the communication and interpretation of price information, is considered. Finally, the interaction of institutional and behavioral factors, notably the bid-ask spread, is discussed.

INSTITUTIONAL BASICS OF THE FOREIGN EXCHANGE MARKET

The foreign exchange market is the international market in which buyers and sellers of currencies "meet."[3] It is largely decentralized: the participants (classified as market-makers, brokers and customers) are physically separated from one another; they communicate via telephone, telex and computer network. Trading volume is large, estimated at $128.9 billion for the U.S. market in April 1989. Most of this trading was between bank market-makers.[4]

The market is dominated by the market-makers at commercial and investment banks, who trade currencies with each other both directly and through foreign exchange brokers (see figure 1).[5] Market-makers, as the name suggests, "make a market" in one or more currencies by providing bid and ask prices upon demand. A broker arranges trades by keeping a "book" of market-maker's limit orders — that is, orders to buy (alternatively, to sell) a specified quantity of foreign currency at a specified price — from which he quotes the best bid and ask orders upon request. The best bid and ask quotes on a broker's book are together called the broker's "inside spread." The other participants in the market are the customers of the market-making banks, who generally use the market to complete transactions in international trade, and central banks, who may enter the market to move exchange rates or simply to complete their own international transactions. Market-makers may trade for their own account — that is, they may maintain a long or short position in a foreign currency — and require significant capitalization for that purpose. Brokers do not contact customers and do not deal on their own account; instead, they profit by charging a fee for the service of bringing market-makers together.

Figure 1
Spot Market Volume by Transactor (4/89)



The mechanics of trading differ substantially between brokered transactions and direct deals. In the direct market, banks contact each other. The bank receiving a call acts as a market-maker for the currency in question, providing a two-way quote (bid and ask) for the bank placing the call. A direct deal might go as follows:

Mongobank: *"Mongobank with a dollar-mark please?"*

(Mongobank requests a spot market quote for U.S. dollars (USD) against German marks (DEM).)

[3]For more thorough descriptions of the workings of the foreign exchange market, see Burnham (1991), Chrystal (1984), Kubarych (1983) and Riehl and Rodriguez (1983).

[4]See Federal Reserve Bank of New York (1989a) and Bank for International Settlements (BIS) (1990). Extending this figure over 251 trading days per year, this implies a trading volume of roughly $32 trillion for all of 1989. Volume has roughly doubled every three years for the past decade.

[5]Federal Reserve Bank of New York (1989a) lists 162 market-making institutions (148 are commercial banks) and 14 brokers; an earlier study, Federal Reserve Bank of New York (1980), lists 90 market-making banks and 11 brokers.

Loans 'n Things: *"20-30"*
(Loans n' Things will buy dollars at 2.1020 DEM/USD and sell dollars at 2.1030 DEM/USD —the 2.10 part of the quote is understood.)

Mongobank: *"Two mine."*
(Mongobank buys $2,000,000 for DEM 4,206,000 at 2.1030 DEM/USD, for payment two business days later. The quantity traded is usually one of a handful of "customary amounts.")

Loans 'n Things: *"My marks to Loans 'n Things Frankfurt."*
(Loans n' Things requests that payment of marks be made to their account at their Frankfurt branch. Payment will likely be made via SWIFT.)[6]

Mongobank: *"My dollars to Mongobank New York."*
(Mongobank requests that payment of dollars be made to them in New York. Payment will most likely be made via CHIPS.)[7]

Spot transactions are made for "value date" (payment date) two business days later to allow settlement arrangements to be made with correspondents or branches in other time zones. This period is extended when a holiday intervenes in one of the countries involved. Payment occurs in a currency's home country.

The other method of interbank trading is brokered transactions. Brokers collect limit orders from bank market-makers. A limit order is an offer to buy (alternatively to sell) a specified quantity at a specified price. Limit orders remain with the broker until withdrawn by the market-maker.

The advantages of brokered trading include the rapid dissemination of orders to other market-makers, anonymity in quoting, and the freedom not to quote to other market-makers on a reciprocal basis, which can be required in the direct market. Anonymity allows the quoting bank to conceal its identity and thus its intentions; it also requires that the broker know who is an acceptable counterparty for whom. Limit orders are also provided in part as a courtesy to the brokers as part of an ongoing business relationship that makes the market more liquid. Because his limit order is often a market-maker's first indication of general price shift, Brooks likens the posting of an order with a broker "to sticking out the chin so as to be acquainted with the moment that the fight starts."[8] Schwartz points out that posting a limit order extends a free option to other traders.[9]

A market-maker who calls a broker for a quote gets the broker's inside spread, along with the quantities of the limit orders. A typical call to a broker might proceed as follows:

Mongobank: *"What is sterling, please?"*
(Mongobank requests the spot quote for U.S. dollars against British pounds (GBP).)

Fonmeister: *"I deal 40-42, one by two."*
(Fonmeister Brokerage has quotes to buy £1,000,000 at 1.7440 USD/GBP, and to sell £2,000,000 at 1.7442 USD/GBP)

Mongobank: *"I sell one at 40, to whom?"*
(Mongobank hits the bid for the quantity stated. Mongobank could have requested a different amount, which would have required additional confirmation from the bidding bank.)

Fonmeister: [A pause while the deal is reported to and confirmed by Loans 'n Things] *"Loans 'n Things London."*
(Fonmeister confirms the deal and reports the counterparty to Mongobank. Payment arrangements will be made and confirmed separately by the respective back offices. The broker's back office will also confirm the trade with the banks.)

Value dates and payment arrangements are the same as in the direct dealing case. In addition to the payment to the counterparty bank, the banks involved share the brokerage fee. These fees are negotiable in the United States. They are also quite low: roughly $20 per million dollars transacted.[10]

[6]The Society for Worldwide Interbank Financial Telecommunication (SWIFT) is an electronic message network. In this case, it conveys a standardized payment order to a German branch or correspondent bank, which, in turn, effects the payment as a local interbank transfer in Frankfurt.

[7]The Clearing House for Interbank Payments System (CHIPS) is a private interbank payments system in New York City.

[8]See Brooks (1985), p. 25.

[9]See Schwartz (1988), p. 239.

[10]See Burnham (1991), p. 141, note 16, and Kubarych (1983), p. 14.

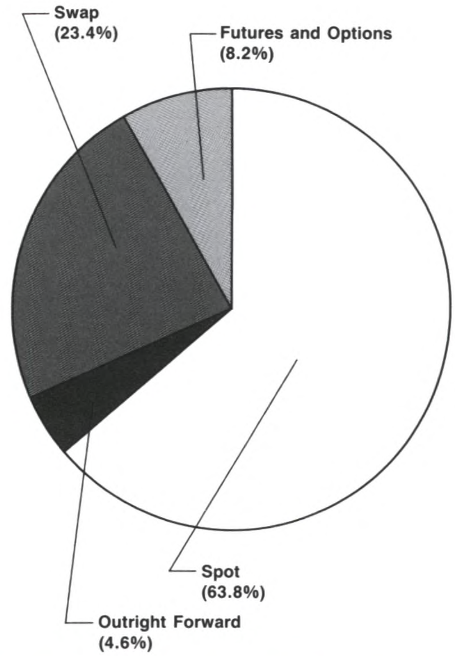
The final category of participants in the foreign exchange market is the corporate customers of the market-making banks. Customers deal only with the market-makers. They never go through brokers, who cannot adequately monitor their creditworthiness. Typically, a customer transacts with a bank with which it already has a well-established relationship, so that corporate creditworthiness is not a concern for the bank's foreign exchange desk, and trustworthiness is not an issue for the customer. The mechanics of customer trading are similar to those of direct dealing between market-makers. A customer requests a quote, and the bank makes a two-way market; the customer then decides to buy, sell or pass. The chief difference between this and an interbank relationship is that the customer is not expected ever to reciprocate by making a market.

Participants in the foreign exchange market also deal for future value dates. Such dealing composes the forward markets. Active forward markets exist for a few heavily traded currencies and for several time intervals corresponding to actively dealt maturities in the money market. Markets can also be requested and made for other maturities, however. Since the foreign exchange market is unregulated, standard contract specifications are matters of tradition and convenience, and they can be modified by the transacting agents.

Forward transactions generally occur in two different ways: outright and swap. An outright forward transaction is what the name implies, a contract for an exchange of currencies at some future value date. "Outrights" generally occur only between market-making banks and their commercial clients. The interbank market for outrights is very small, because outright trading implies an exchange rate risk until maturity of the contract. When outrights are concluded for a commercial client, they are usually hedged immediately by swapping the forward position to spot. This removes the exchange rate risk and leaves only interest rate risk.

A swap is simply a combination of two simultaneous trades: an outright forward contract and an opposing spot deal. For example, a bank might "swap in" six-month yen by simultaneously buying spot yen and selling six-month forward

Figure 2
Market-Maker Volume by Type (4/89)

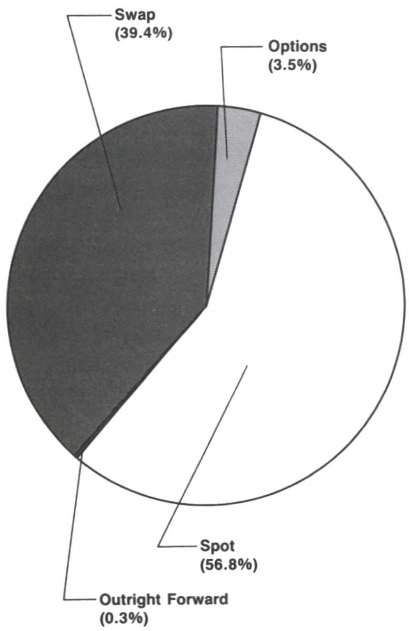


yen. Such a swap might be used to hedge an outright purchase of six-month yen from a bank customer.[11] In effect, the swapping bank is borrowing yen for the six months of the outright deal. The foreign exchange market-maker swaps in yen — rather than simply borrow yen on a time deposit — because banks maintain separate foreign exchange and money market accounts for administrative reasons. Swapping is generally the preferred means of forward dealing (see figures 2 and 3).

In practice, the vast majority of foreign exchange transactions involve the U.S. dollar and some other currency. The magnitude of U.S. foreign trade and investment flows implies that, for almost any other currency, the bilateral dollar exchange markets will have the largest volume. Consequently, the dollar markets are the most liquid. The possibility of triangular arbitrage enforces the law of one price for the cross rates. The upshot is that liquidity considerations outweigh transaction costs. A German wanting

[11]Hedging an outright purchase of currency with an opposing swap deal still leaves an open spot purchase of the currency. This can be easily covered in the spot market.

Figure 3
Broker Volume by Type (4/89)



pounds, for example, will typically convert marks to dollars and then dollars to pounds, rather than trading marks for pounds directly. Though this is especially true in the American market, it holds for foreign markets as well.

CLASSIFYING MARKETS

The microstructure literature is by nature market-specific, and much of it concerns U. S. equity markets. This specificity has the advantage of realism, but it makes the immediate applicability of some microstructural models to the foreign exchange market questionable. The first task is to define some basic microstructural concepts, identifying where the foreign exchange market fits into the context they provide. Such a taxonomy is important, because one of the fundamental lessons of the microstructure literature is that institutional differences can affect the efficiency of pricing and allocation.

As described above, the foreign exchange market combines two disparate auction structures for the same commodity: the interbank direct market and the brokered market. Defying a naive application of institutional Darwinism, whereby only the fitter of the two systems would survive, these trading methods appear to coexist comfortably.[12] The direct market can be classified as a decentralized, continuous, open-bid, double-auction market. The brokered market is a quasi-centralized, continuous, limit-book, single-auction market. The meanings of these classifications are explained below.

Centralization

In a centralized market, "trades are carried out at publicly announced prices and all traders have access to the same trading opportunities." In a decentralized market, in contrast, "prices are quoted and transactions are concluded in private meetings among agents."[13] A New York Stock Exchange's (NYSE) specialist system is a centralized market; the interbank direct market for foreign exchange is a decentralized one.

The distinction between centralized and decentralized markets might seem to provide a neat dichotomy of possible market structures. The multiplicity of brokers in the foreign exchange market violates this simple taxonomy, however. Each foreign exchange broker accumulates a subset of market-makers' limit orders. This network of "brokerage nodes" is as different from a fully centralized system as it is from a fully decentralized one. This arrangement is labeled here as "quasi-centralized."

Most microstructural studies have confined themselves to centralized markets, especially the NYSE's specialist system and the National Association of Securities Dealers Automated Quotation (NASDAQ) System on the over-the-counter (OTC) market.[14] Although there are a number of important decentralized markets, including the interbank direct foreign exchange market, rela-

[12]A similar situation obtains on the New York Stock Exchange, where specialists act as either brokers or market-makers, depending on the level of activity in the market.

[13]See Wolinsky (1990), p. 1. He goes on to analyze theoretically the difference in the price discovery process between centralized and decentralized markets. Schwartz (1988), pp. 426-35, refers to centralization as "spatial consolidation."

[14]For models of specialist systems, see Demsetz (1968), Tinic (1972), Garman (1976), Bradfield (1979), Amihud and Mendelson (1980), Conroy and Winkler (1981), Glosten and Milgrom (1985) and Sirri (1989). For studies of the OTC market, see Tinic and West (1972), Benston and Hagerman (1974), Ho and Macris (1985) and Stoll (1989).

tively few studies have focused on the impact of decentralization.

There is some evidence that differences in the degree of centralization between various markets cause differences in market performance. Garbade, in studying the largely decentralized Treasury securities market, concludes that because brokerage tends to centralize trading and price information, it "uses time more efficiently," "eliminates the most important arbitrages," and benefits dealers by ensuring that orders are executed according to price priority.[15]

The efficiency gains of centralized price information may imply economies of scale and, thus, a natural monopoly for brokers in securities markets. This is entirely consistent with the textbook presentation of the relatively greater operational efficiency of centralized markets.[16] Thus, the fact that a number of brokers service the foreign exchange market seems to represent a discrepancy between theory and reality. Brokers do communicate among themselves, however, to eliminate the possibility of arbitrage between limit order books. While this helps explain the multiplicity of brokers, it does not fully resolve the issue of decentralization in the interbank direct market.

Temporal Consolidation

The distinction between a continuous market and a call market involves what Schwartz refers to as the degree of "temporal consolidation."[17] In a call market, trading occurs at pre-appointed times (the "calls"), with arriving transaction orders detained until the next call for execution. In continuous markets, like the foreign exchange market, trading occurs at its own pace, and transaction orders are processed as they arrive. A range of intermediate arrangements falls between these two extremes.

Most microeconomic models assume call markets. In a Walrasian *tâtonnement* model, for example, an auctioneer calls out a series of prices and receives buy and sell orders at each price. When a price is found for which the quantities supplied and demanded are equal, all transactions are consummated at that price. Interestingly enough, Walras based this price discovery model on the mechanics of the Paris Bourse.

Temporal consolidation can affect the performance of a market. Theoretical work indicates how continuous trading can alter allocations, the process of price discovery and even the ultimate equilibrium price.[18] The basic thrust of these arguments is that, with continuous trading, earlier transactions satisfy some consumers and producers, causing shifts in supply and demand that affect prices for later transactions. As a result, the Pareto-efficiency characteristic of Walrasian equilibria does not necessarily obtain in continuous markets.[19]

On the other hand, the periodic batching of orders that occurs in a call market also has disadvantages. The difference in time between order placement and execution can impose real costs on investors. A recurring argument in the literature is the willingness of investors to pay more — a liquidity premium — for the ability to trade immediately. Similarly, periodic calls delay any information conveyed by prices until the time of the call, introducing price uncertainty in the period between the calls.

In sum, a trade-off exists between the allocational efficiency of the nearly Walrasian call market system and the informational efficiency and immediacy of the continuous market system.[20] It is not clear whether the microstructure of the foreign exchange market represents a globally optimal balance of these relative ad-

[15]See Garbade (1978), p. 497.

[16]The textbook argument counts trips to market. Briefly, if there are N traders, then a total of N trips to a central marketplace are required for each to haggle with everyone else; to pair them bilaterally requires a total of N(N-1)/2 trips. If trips are costly, then centralization is more efficient.

[17]See Schwartz (1988), pp. 435-47. Garman (1976), pp. 257-58, also describes continuous and call markets; he refers to these as asynchronous and synchronous markets, respectively.

[18]See Hahn (1984), Negishi (1962), Beja and Hakansson (1977), as well as the references therein.

[19]A continuous market cannot be viewed as a continuum of infinitesimally lived call markets. Clearing supply and demand in each such call market would require an infinite trading volume over the course of a day. Cohen and Schwartz (1989) recommend an electronic order-routing system for the stock exchanges, to facilitate the placement and revision of orders. This would encourage additional trading volume, making more frequent calls feasible.

[20]See Stoll (1985), p. 72, and especially Schwartz (1988), pp. 442-53, for a more thorough exposition of the pros and cons of temporal consolidation. Intermediate arrangements are also possible. For example, Schwartz argues that many of the problems caused by infrequent batching in a call market might be overcome by expanding access to the market with computer technology, whereby the increased number of traders would allow for more frequent calls.

vantages. A persistent deviation from optimality might be explained, for example, by arguing that the allocational benefits of a call market system are a public good.

Communication of Prices

The terms "open-bid" and "limit-book" refer to ways in which price information is communicated. In an open-bid market — the open outcry system on the futures exchanges, for example — offers to buy or sell at a specified price are announced to all agents in the market. At the opposite extreme, in a sealed-bid market, orders are known only to the entity placing the order and perhaps to a disinterested auctioneer.

Direct trading in foreign exchange approximates the standard open-bid structure. The salient difference between the foreign exchange market and the standard arrangement is the bilateral pairing of participants in the foreign exchange market. In principle, any participant can contact a market-maker at any time for a price quote. The bilateral nature of such contacts and the time consumed by each contact together imply, however, that all participants cannot be simultaneously informed of the current quotes of a market-maker. This practical constraint on the dissemination of price information is significant: it introduces the possibility of genuine arbitrage, that is, of finding two market-makers whose current bid-ask spreads do not overlap.

The limit order book, which is used by both foreign exchange brokers and stock exchange specialists, is another intermediate form of price communication. Although it would be possible in principle for foreign exchange brokerage books to be fully open for public inspection, in practice only certain orders — namely, the best bid and ask on each book — are revealed to market-makers, while the others remain concealed. As in the direct market, market-makers must contact brokers bilaterally to get these "inside spreads." Knowledge of the concealed limit orders would be of speculative value to market-makers, because an imbalanced book suggests that large future price movements are more likely in one direction than the other.

More generally, price communication is intimately related to the role of market-makers as providers of "predictable immediacy."[21] Market participants are willing to pay a liquidity premium, usually embedded in a market-maker's spread, for the reduction in search costs implied by constant access to a counterparty. The costs of "finding" the other side of a transaction can be further broken down into the liquidity concession, the cost of communicating the information and the cost of waiting for potential counterparties to respond.[22] Other things equal, an efficient system of price communication is one that minimizes such transaction costs. While the communication of price information is a central function of securities markets, the fact that the systems of price communication in the foreign exchange market are not fully centralized suggests that these systems do not represent a cost-minimizing arrangement.

Structure of Prices

The terms "double-auction" and "single-auction" refer to the nature of the prices quoted. In a double-auction market, certain participants provide prices on both sides of the market, that is, both bid and ask prices. Participants providing double-auction quotes upon demand are known as market-makers, and they must have sufficient capitalization to back up their quotes. In a single-auction market, prices are specified either to buy or to sell, but not both. In the foreign exchange market, market-makers provide double-auction prices, while brokers try to aggregate single-auction quotes into two-way (inside) spreads. A broker's book may occasionally be empty on one or both sides. Rather than make a market in such cases, the broker provides, respectively, a single-auction quote or none at all.

Thus, whether double or single-auction prices are quoted depends largely on whether the agent quoting prices is providing market-making services or simply attempting to acquire (or sell) the commodity. This issue is related to the degree of centralization in the market. The absence of market-makers in a single-auction market, together with the presence of search costs, results in a tendency toward centralization of price information, thus facilitating the search for a counterparty. Inversely, decentralization of price information leads to a tendency

[21]This term is due to Demsetz (1968), p. 35. Tinic (1972), p. 79, calls in "liquidity services."

[22]See Logue (1975), p. 118.

toward double-auction prices, again to facilitate the search for a counterparty.[23]

MODELING TRADERS' BEHAVIOR

The microstructure literature extends well beyond a simple description of market institutions. Modeling the behavior of market participants is central to almost all discussions of microstructure. Although numerous approaches to such modeling have been taken, two common concerns are of special interest. These are the treatment of price information by market participants, and determination of the bid-ask spread. The latter raises the interrelated issues of inventory and quantity transacted.

Price Expectations

Modeling the interpretation of price information is a crucial step in constructing microstructural models of price discovery.[24] Many diverse approaches have been taken in such modeling. An almost universal simplification is to model securities markets in partial equilibrium, so that prices are not determined endogeneously in the traditional general equilibrium sense. This allows the modeler to focus on the microstructure's finer details. Another common simplification is to assume that agents ignore the impact of their own behavior on the market.[25]

Rather than explicitly model such forces as general equilibrium or recursive beliefs, models posit probability distributions that produce the prices of orders in the market. Modelers have included randomness at one or both of two levels, depending on their focus. First, order prices can be generated by objective distributions, that is, by stochastic processes exogenous to the market. For example, there may be a stochastic process that generates the "true" equilibrium price. Second, probability models of participants' subjective beliefs about prices can be used. Conroy and Winkler, for example, attribute subjective normal price distributions to market-makers, who use Bayesian updating to learn about the prices of incoming limit orders.[26] Objective processes can coexist with subjective beliefs about those processes. Harsanyi suggests a consistency requirement for the subjective price distributions of multiple agents; these distributions are each equated with a conditional distribution of a single distribution known to all.[27]

Models can be further classified according to how they relate supply and demand. In particular, there are both models with single price processes and with dual price processes. In dual price models, purchase orders (whether market or limit orders) are generated by one process, while sale orders are generated by another.[28] The salient point here is that purchase and sale orders come from independent distributions. This independence is especially clear in Conroy and Winkler, where the distributional assumptions are explicit; there, independence implies that any sequence of buy orders, regardless of their prices and quantities, has no effect on the subjective probability of a sell order at any price.

Statistical independence implicitly restricts the ways in which orders can be generated. Purchase and sale orders are somehow motivated independently, although the cause of this separation is not always specified. Statistical independence is not a necessary component of a dual price process, however. Cohen, Maier, Schwartz and Whitcomb (1981), for example, assume that actual market bid and ask prices are

[23]Note that the converse does not appear to hold. That is, centralization does not tend to eliminate double-auction quoting. For example, the NASDAQ system on the OTC stock market centralizes price information while still supporting numerous market-makers for every stock.

[24]Notably, the term "price" is generally too inexact in a microstructural context. One must often distinguish at a minimum between quoted prices, transaction prices and equilibrium prices. There are also reservation prices, market-clearing prices and closing prices (see Schwartz (1988), chapter 9, for the distinction between equilibrium and clearing prices). If unspecified here, the intended definition should be clear from the context.

[25]The alternative, which dates at least to Keynes' "beauty contest," is recursive beliefs, in which an agent considers the feedback of her own actions on the beliefs of others, and thence how the behavior on the other agents might affect her own beliefs, etc. See Keynes (1936), p. 156. The limiting case—an infinite recursion of beliefs—presumes extreme informational and computational resources on the part of agents, and models based on it are usually intractable. Intermediate approaches allowing a finite degree of recursion must somehow justify the truncation of recursive beliefs, just as the standard model of atomistic agents allows no beliefs about beliefs and is justified by an assumption on the relative size of individual agents.

[26]See shaded insert on opposite page.

[27]See Harsanyi (1982), especially chapter 9, and the references therein. His consistency requirement identifies a unique equilibrium for the game.

[28]A market order is an order to trade at the best price available; a limit order specifies a price. These models represent a strain of the literature that was pioneered by Demsetz using straightforward supply and demand schedules (see shaded insert on page 63). Similar approaches were later taken by Garman (1976), Amihud and Mendelson (1980) and Conroy and Winkler (1981), among others.

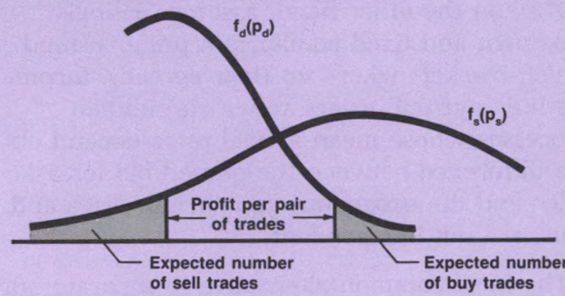
Bayesian Learning of Price Information

Conroy and Winkler (1981) developed a Bayesian model of market-maker price expectations, which is outlined here. Consider an expected-profit-maximizing, monopolistic market-maker who faces streams of buy and sell limit orders from investors.[1] All orders are for a single round lot. Assume that the market-maker believes that reservation prices of buy orders, p_d, are generated by a normal distribution, $F_d(p_d:\mu_d,\sigma_d)$; reservation prices of sell orders, p_s, are generated by a second, independent, normal distribution, $F_s(p_s:\mu_s,\sigma_s)$. That is, the market-maker has two independent, normal, subjective price distributions (with corresponding densities f_d and f_s). Further assume that the market-maker currently holds his desired inventory level. How should he set his spread?

The inventory condition implies that the chosen bid and ask rates, B and A, must satisfy the constraint, $F_s(B) = 1-F_d(A)$, so that the expected change in inventory is zero.[2] Given this constraint, the expected profit per period is: $E(\pi) = (A-B)\,F_s(B) = (A-B)\,(1-F_d(A))$. Maximizing this over B and A yields: $B^* = M - \sigma_s\Phi(B')/\phi(B')$ and $A^* = M + \sigma_d\Phi(A')/\phi(A')$, where $M \equiv (\sigma_d\mu_s+\sigma_s\mu_d)/(\sigma_d+\sigma_s)$, $B' \equiv (B-\mu_s)/\sigma_s$, $A' \equiv (A-\mu_d)/\sigma_d$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions. It can be shown that this optimal spread shrinks (i.e., A^*-B^* decreases), *ceteris paribus*, as the subjective variances, σ_s and σ_d decrease.

The important aspect of this study is that it provides an explicit mathematical model for a market-maker's interpretation of price information. The market-maker is assumed to behave in a Bayesian fashion, using the observed prices on incoming limit orders to

Profit Per Pair of Trades and Expected Number of Trades Per Unit Time



refine the parameters of his subjective distributions. For example, assume that the market-maker views purchase prices as coming from a normal distribution $g_d(p_d:\mu_d,\sigma_d)$, but is unsure about the mean of this distribution. Represent this uncertainty by a normal prior density $h'(\mu_d:m',v')$ over the possible values for the mean, μ_d. Given this, the marginal subjective density over the prices of incoming limit orders, $f_d(p_d) = \int g_d(p_d)\cdot h'(\mu_d)$, is normal with mean m′ and variance $(\sigma_d^2+v'^2)$. Following a sample of n buy orders with mean price m, the market-maker is able to refine his subjective distribution of the mean. The posterior parameters of $h''(\mu_d:m'',v'')$ are $m'' = (m'/v'^2+nm/\sigma_d^2)/(1/v'^2 + n/\sigma_d^2)$ and $v'' = 1/(1/v'^2 + n/\sigma_d^2)$. The upshot of this refined estimate is that the variance of the marginal subjective price density, $f_d(p)$, is now smaller, and the market-maker's optimal spread, (B^*, A^*), is narrower.

[1]Conroy and Winkler (1981) also consider a risk-averse market-maker and the information conveyed by a market order, which does not specify a price. They do not incorporate the impact of inventory on pricing, nor do they generalize beyond the unrealistic assumption of normally distributed prices.

[2]This is depicted in the figure above, where price is on the horizontal axis, and the relative frequency of orders is on the vertical axis. The market-maker's spread is the interval from B to A, and the inventory constraint is satisfied when the two shaded tails have equal area. Their optimization problem is similar in spirit to that of Allen (1977), although the latter does not consider learning.

independent Poisson processes and give investors joint subjective distributions over those prices. For the latter distributions, probabilistic independence of bid and ask prices is not explicitly required. Black (1989) models quantities (independent of prices) of market orders. Quantities supplied and demanded are drawn from different distributions, but the distributions are constrained to have the same mean. Garbade (1978), on the other hand, assumes a single, unknown and fixed equilibrium price, around which market-makers set their spreads. Incoming buy and sell orders arrive via random processes whose mean arrival rates depend on the difference between the quoted bid (or ask) price and the exogenous equilibrium price and, thus, are not independent.

The most common alternative to separate purchase and sale processes is to model prices as some function of a single scalar process. This approach is in the spirit of the efficient markets literature, which posits a unique value for a security conditional on the available information. Ross (1987) points out that this approach can be regarded conceptually as a special case of the dual price process, with supply and demand infinitely elastic at a common price. Many authors reveal their theoretical roots by using terminology drawn from the literature on efficient markets. Thus, for example, Barnea describes a stock's "intrinsic value," which follows a random walk.[29] Similarly, Copeland and Galai posit a "'true' underlying asset value ... known (ex ante) to all market participants."[30] In contrast, Garbade's (1978) exogenous equilibrium price is unknown.

It is possible to extend the single price approach beyond the efficient markets tradition by modeling the value of a security subjectively rather than as an objective fact. Glosten and Milgrom (1985), for example, begin with an exogenous random value representing the consensus value of a stock given all public information. Investors do not act on this exogenous value directly; instead, they act on their expectation of it, conditional on their information set. Ho and Stoll personalize price expectations in a similar fashion:[31]

> We take the dealer's opinion of the "true" price of the stock to be exogenously determined by his information set and ask how the dealer prices relative to his "true" price...

This subjectivization of the pricing process is significant, because it allows for heterogeneous expectations and thus for more realistic modeling of price discovery.

Research into the microstructure of the foreign exchange market should presume such heterogeneity among market-makers. There are numerous market-makers for foreign exchange: The Federal Reserve Bank of New York (FRB-NY) (1989a) lists 162 dealing institutions in the U.S. interbank market. There would be little point in such superfluity if all market-makers were identical. Furthermore, it is well known that "taking a view," that is, speculating on future prices, is routine for many participants.[32] To omit this heterogeneity from a model is to ignore an important characteristic of the market.

The large proportion of market-makers in the foreign exchange market has another important modeling implication. It implies that a single-price process is more appropriate as a theoretical representation of agents' expectations. Market-makers consistently face other market-makers, who can hold positive or negative inventories of foreign currency with equal ease. A quote that is "off the market" on the high side will be hit (i.e., traded upon) just as surely as a quote that is off on the low side. This is also true of customers, who normally enter the market with a predilection to either buy or sell. As Burnham notes:[33]

> The customer knows that if the first marketmaker is too far off the market price, he can unexpectedly take the other side of the quote and resell the position to a second marketmaker.

The point is that the market-maker must expect to be penalized for underestimating as well as overestimating his counterparty's valuation of the currency. From the perspective of the market-maker, who quotes a spread and observes a response, the forces determining short-run effective demand and supply are not merely related, but indistinguishable.

---

[29]See Barnea (1974), pp. 512-14.

[30]See Copeland and Galai (1983), p. 1458.

[31]See Ho and Stoll (1981), p. 48. For a similar example, see Stoll (1978), especially p. 1136.

[32]See, for example, Kubarych (1983), p. 29, or Burnham (1991), p. 139.

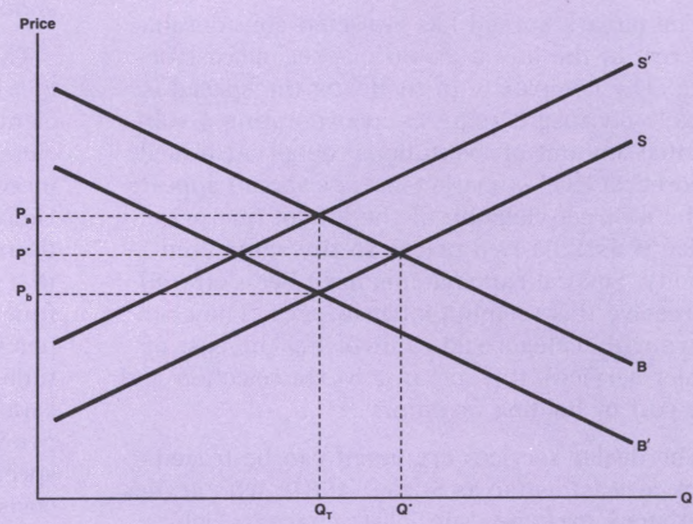
[33]See Burnham (1991), p. 136.

Dealer Services and the Bid-Ask Spread

Traditional wisdom refers to the bid-ask spread as the "jobber's turn," suggesting that it provides compensation to the dealer for the provision of services.[1] Demsetz (1968) formalized this rationale for the spread, defining the particular service provided as "predictable immediacy" and offering a simple model to describe the spread.

Consider a continuous market with aggregate supply (sell) and demand (buy) schedules, S and B, for a security (see figure at right). In an idealized world, investors would come together simultaneously, and the market would clear at price P^* and quantity Q^*. In this marketplace, however, such coordination of trading is impossible. By assumption, the market is continuous, and there is no mechanism (e.g., a limit order book) for holding orders over time. Thus, S and B do not represent standard static supply and demand schedules, but rather time rates of supply and demand. At any given instant, there may be no orders on either side.

Market Buy and Sell Curves with and without the Provision of Immediacy



Instead, we introduce a monopolistic market-maker who allows the trading to occur by standing in as a counterparty to all trades. In the process, he provides a service to investors that Demsetz labels "predictable immediacy." The market-maker knows the aggregate supply and demand propensities. The supply and demand curves that he presents to the public, S′ and B′, however, are both shifted leftward. Investor purchases clear for price P_a, at the intersection of the market demand schedule, B, with the market-maker's supply schedule, S′. Similarly, investor sales clear at the intersection of S and B′, for price P_b. The differences: $P_a - P^*$ and $P^* - P_b$ are liquidity premia. In the figure, the quantities, Q_T, purchased and sold by the market-maker happen to be equal, so that no market-maker inventory is accumulated. His profit thus equals $Q_T(P_a - P_b)$; this is the "jobber's turn."

[1]See, for example, Keynes (1936), p. 158, or Stigler (1964), p. 129.

A market-maker's constant contact with other well-capitalized market-makers implies that this is not a theoretical fine point. In the words of one market-maker:[34]

> Ninety percent of what we do is based on perception. It doesn't matter if that perception is right or wrong or real. It only matters that other people in the market believe it. I may know it's crazy. I may think it's wrong. But I lose my shirt by ignoring it.

In other words, as a direct implication of their readiness to buy or sell, market-makers must strive first to achieve a price consensus. The im-

[34]James Hohorst, as quoted by Mossberg (1988), p. 29R. Mr. Hohorst directed foreign exchange trading in North America for Manufacturers Hanover.

perative of arbitrage avoidance must be regarded as the first priority in individual market-maker pricing, to which all other factors (e.g., purchasing power parity) must be subordinated.

Market-makers' Bid-Ask Spreads

The bid-ask spread has attracted considerable interest in the literature on market microstructure. The complexity of modeling the spread is largely because it requires incorporating a substantial amount of institutional detail. At a facile theoretical level, a market-maker's spread appears to be a direct violation of the law of one price, since it assigns two prices to the same commodity. Several explanations have been offered to resolve this seeming inconsistency. They can be roughly categorized as involving the cost of dealer services, the cost of adverse selection and the cost of holding inventory.[35]

The dealer services argument can be traced back at least as far as Stigler (1964), who argues that stock exchange specialists charge a "jobber's turn" as compensation for the costs of acting as a specialist. The analysis of dealer services was formalized by Demsetz (1968), who identified "predictable immediacy" as the particular service for which investors are willing to pay. This identification hints at the complex question of what liquidity is and where it comes from. In a busy market, liquidity is a public good: a continuous stream of buyers and sellers generates predictable immediacy as a by-product of their trading.

The determinants of the level of compensation are themselves a topic of debate. Stigler argues that, because centralization of exchange limits fixed costs and aggregates separate transaction orders into less risky actuarial order flows, it implies economies of scale and thus a natural monopoly for market-making.[36] Smidt (1971) counters that barriers to entry among NYSE specialists allow them to exact monopoly rents from other investors. In his view, the natural monopoly argument, while used as an apology for barriers to entry, remains unsupported empirically: "There is no empirical evidence to support the proposition that [market-making] is, in fact, a natural monopoly."[37] Indeed, if market-making is a natural monopoly, barriers to entry should be unnecessary.

The foreign exchange market has no apparent barriers to entry other than the need for sufficient capitalization. It also has no apparent barriers to exit. The market supports a large and increasing number of competing market-makers. Unless it can be shown that there is some subtle restriction in the foreign exchange market that prevents consolidation of the market-making function, one must conclude that market-making per se is not a natural monopoly.[38] The multitude of market-makers also implies that they cannot earn monopoly rents by embedding a premium for predictable immediacy in the spread, although the spread may still cover the costs of processing orders.

Other research suggests that a market-maker's job is more complex than the mere sale of counterparty services. A second explanation for the bid-ask spread — adverse selection — can be traced to Bagehot (1971). He starts with "liquidity-motivated transactors" who pay the market-maker the price of the spread in exchange for the service of predictable immediacy. The market-maker also confronts traders who have inside information, however, and who can therefore speculate profitably at the expense of the market-maker.[39] The market-maker must charge everyone a wider spread to compensate for losses to the information-motivated traders.

Because of the relatively abstract nature of currencies as commodities, it is difficult to construct examples of "inside" information on foreign exchange rates. One exception is money supply announcements, which, if known before

---

[35]This is essentially the same taxonomy as provided by Barnea and Logue (1975), although they use the terms "liquidity theory," "adversary theory," and "dynamic price/inventory adjustment theory," respectively.

[36]See Stigler (1964), p. 129.

[37]See Smidt (1971), p. 64.

[38]For example, in the context of the OTC stock market, Benston and Hagerman (1974), p. 362, conjecture that, "dealers may face positively sloped marginal cost curves which shift down as industry output increases." The idea is that market-making per se is not a natural monopoly, even though the industry as a whole experiences economies of scale. Hamilton (1976) also addresses the natural monopoly question; Reinganum (1990) provides evidence on liquidity premia for NYSE vs. NASDAQ stocks.

[39]This situation is called adverse selection, because, in a market with competing market-makers, the one who gets the insider's business is a loser rather than a winner. Bagehot also posits a third class of investors, who only think they have inside information; they speculate, but lose on average, and are indistinguishable to the market-maker from the liquidity-motivated traders.

publicly distributed, might provide a basis for profitable speculation. Another form of information that can be construed as inside information is knowledge of an arbitrage opportunity. Consider a hypothetical market in which there are numerous decentralized market-makers who do not quote spreads, but single prices at which they are willing both to buy and sell. Unless there were a perfect consensus among the market-makers on the value of the foreign currency, all of them would be vulnerable to arbitrage. A decentralized market makes a perfect consensus difficult to achieve. Without centralizing price information, it is impossible to know if no arbitrage opportunities exist. A bid-ask spread, in contrast, allows a market-maker to include an error tolerance in her prices, thus facilitating a price consensus: it is easier to get bid-ask spreads to overlap than to get scalar prices to coincide. The spread also provides the market-maker with some degree of protection from adverse selection in the form of arbitrage.

The bid-ask spread is also affected by inventory considerations. This idea dates back at least as far as Barnea and Logue (1975).[40] The notion of a desired inventory level for the market-maker underlies all of these models. In the simplest case, the desired level is set at zero, and a constant spread is shifted up and down on a price scale to equalize the probability of receiving a purchase order with that of receiving a sale order. The result is that the expected change in inventory is always equal to zero, and (with all trades for one round lot) the inventory level follows a simple random walk.

An undesirable implication of random-walk models of inventory is the inevitable bankruptcy of the market-maker. Finite capitalization levels for market-makers impose upper and lower bounds on allowable inventories. Because inventory follows a random walk, with probability one it will reach either its upper or lower bound in a finite number of trades.[41] The dynamic optimization models of Bradfield (1979), Amihud and Mendelson (1980) and Ho and Stoll (1981) resolve this problem. They conclude that a market-maker, optimizing his bid and ask prices over time in the face of a stochastic order flow, will shift both bid and ask rates downward (upward) and increase the width of the spread when a positive (negative) inventory has accumulated.[42]

We should expect two of these three rationales for the spread to apply to market-makers' bid-ask spreads in the foreign exchange market. Because there are numerous market-makers, competition should eliminate their ability to earn monopoly rents by charging a premium for predictable immediacy per se. The adverse selection argument does apply in the foreign exchange market, however, since the spread allows market-makers some protection against arbitrage opportunities. Arbitrage opportunities can be construed as a form of inside information in a market where price information is not centralized. In accordance with the dynamic optimization models, a market-maker's inventory level should affect the spread, widening and shifting it as inventories accumulate.

Brokers' Spreads

So far, the discussion of the bid-ask spread has focused on models in which bid and ask prices are set by individual market-makers. The dual role of the stock exchange specialist suggests that this is only part of the story. Spreads are produced in two fundamentally different ways. It is only when limit orders are sparse that a NYSE specialist must step in as a market-maker to provide an "orderly market."[43] When limit order volume is sufficient, the specialist acts as a broker, accounting for incoming limit orders on the limit order book, and pairing market orders against them. Cohen, Maier, Schwartz and Whitcomb (1979) note that inadequate attention has been given to the fact that not all prices are market-maker spreads. The market often makes itself without specialist assistance, through the aggregation of limit orders on the book.

The foreign exchange market differs from the NYSE in that the market-making and brokerage roles are separated: market-makers do not act as brokers, and brokers do not make markets.

[40]Barnea and Logue attribute it to Smidt (1971), although Smidt's paper does not explicitly develop the connection between the market-maker's inventory and his spread. Formal models of the relationship between inventories and spreads can be found in Stoll (1978), Amihud and Mendelson (1980), Ho and Stoll (1981) and Sirri (1989), among others.

[41]See, for example, Ross (1983), pp. 106-07.

[42]See shaded insert on page 66.

[43]The NYSE defines this role in rule 104: "the specialist should maintain a continuous market with price continuity and close bid and asked prices, and minimize the effect of temporary disparity between public supply and demand." See Leffler and Farwell (1963), pp. 211-12.

Dynamic Price-Inventory Adjustment Models

Amihud and Mendelson (1980, 1982) provide a model of market-maker spread-setting that takes inventory into account. Assume that a market-maker faces order flows of buy and sell market orders that arrive according to independent Poisson processes. The buy and sell arrival rates (i.e., process intensities), d and s, respectively, depend on the ask and bid prices, P_a and P_b, that the market-maker quotes: $d = D(P_a)$ and $s = S(P_b)$. Denote the inventory level by $k \in \{-\Sigma,\ldots,\Lambda\}$, where Σ and Λ are the largest allowable short and long positions, respectively. Let d_k and s_k denote the order arrival rates when prices are set as functions of the inventory level: $d_k = D(P_a(k))$ and $s_k = S(P_b(k))$.

The expected sojourn at k (i.e., the time until an order arrives) is given by the Poisson processes known as $1/(d_k+s_k)$. The probability that the next order will be a buy order is $d_k/(d_k+s_k)$, and the probability that it will be a sell order is $s_k/(d_k+s_k)$. Thus, the expected cash flow per unit time at position k is given by:

$$Q(k) = \left[\frac{d_k}{d_k+s_k} \cdot P_a(k) - \frac{s_k}{d_k+s_k} \cdot P_b(k) \right] \cdot (d_k+s_k)$$

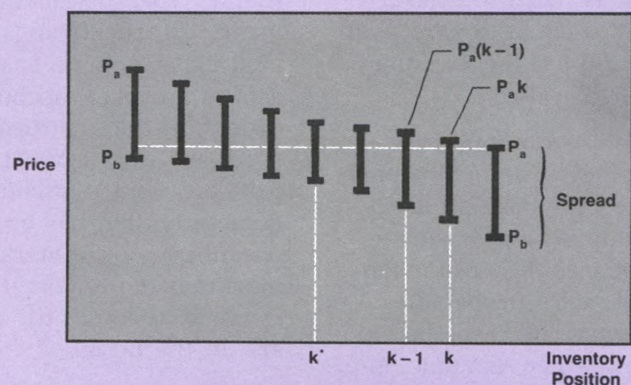
$$= d_k \cdot P_a(k) - s_k \cdot P_b(k).$$

The market-maker's objective is to maximize the expected profit per unit time, given by:

$$\pi = \sum_{k=-\Sigma}^{\Lambda} \Phi_k Q(k).$$

Where Φ is the probability of being at inventory level k. The solution to this optimization problem gives the values for $P_a(k)$ and $P_b(k)$, which are depicted in the figure below. The market-maker controls inventory by adjusting prices up (down) to make an investor sale (purchase) more likely when inventory is low (high). The spread must widen as the inventory nears its bounds, in order to avoid the problem of a random walk for inventory; at the extremes, $d_{-\Sigma} = s_{\Lambda} = 0$.

Optimal Prices and Bid-Ask Spread as a Function of Inventory Position



Therefore, it is even more appropriate to model brokered spreads as determined in a fundamentally different way from market-maker spreads. The separation of roles also has other implications for modeling foreign exchange brokerage.

A brokered spread is the combination of the best bid and best ask, received by the broker as separate limit orders. This arrangement might be modeled as a pair of extreme order statistics from independent distributions of purchase and sale limit orders. The distribution of these statistics would have to be conditional on limit order volume and on the fact that the best ask must always exceed the best bid, since crossing orders transact immediately and are removed from the book.[44] Perhaps because of its complexity, such a derivation has not been attempted.

Cohen, Maier, Schwartz and Whitcomb (1979) model limit orders as generated by "yawl" distributions. These distributions satisfy three heuristics for the incentives of investors placing limit orders.[45] The heuristics are motivated by a notion of the centralized exchange as a market for immediacy; placers of limit orders produce immediacy, and placers of market orders consume it. This relationship between limit and market orders is formalized in Cohen, Maier, Schwartz and Whitcomb (1981), where each half of the brokered spread is assumed to be generated by a compound Poisson process. A minimum brokered spread results: if the limit order's bid (ask) price is sufficiently close to the specialist's ask (bid), the benefit to the investor of being able to specify the price of a limit order is overwhelmed by the cost of foregone immediacy.

Because models of the informational content of brokered spreads are few, the literature offers little guidance in modeling brokered quotes in the foreign exchange market. The yawl distribution is the only explicit distributional form for brokered spreads in the literature. Unfortunately, its heuristic basis cannot be transferred directly to the foreign exchange market, because market-makers there differ from stock market investors. Indeed, this may be an instance in which the foreign exchange market informs microstructure theory rather than the other way around. The extant approaches to brokerage treat it as a service facilitating predictable immediacy. This aspect of brokerage is redundant in the foreign exchange market, because of the multitude of market-makers, each providing immediacy. This redundancy suggests instead that foreign exchange brokerage serves some other function.

One motive for trading through a foreign exchange broker is to maintain anonymity — the name of the bank placing a limit order is not revealed unless a deal is consummated and then only to the counterparty.[46] Anonymity is valuable, because revealing a need to buy or sell a currency puts a market-maker at a bargaining disadvantage. In addition, anonymity can help pair market-makers who ordinarily would not contact each other directly. These issues have not been explored at a theoretical level. Until an adequate microstructural model of the strategic benefits of anonymity is developed, the theoretical understanding of foreign exchange brokerage will be limited.

CONCLUSIONS

Students of the foreign exchange market can draw several lessons from the literature on market microstructure. The most fundamental of these is that the institutional details of exchange in a market can affect all aspects — price, allocational, informational and operational — of the market's efficiency. A multitude of market-makers who can provide liquidity, or predictable immediacy, arises in response to the decentralization of the market. As a result, search costs are reduced relative to a world without market-makers, because finding one of many market-makers amounts to finding a counterparty. Brokerage also reduces search costs by achieving a degree of centralization in price information.

An unanswered question is why the specific combination of trading structures characteristic of the foreign exchange market — a decentralized, open-book, direct arrangement and a quasi-centralized, limit-book, brokered arrangement — should coexist. Apparently, each structure has relative advantages, but a full analysis of these advantages is lacking. Is there a single microstructure that would combine the relative advantages of the direct and brokered arrangements? Put another way, why does the microstructure of the foreign exchange market differ from that of the stock exchanges, the futures pits and the OTC stock market? Answering these questions will require a fuller specification of the objectives of a trading system and a better understanding of the impact of microstructural arrangements on those goals.

These issues provide a motive for deeper investigation of the behavior of the foreign ex-

[44]An order statistic is defined as follows: the sample realizations of a finite number of independent random variables are ranked in increasing order, and the kth order statistic is the kth number in that list. For the foreign exchange market, the modeling is still more complex, since brokers compare books amongst themselves in the sense that incoming orders can cross against any book.

[45]The yawl distribution, named for its resemblance to a sailboat, is a probability distribution contrived for modeling the generation of buy (or sell) limit orders. See Cohen, Maier, Schwartz and Whitcomb (1979, 1983, 1986) for details.

[46]See Kubarych (1983), p. 16, Burnham (1991), p. 141, and Federal Reserve Bank of New York (1989b), p. 41-3.

change market and its participants. Market-makers are the crucial element: they provide all transaction prices in the market and are involved in at least one side of every deal. The microstructure literature has developed numerous models of the interpretation and setting of prices by traders. The diversity of expectations models used in the literature illustrates the importance of tailoring such models to the specific environment confronted by market participants. Given that a foreign exchange market-maker's double-auction quote can be hit on either side (bid or ask) with equal ease, he must try to maneuver his spread to bracket the market's consensus valuation of the foreign currency. In other words, suppliers and demanders of currency are indistinguishable to the market-maker ex ante. The inability to separate the forces determining effective demand from those determining effective supply in the very short run imply that a single-price expectations process (rather than a dual-price process) is appropriate in modeling market-makers in the foreign exchange market.

A market-maker's bid-ask spread serves several purposes. Competition among market-makers in the foreign exchange market implies that they should be unable to charge a monopoly premium for the service of predictable immediacy. Instead, the spread obviates the need for perfect price consensus by giving the market-maker some protection from arbitrageurs with superior price information. While arbitrage avoidance must be considered a primary goal in setting a market-maker's bid and ask quotes, the spread provides flexibility elsewhere. Just as arbitrage avoidance is concerned with accurately estimating current prices, speculation is concerned with estimating future prices. By changing in size and shifting up or down, the spread can control stochastically the market-maker's foreign currency inventory in the face of random order flows. Systematic empirical study of the effect of inventories on market-makers' spreads is still needed, however.

The brokered spread is less well understood than the market-maker's spread, and certain areas are ripe for further research. Theoretical models of brokered spreads are few. The existing rationales for brokerage maintain that it provides liquidity services. In the foreign exchange market, however, numerous market-makers make the liquidity services provided by brokerage superfluous. Descriptions of the foreign exchange market suggest instead that anonymity is an important motive for trading in the brokered market. Yet the strategic value of anonymity in foreign exchange quoting is not well understood at a theoretical level. In addition, there is not a clear understanding of the differences in price information between a market-maker's spread and a broker's spread; this too remains a topic for future research.

From a broader perspective, a better understanding of institutional choice and change as regards securities market microstructure is necessary. Most microstructural research has been devoted to analyzing the impact of microstructural factors on important economic variables, such as price and allocation. Relatively little attention has been paid to the effect of economic factors on the choice of an institutional microstructure.

REFERENCES

Allen, Wiliam A. "A note on uncertainty, transaction costs and interest parity," *Journal of Monetary Economics* (July 1977), pp. 367-73.

Amihud, Yakov, Thomas S. Y. Ho, and Robert A. Schwartz, eds. *Market Making and the Changing Structure of the Securities Industry*, (Lexington Books, 1985).

Amihud, Yakov, and Haim Mendelson. "Dealership Market: Market-Making with Inventory," *Journal of Financial Economics* (March 1980), pp. 31-53.

________ . "Asset Price Behavior in a Dealership Market," *Financial Analysts Journal* (May/June 1982), pp. 50-59.

________ . "Asset Pricing and the Bid-Ask Spread," *Journal of Financial Economics* (December 1986), pp. 223-49.

________ . "Liquidity and Asset Prices: Financial Management Implications," *Financial Management* (Spring 1988), pp. 5-15.

Bagehot, Walter. "The Only Game in Town," *Financial Analysts Journal* (March/April 1971), pp. 12-14, 22.

________. *Bank for International Settlements*. "Survey of Foreign Exchange Market Activity," (February 1990).

Barnea, Amir. "Performance Evaluation of New York Stock Exchange Specialists," *Journal of Financial and Quantitative Analysis* (September 1974), pp. 511-35.

Barnea, Amir, and Dennis E. Logue. "The Effect of Risk on the Market Maker's Spread," *Financial Analysts Journal* (November/December 1975), pp. 45-49.

Beja, Avraham, and Nils H. Hakansson. "Dynamic Market Processes and the Rewards to Up-to-date Information," *Journal of Finance* (May 1977), pp. 291-304.

Benston, George J., and Robert L. Hagerman. "Determinants of Bid-Asked Spreads in the Over-the-Counter Market," *Journal of Financial Economics* (1974), pp. 353-64.

Black, Stanley W. "Transaction Costs and Vehicle Currencies," *IMF Working Paper* WP/89/96, International Monetary Fund (November 21, 1989).

Bossaerts, Peter, and Pierre Hillion. "Market Microstructure Effects of Government Intervention in the Foreign Exchange Market," *Review of Financial Studies* (forthcoming).

Bradfield, James. "A Formal Dynamic Model of Market Making," *Journal of Financial and Quantitative Analysis* (June 1979), pp. 275-91.

Brooks, J. A. "The dealing room and the dealer," in Rudi Weisweiller, ed., *Managing a Foreign Exchange Department* (Woodhead-Faulkner, Cambridge, 1985), pp. 18-33.

Burnham, James B. "Current Structure and Recent Developments in Foreign Exchange Markets," in Sarkis J. Khoury, ed., *Recent Developments in International Banking and Finance, Volume IV* (Elsevier Science Publishers, Amsterdam, 1991), pp. 123-53.

Chrystal, K. Alec. "A Guide to Foreign Exchange Markets," this *Review* (March 1984), pp. 5-18.

Cohen, Kalman J., Gabriel A. Hawawini, Steven F. Maier, Robert A. Schwartz, and David K. Whitcomb. "Implications of Microstructure Theory for Empirical Research on Stock Price Behavior," *Journal of Finance* (May 1980), pp. 249-57.

Cohen, Kalman J., Steven F. Maier, Walter L. Ness, Jr., Hitoshi Okuda, Robert A. Schwartz, and David K. Whitcomb. "The impact of designated market makers on security prices," *Journal of Banking and Finance* (November 1977), pp. 219-47.

Cohen, Kalman J., Steven F. Maier, Robert A. Schwartz, and David K. Whitcomb. "The Returns Generation Process, Returns Variance, and the Effect of Thinness in Securities Markets," *Journal of Finance* (March 1978a), pp. 149-67.

________ . "Limit Orders, Market Structure, and the Returns Generation Process," *Journal of Finance* (June 1978b), pp. 723-36.

________ . "Market Makers and the Market Spread: A Review of Recent Literature," *Journal of Financial and Quantitative Analysis* (November 1979), pp. 813-35.

________. "Transaction Costs, Order Placement Strategy, and Existence of the Bid-Ask Spread," *Journal of Political Economy* (April 1981), pp. 287-305.

________ . "A simulation model of stock exchange trading," *Simulation* (1983), pp. 181-91.

________ . *The Microstructure of Securities Markets* (Prentice-Hall, Englewood Cliffs, 1986).

Cohen, Kalman J., and Robert A. Schwartz. "An Electronic Call Market: Its Design and Desirability," in Henry C. Lucas, Jr. and Robert A. Schwartz, eds., *The Challenge of Information Technology for the Securities Markets: Liquidity, Volatility, and Global Trading* (Dow Jones-Irwin, Homewood, 1989) pp. 20-58.

Conroy, Robert M., and Robert L. Winkler. "Informational Differences between Limit and Market Orders for a Market Maker," *Journal of Financial and Quantitative Analysis* (December 1981), pp. 703-24.

Copeland, Thomas E. "A Model of Asset Trading under the Assumption of Sequential Information Arrival," *Journal of Finance* (September 1976), pp. 1149-68.

Copeland, Thomas E., and Dan Galai. "Information Effects on the Bid-Ask Spread," *Journal of Finance* (December 1983), pp. 1457-69.

Demsetz, Harold. "The Cost of Transacting," *Quarterly Journal of Economics* (February 1968), pp. 33-53.

Fama, Eugene F. *Foundations of Finance* (Basic Books, 1976).

Federal Reserve Bank of New York. "Summary of Results of U.S. Foreign Exchange Market Turnover Survey Conducted in April 1989," (September 13, 1989a).

________. "The Foreign Exchange Committee: Annual Report 1988," (June 1989b).

________. "Gross Foreign Exchange Transactions by Banks in U. S. Market Exceeds $23 Billion Daily: Broker Turnover Separately Amounts to $8.5 Billion Daily, New York Fed Survey Shows," (June 23, 1980).

Flood, Mark D. "Market Structure and Inefficiency in the Foreign Exchange Market," Working Paper 91-001A, Federal Reserve Bank of St. Louis (July 1991).

Garbade, Kenneth D. "The Effect of Interdealer Brokerage on the Transactional Characteristics of Dealer Markets," *Journal of Business* (July 1978), pp. 477-98.

Garman, Mark B. "Market Microstructure," *Journal of Financial Economics* (June 1976), pp. 257-75.

Glassman, Debra. "Exchange Rate Risk and Transaction Costs: Evidence from Bid — Ask Spreads," *Jornal of International Money and Finance* (December 1987), pp. 479-490.

Glosten, Lawrence R., and Paul R. Milgrom. "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics* (March 1985), pp. 71-100.

Hahn, Frank H. *Equilibrium and Macroeconomics,* (Basil Blackwell, Oxford, 1984).

Hamilton, James L. "Competition, Scale Economies, and Transaction Cost in the Stock Market," *Journal of Financial and Quantitative Analysis* (December 1976), pp. 779-802.

Harsanyi, John C. *Papers in Game Theory,* (D. Reidel, Dordrecht, 1982).

Ho, Thomas S. Y., and Richard G. Macris. "Dealer Market Structure and Performance," in Yakov Amihud, Thomas S. Y. Ho, and Robert A. Schwartz, eds., *Market Making and the Changing Structure of the Securities Industry* (Lexington Books, Lexington, 1985), pp. 41-66.

Ho, Thomas S. Y., and Hans R. Stoll. "On Dealer Markets under Competition," *Journal of Finance* (May 1980), pp. 259-68.

________ . "Optimal Dealer Pricing under Transactions and Return Uncertainty," *Journal of Financial Economics* (March 1981), pp. 47-73.

Keynes, John M. *The General Theory of Employment, Interest and Money* (MacMillan, London, 1936).

Kubarych, Roger M. *Foreign Exchange Markets in the United States* (Federal Reserve Bank of New York, 1983).

Leffler, George L., and Loring C. Farwell. *The Stock Market*, (Ronald Press, New York, 1963).

Logue, Dennis E. "Market-Making and the Assessment of Market Efficiency," *Journal of Finance* (March 1975), pp. 115-23.

Lyons, Richard K. "Private Beliefs and Information Externalities in the Foreign Exchange Market," Working Paper, Columbia University (November 1991).

Mossberg, Walter S. "Making Book on the Buck," *Wall Street Journal*, September 23, 1988.

Negishi, Takashi. "The Stability of a Competitive Economy: A Survey Article," *Econometrica* (October 1962), pp. 635-69.

Plott, Charles R. "Industrial Organization Theory And Experimental Economics," *Journal of Economic Literature* (December 1982), pp. 1485-527.

_______ . "Will Economics Become an Experimental Science?" *Southern Economic Journal* (April 1991), pp. 901-19.

Rai, Anoop K. "A Market Micro-Structure Investigation of Dealer Quotes and Risk Premium in the Foreign Exchange Markets," Working Paper, (1991).

Reinganum, Marc R. "Market Microstructure and Asset Pricing: An empirical investigation of NYSE and NASDAQ securities," *Journal of Financial Economics* (November/December 1990), pp. 127-47.

Riehl, Heinz, and Rita M. Rodriguez. *Foreign Exchange and Money Markets: Managing Foreign and Domestic Currency Operations* (McGraw-Hill, 1983).

Ross, Sheldon M. *Stochastic Processes*, (John Wiley & Sons, 1983).

Ross, Stephen A. "The Current Status of the Capital Asset Pricing Model (CAPM)," *Journal of Finance* (June 1978), pp. 885-901.

_______ . "The Interrelations of Finance and Economics: Theoretical Perspectives," *American Economic Review* (May 1987), pp. 29-34.

Schreiber, Paul S., and Robert A. Schwartz. "Efficient Price Discovery in a Securities Market: The Objective of a Trading System," in Yakov Amihud, Thomas S. Y. Ho, and Robert A. Schwartz, eds., *Market Making and the Changing Structure of the Securities Industry* (Lexington Books, Lexington, 1985), pp. 19-39.

_______ . "Price Discovery in Securities Markets," *Journal of Portfolio Management* (Summer 1986), pp. 43-48.

Schwartz, Robert A. *Equity Markets: Structure, Trading, and Performance* (Harper & Row, New York, 1988).

Sirri, Erik R. "Bid/Ask Spread, Price, and Volume in a Specialist Market," Working Paper, University of California, Los Angeles (February 1989).

Smidt, Seymour. "Which Road to an Efficient Stock Market," *Financial Analysts Journal* (September/October 1971), pp. 18-20, 64-69.

Stigler, George J. "Public Regulation of the Securities Markets," *Journal of Business* (April 1964), pp. 117-42.

Stoll, Hans R. "The Supply of Dealer Services in Securities Markets," *Journal of Finance* (September 1978), pp. 1133-51.

_______ . "Alternative Views of Market Making," in Yakov Amihud, Thomas S. Y. Ho, and Robert A. Schwartz, eds., *Market Making and the Changing Structure of the Securities Industry* (Lexington Books, Lexington, 1985), pp. 67-91.

_______ . "Inferring the Components of the Bid-Ask Spread: Theory and Empirical Tests," *Journal of Finance* (March 1989), pp. 115-34.

Tinic, Seha M. "The Economics of Liquidity Services," *Quarterly Journal of Economics* (February 1972), pp. 79-93.

Tinic, Seha M., and Richard R. West. "Competition and the Pricing of Dealer Service in the Over-the-Counter Stock Market," *Journal of Financial and Quantitative Analysis* (June 1972), pp. 1707-27.

Varian, Hal R. "Differences of Opinion in Financial Markets," in Courtenay C. Stone, ed., *Financial Risk: Theory, Evidence and Implications* (Kluwer Academic Publishers, Boston, 1989), pp. 3-37.

Wei, Shang-Jin. "Anticipations of Foreign Exchange Volatility and Bid — Ask Spreads," International Finance Discussion Paper No. 409, Board of Governors of The Federal Reserve System (August 1991).

Wolinsky, Asher. "Information Revelation in a Market with Pairwise Meetings," *Econometrica* (January 1990), pp. 1-23.

FEDERAL RESERVE BANK OF ST. LOUIS
REVIEW INDEX 1991

JANUARY/FEBRUARY

R. Alton Gilbert, "Do Bank Holding Companies Act as 'Sources of Strength' for Their Bank Subsidiaries?"

Cletus C. Coughlin, "A Consumer's Guide to Regional Economic Multipliers"

Keith M. Carlson, The Future of Social Security: An Update"

James B. Bullard, Learning, Rational Expectations and Policy: A Summary of Recent Research"

MARCH/APRIL

John A. Tatom, "Should Government Spending on Capital Goods Be Raised?"

Alison Butler, "A Case for a Bilateral Trade Deficit"

Mark D. Flood, "An Introduction to Complete Markets"

David A. Dickey, Dennis W. Jansen and Daniel L. Thornton, "A Primer on Cointegration with an Application to Money and Income"

MAY/JUNE

John A. Tatom, "Public Capital and Private Sector Performance"

R. Alton Gilbert, "Supervision of Undercapitalized Banks: Is There a Case for Change?"

James B. Bullard, "The FOMC in 1990: Onset of Recession"

Allan H. Meltzer, "U.S. Policy in the Bretton Woods Era"

JULY/AUGUST

Cletus C. Coughlin, "U.S. Trade-Remedy Laws: Do They Facilitate or Hinder Trade?"

Jeffrey D. Karrenbrock, "The Behavior of Retail Gasoline Prices: Symmetric Or Not?"

Michael T. Belongia, "Monetary Policy and the Farm/Nonfarm Price Ratio: A Comparison of Effects in Alternative Models"

Michelle R. Garfinkel and Daniel L. Thornton, "The Multiplier Approach to the Money Supply Process: A Precautionary Note"

Cletus C. Coughlin and Thomas B. Mandelbaum, "Measuring State Exports: Is There a Better Way?"

SEPTEMBER/OCTOBER

Keith M. Carlson, "The U.S. Balance Sheet: What Is It and What Does It Tell Us?"

Piyu Yue and Robert Fluri, "Divisia Monetary Services Indexes for Switzerland: Are They Useful for Monetary Targeting?"

Steven Russell, "The U.S. Currency System: A Historical Perspective"

NOVEMBER/DECEMBER

John A. Tatom, "The 1990 Oil Price Hike in Perspective"

Mark D. Flood, "Microstructure Theory and the Foreign Exchange Market"

Piyu Yue, "A Microeconomic Approach to Estimating Money Demand: The Asymptotically Ideal Model"

Michelle R. Garfinkel and Daniel L. Thornton, "Alternative Measures of the Monetary Base: What Are the Differences and Are They Important?"

Federal Reserve Bank of St. Louis
Post Office Box 442
St. Louis, Missouri 63166

The* Review *is published six times per year by the Research and Public Information Department of the Federal Reserve Bank of St. Louis. Single-copy subscriptions are available to the public free of charge. Mail requests for subscriptions, back issues, or address changes to: Research and Public Information Department, Federal Reserve Bank of St. Louis, P.O. Box 442, St. Louis, Missouri 63166.