Vol. 25, No. 3

ECONOMIC REVIEW

1989 Quarter 3

FEDERAL RESERVE BANK
OF CLEVELAND

ECONOMIC REVIEW

1989 Quarter 3
Vol. 25, No. 3

Economic Review is published quarterly by the Research Department of the Federal Reserve Bank of Cleveland. Copies of the *Review* are available through our Public Affairs and Bank Relations Department, 216/579-2157.

Coordinating Economist:
Randall W. Eberts

Editors: Paul J. Nickels
Robin Ratliff
Design: Michael Galka
Typography: Liz Hanna

ISSN 0013-0281

The Stability of Money Demand, Its Interest Sensitivity, and Some Implications for Money as a Policy Guide

by John B. Carlson

John B. Carlson is an economist at the Federal Reserve Bank of Cleveland. The author acknowledges helpful discussions and comments from Charles Carlstrom, William Gavin, and George Tavlas. Thanks also to Christine Dingledine for meticulous research assistance.

Introduction

The money demand function is one of the most closely studied relationships in economics. One reason is that the question of the stability of money demand has long been central to issues of monetary theory. This largely reflects the influential restatement of the quantity theory of money by Milton Friedman (1956): "The quantity theory is in the first instance a theory of money demand." Further, he argued, "The quantity theorist accepts the empirical hypothesis that the demand for money is highly stable—more stable than functions such as the consumption function that are offered as alternative key relations."

Friedman did not specify precisely the meaning of "highly stable" or "more stable." Presumably, highly stable implies that the parameters of the money demand function do not change over time. Thus, one would expect that any reasonable specification of money demand might satisfy some sort of in-sample stability test (for example, Chow test) at a minimum. The notion that money demand is more stable than other "key" relationships has been interpreted in the context of a simple IS-LM framework by Poole (1970). In essence, "more stable" implied that the variance of the money demand function was *relatively* smaller than the variance of the IS curve.

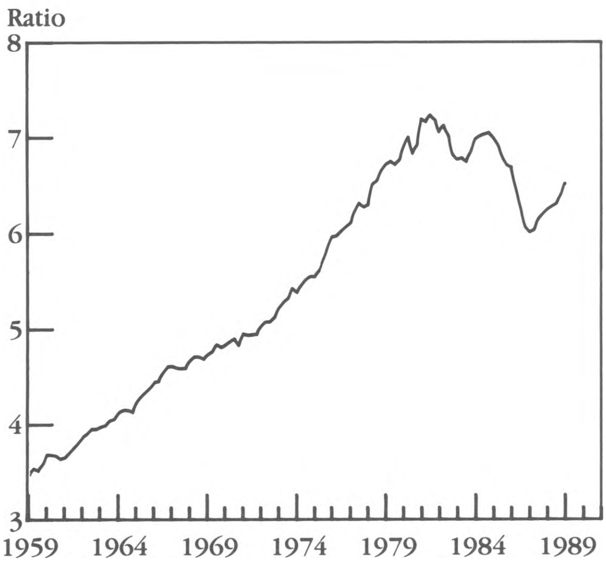
For years, the question of stability was simply examined by estimating various specifications of money demand, including both long-run and short-run models. It was commonly affirmed that money demand was a function of relatively few variables, including income and interest rates. By the mid-1970s, a consensus seemed to emerge that money demand was indeed one of the more stable relationships in economics, reliable enough to serve as a basis for formulating monetary policy.

Unfortunately, just as a consensus seemed to develop, many of the estimated relationships broke down, first around 1974, and again around 1982. By the mid-1980s, it appeared as though many economists had given up on finding a specification of money demand that might be stable, in either the short or the long run.

Recently, however, several researchers have found evidence that some specifications of money demand have remained stable through events of the 1970s and 1980s. One common conclusion of these studies is that money demand is highly interest sensitive—more so than many economists previously thought, particularly in

FIGURE 1

M1 Velocity



SOURCE: Board of Governors of the Federal Reserve System.

the long run. The magnitude of the interest elasticity of money demand has important implications for the role of money in the economy and hence for the conduct of monetary policy.

Much of the early debate about the role of money centered on how interest rates affected the velocity of money. Some analysts argued that interest-rate changes had little effect on velocity in the short or long run. Moreover, some presumed that M1 velocity had an inherent trend growth rate of about 3 percent. These assumptions now appear to be clearly refuted by the experience of the 1980s.

This paper reviews some recent findings of the research on money demand and considers the implications of these findings for monetary policy and rules. Section I reviews briefly a common specification of M1 demand that misled many economists about the importance of interest rates. Section II examines recent evidence that long-run equilibrium demand for the narrow money measure continues to be a stable function of relatively few variables.

The implications of these findings for the apparent shift in M1 velocity are discussed in section III. Section IV reviews the evidence that M2 demand is stable in the short run. In section V, the findings on M2 demand are reconciled with evidence that M2 velocity is trend stationary. The policy implications of the common finding that money demand is substantially interest sensitive are analyzed in sections VI and VII. Section VIII offers some concluding thoughts.

I. The Demand for M1 Before 1980

Until the 1980s, most attention in the money demand literature was given to M1—the money measure that then included currency and non-interest-bearing demand deposits. Focus on this measure reflected both theoretical and pragmatic considerations. First, M1 was the closest measure of pure transactions balances and hence conformed well to the concepts embodied in the inventory-theoretic model of Baumol (1952) and the portfolio-choice theory of Tobin (1958). These approaches essentially explained why individuals would hold the non-interest-bearing components of M1 instead of interest-bearing alternatives.

Perhaps more important, the focus on M1 seemed justified on empirical grounds. Of the various money measures, M1 appeared to be most closely related to economic activity, particularly in the short run. Movements in M1 served as a relatively useful indicator of current and future changes in economic activity. Moreover, the velocity of M1 exhibited a high degree of stability. From 1959 to 1980, M1 velocity increased at a trend rate of around 3 percent, deviating only a few tenths of a percent from year to year (see figure 1).

By the 1970s, a conventional empirical model for M1 demand had evolved.[1] Desired real M1 balances, m^*, were a function of some scale variable, y, either real income or wealth; and a measure of the opportunity cost of holding money, r, the level of interest rates:

$$(1) \quad m^* = \alpha_0 + \alpha_1 y - \alpha_2 r .$$

Earlier studies used annual data (see Meltzer [1963], Laidler [1966], and Chow [1966]). In these studies, the scale variable was typically some measure of wealth, and the opportunity cost was most often a measure of the long-term interest rate. The interest elasticities for M1 ranged between –0.7 and –0.9.[2]

■ 1 See, for example, Goldfeld (1973).

■ 2 For a more complete discussion of earlier studies, see Havrilesky and Boorman (1978), chapters 7 and 8.

Later studies in money demand used quarterly data, perhaps motivated by the increasing availability of such data and the development of quarterly econometric models (see Goldfeld [1973]). It became more common to use real income as the scale variable and to use a measure of the short-term interest rate as the measure of opportunity cost. It was often assumed that in any given quarter, money balances adjusted only partially to their desired (equilibrium) level. The adjustment process was specified as

$$(2) \quad m_t - m_{t-1} = \lambda(m^* - m_{t-1}),$$

where λ is the speed of adjustment to equilibrium. Substituting equation (1) into (2) yields

$$(3) \quad m_t = \lambda\alpha_0 + \lambda\alpha_1 y_t - \lambda\alpha_2 r_t + (1 - \lambda)m_{t-1}.$$

Equation (3) was sometimes estimated in first-difference form.[3]

The speed of adjustment of M1 balances to equilibrium levels was typically estimated to be between 0.25 and 0.5 per quarter. The estimates of income elasticities of this specification were typically around 0.2 in the short run and less than unity in the long run. Estimates for interest-rate elasticities were around -0.02 in the short run and ranged between -0.05 and -0.15 in the long run.[4]

The estimates of *long-run* interest elasticities seemed lower than the theories predicted and were substantially lower than earlier estimates. Given the absence of any evident interest-rate effects on M1 velocity and the apparent stability of the short-run specifications through the early 1970s, the smaller estimates of interest elasticity appeared to have gained greater acceptance.

By the 1980s, however, the quarterly specifications for M1 demand failed miserably. This was evident in the sharp change in the behavior of M1 velocity, which has varied substantially since 1980 and exhibits no clear trend. The breakdown in the conventional relationship is believed to be largely a consequence of disinflation and financial deregulation.[5]

Disinflation and financial deregulation greatly affected the opportunity cost of M1. Disinflation resulted in sharply falling interest rates, reversing the secular trend that dated back to the 1950s. Deregulation allowed banks to compete more effectively for funds by offering interest-bearing checking accounts and market rates of interest on savings and time deposits. The opportunity cost of most bank deposits fell markedly after 1982 when market rates fell and when banks priced deposits more competitively.

II. M1 Demand Revisited

While attempts have been made to rectify M1 demand in the short run, no consensus appears to be forming on any particular specification (see Moore, Porter, and Small [1988]). Many analysts now question whether a short-run demand function can ever be identified for M1.[6] On the other hand, recent studies by Poole (1988) and by Hoffman and Rasche (1989) suggest that the long-run (equilibrium) relationship may have endured through the past decade. Their specifications find that the long-run equilibrium interest elasticity of M1 demand is substantial.

Poole offers an explanation for why some economists may have been misled from models estimated in first-difference form. Such models often included a constant term, which made it equivalent to a linear-time-trend specification in a regression using the levels of the data. He concludes that in the postwar period, the constant term incorrectly picked up the trend in velocity, which should have been attributed to the postwar trend in interest rates.

This argument fails to explain, however, why the regressions for M1 in levels form (without time-trend variables) also underestimated interest elasticities. Closer inspection of the conventional relationships reveals that part of the trend effect of interest rates on M1 may have mistakenly been attributed to the trend in income. As noted above, the long-run income elasticity was typically estimated to be less than one—often around one-half. This, in turn, implied that over long periods, velocity would increase at approximately half the rate of increase in income, other things being equal. Since the conventional estimate of income elasticity concurred with the

■ 3 The inclusion of lagged money was also rationalized on an expectational basis (see Havrilesky and Boorman).

■ 4 Some specifications included interest paid on passbook savings deposits as an additional measure of opportunity cost.

■ 5 Some economists believe that the breakdown in the conventional relationship was also a consequence of the change in the Federal Reserve's operational procedure in October 1979 and the implications of that regime change on structural coefficients.

■ 6 Poole (1988) discusses the difficulties of identification from a buffer-stock perspective of money demand and concludes that the econometric problems may well be insurmountable. For a review of the buffer-stock approach to money demand, see Laidler (1984).

inventory-theoretic models of transactions balances, many analysts accepted the low estimate as a confirmation of the theory.[7]

To estimate long-run money demand, Poole advocates a simple regression of the level of velocity on the level of a long-term interest rate using annual data. By excluding income as an explanatory variable, Poole implicitly constrains the income elasticity to be unitary; hence, any potential trend in velocity must be independent of any trend in income.

Poole's case for using a long-term interest rate is predicated on the assumption that equilibrium money demand would not likely be affected by *temporary* changes in interest rates in the long run. Investment in cash management techniques is costly and hence only profitable when interest-rate increases are sustained. Since long-term rates are believed to embody expectations about future short-term rates, a rise in long-term rates is likely to indicate a more permanent rise in the general level of interest rates. Thus, Poole concludes, long-term rates better measure the opportunity cost of cash.

Finally, Poole argues that adequate estimates of a money-demand function cannot be obtained by using postwar data alone. During this period, both short- and long-term rates rose secularly. Thus, he uses an extensive sample period, 1915-1986, and three different subsamples. He estimates that the interest elasticity is around –0.6 for the whole period and for various subsamples, which is substantially larger than conventional estimates.

Hoffman and Rasche obtain estimates of a similar order of magnitude using a different estimation and testing method. Unlike Poole, they do not constrain the income elasticity to be unitary. Their approach—based on the notion of cointegration—addresses a potential problem related to the statistical properties of the variables included in money demand.

As with most economic variables, M1, interest rates, and income are nonstationary in levels. In such variables, there is no tendency to systematically return to a unique level or trend over time. It is now well known that standard regression analysis can yield spurious relationships between variables when the variables drift over time.

Methods initially developed by Engle and Granger (1987) allow one to examine whether equilibrium relationships exist between nonstationary variables. Such variables are said to be cointegrated, if some linear combination of them is stationary. Thus, cointegration implies a long-run equilibrium relationship between variables, and one can obtain long-run elasticities from the cointegrating vector.[8]

Hoffman and Rasche test for cointegration and find that 1) real M1 balances and real income are not cointegrated by themselves; 2) real M1, real income, and the interest rate are cointegrated with one cointegrating vector; and 3) one cannot reject the hypothesis that the coefficients of real money and real income in the cointegrating vector are equal in value but opposite in sign.[9]

The first result is consistent with the common finding that M1 velocity is nonstationary. Since both income and money are nonstationary, but not cointegrated, their difference will be nonstationary. The second result, however, implies a stable long-run relationship between money, income, and interest rates. The third result implies that it is appropriate to interpret the cointegrating vector as a linear combination of M1 velocity and interest rates or, equivalently, that the equilibrium real income elasticity of demand for real balances is unity.

To estimate the equilibrium interest-rate elasticity, Hoffman and Rasche consider both a short-term rate (three-month Treasury bill) and a long-term rate (10-year Treasury bond). Like Poole, they find that the interest elasticity on the long-term rate is about –0.6, while somewhat less, –0.4, for the short-term rate. Moreover, they find that cointegration holds for either of the long- or short-term measures. These results are robust across subsample periods investigated.

III. M1 Velocity in the 1980s

The Hoffman and Rasche findings imply that any observed drift in the velocity of M1 should be proportional to any drift in nominal interest rates. Thus, any shift in the drift of velocity should be the mirror image of any shift in the drift of nominal interest rates. Rasche (1989) investigates this last property by examining regressions of the changes in the log of M1 velocity and changes in the nominal interest rate, each against a constant and a dummy variable, which is zero through December 1981 and 1.0 thereafter.

■ **7** Other economic explanations for why an income elasticity might be less than one include improvements in cash management technology.

■ **8** For a more precise description of the concepts of cointegration, see Engle and Granger (1987).

■ **9** All variables are in log form.

The results indicate significant shifts in the interest-rate equation and in the velocity equation, both in the same direction. Again, the results hold for both long- and short-term rates; but, because of the high variance in the short-term rates, the shift is not measured with any precision. Rasche concludes that the abrupt change in the pattern of M1 velocity in the early 1980s was indeed associated with a coincidental change in the drift in interest rates.

Rasche further investigates the hypothesis that the observed change in velocity behavior is a result of a break in inflationary expectations. He argues that if the postwar period through 1980 is characterized by a steady upward drift in inflation, then it is reasonable to conjecture that it has been associated with the observed positive drift in nominal interest rates. Moreover, he argues that if inflation expectations stabilized at a lower rate in the early 1980s, it is reasonable to conclude that there has been no drift in interest rates over this period.

As evidence for a break in the drift of inflationary expectations, Rasche notes the general consistency of the Livingston Survey data. These data, which begin in the late 1940s, provide annual inflation forecasts formed at the end of the previous year. The survey reveals a general upward trend through 1980 and then a break sharply downward. Rasche notes that since 1982, the Livingston series has fluctuated without a trend in the 3 percent to 5 percent range.

To summarize, the recent evidence of large *long-run* interest elasticities of M1 demand provides a basis for understanding the recent shift in the trend in velocity. While the evidence points to a reasonably stable long-run M1 demand function, no one yet seems to have identified a satisfactory short-run model. Without a reliable short-run model of M1, little can be said about M1 velocity in the short run.

IV. The Demand for M2

Recent research on M2 demand provides evidence of stable specifications for M2 in the short run, at least in the postwar period. Moore, Porter, and Small (1988) estimate a short-run M2 demand function over the period 1964:IQ to 1986:IIQ.[10] The model is specified in two parts. One is an equilibrium money demand function, similar to equation (1):

$$(4)\quad m_t = \alpha + y_t + \beta s_t + e_t,$$

where m_t = log (M2), y_t = log (nominal GNP), and s = log (opportunity cost). Note that the unitary coefficient on nominal GNP assures that this also specifies a velocity relationship.[11] The second component is a dynamic specification based on an error-correction adjustment:

$$(5)\quad \Delta m_t = a + be_{t-1} + \sum_{i=1}^{u} c_i \Delta m_{t-i} + \sum_{i=0}^{v} d_i\, \Delta s_{t-i} + \sum_{i=0}^{w} f_i\, \Delta y_{t-i} + \epsilon_t,$$

where e_{t-1} is the deviation of money from its long-run equilibrium value (derived from [4]) and ϵ_t is white noise.

Equation (5) essentially specifies the short-run convergence process of M2 to its equilibrium value. When the coefficient b is negative, convergence is assured. Substituting equation (4) into (5) yields

$$(6)\quad \Delta m_t = a - b\alpha - b\beta s_{t-1} + b(m_{t-1} - y_{t-1}) + \sum_{i=1}^{u} c_i \Delta m_{t-i} + \sum_{i=0}^{v} d_i \Delta s_{t-i} + \sum_{i=0}^{w} f_i \Delta y_{t-i} + \epsilon_t.$$

Moore et al. estimate a version of equation (6). Simulations, both in-sample and out-of-sample, support the hypothesis that M2 demand has been and continues to be reasonably stable over the whole sample period.

One key feature of Moore et al. is the way opportunity cost is measured. By definition, the opportunity cost of money is the forgone interest income of holding a monetary asset. Over the years, it has been common to use a market yield on a relatively risk-free asset, such as a Treasury bill, to measure opportunity cost. For much of the postwar period, this seemed appropriate for the narrow money measures, since holders of currency and demand deposits did not receive explicit interest payments on these instruments.

Many instruments in the broader monetary aggregates like M2, however, have yielded explicit interest. Their yields, when not exceeding interest-rate ceilings, responded at least partially to market conditions. Moore et al. measure the opportunity cost of these instruments as the

■ **10** For further evidence concerning the stability of M2, see Hetzel and Mehra (1987).

■ **11** Moore et al. include a time index as a regressor to directly estimate any drift in M2 velocity. While they find the coefficient to be significant, the drift is negligible at around .003 percent per year.

difference between their yield and the yield of a Treasury bill. The opportunity cost of M2 then is the weighted average of the opportunity costs of each M2 component, where the weights are equal to the component's share of M2.

The response of money demand to changes in market interest rates in this model requires a specification of the relationship of deposit rates to the market rates.[12] Thus, the interest elasticity of money demand now depends on how rapidly banks adjust their deposit rates in response to changing market rates. To illustrate, consider the extreme case where deposit rates respond instantaneously to changes in market rates so as to maintain a constant spread between them. In such a case, money demand and velocity would be unaffected by changes in market interest rates because the opportunity cost of money would not change.

If, on the other hand, deposit rates adjust instantaneously but only partially to a change in interest rates (that is, not point-for-point), then the interest elasticity would be proportional but less than the opportunity cost elasticity. Any trend in interest rates would also be associated with a trend in the opportunity cost of those deposits. Equilibrium money demand would hence be affected, and the trend in velocity would be proportional to the trend in the opportunity cost of M2.

Finally, consider a case where deposit rates respond sluggishly to changes in open market rates. A permanent increase in market interest rates would initially be associated with an increase in opportunity cost, as market rates moved above deposit rates, followed by a decrease as deposit rates caught up. If the deposit rates ultimately adjusted point-for-point, the long-run equilibrium level of opportunity cost would be unaffected.

Moore et al. specify deposit-rate equations to be simple linear functions of the federal funds rate. They assume that competitive forces ultimately drive the slope coefficients to equal one minus the marginal reserve ratio, and the intercept to equal some negative value to reflect transactions costs that are not recovered as fees assessed to the depositor. As with M2 demand, the short run is formulated within an error-correction framework. Changes in deposit rates are assumed to be related to deviations of the rates from their long-run equilibrium values, and to changes in the current and past values of interest rates.

Moore et al. find that for many components of M2, own rates have been relatively slow to adjust. This is particularly evident for instruments with transactions features such as NOW accounts and, to a lesser extent, money market deposit accounts. On the other hand, some deposit rates, such as those on time deposits, have adjusted relatively quickly and fully to changes in market rates.[13] However, because a significant share of M2 deposit rates adjust sluggishly, changes in market interest rates have substantial short-run effects on the opportunity cost of M2, and consequently on its demand.

Indeed, the model estimated by Moore et al. suggests that the *short-run* interest elasticity of M2 demand is substantial. What is curious is that some bank deposits appear more interest sensitive than before deregulation. One might expect just the opposite, as deregulation allows banks to compete more effectively for funds, even if they adjust only slowly.

Some analysts have speculated that the increased sensitivity of some deposits may reflect the increased sophistication of most deposit-holders and the improved communications technologies that have made funds transfers more convenient. Even if opportunity costs are less affected by changes in interest rates now than before, deposit-holders are much more aware of alternative assets and therefore are more likely to respond to changes in the opportunity cost of some deposits.[14]

V. M2 Velocity

The treatment of opportunity cost as distinct from the market interest rate helps to reconcile why M2 velocity is trendless despite the observed trends in interest rates. This is easiest to understand in the case where deposit rates ultimately adjust point-for-point with changes in market rates. In such a case, opportunity cost is by definition stationary around some trendless differential, and hence would be independent of any trend in interest rates. Thus, the velocity of these deposits would be insulated from changing inflationary expectations.

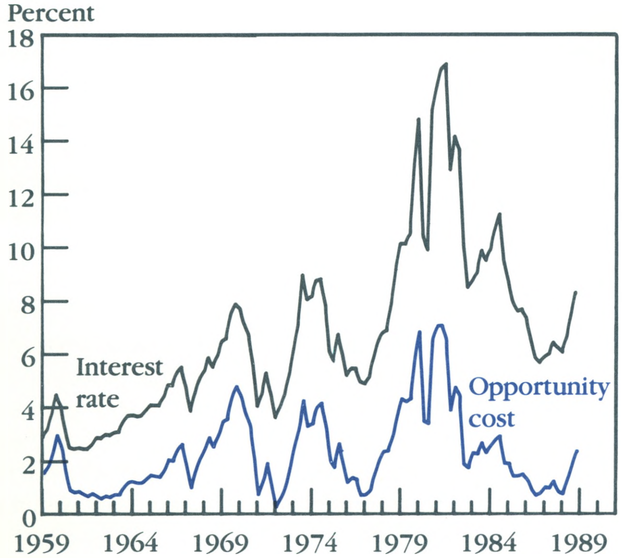
■ 12 The advantages of measuring opportunity cost as a differential in yields are in principle greater since deregulation than before. Currently, there are no interest-rate ceilings on any of M2's noncurrency and non-demand-deposit components, which are 83 percent of the total.

■ 13 Moore et al. also conclude that deposit-rate adjustments are asymmetric, adjusting more rapidly to upward movements in market rates than to downward movements.

■ 14 However, there appears to be no shift in the opportunity cost elasticity of the M2 aggregate after deregulation.

FIGURE 2

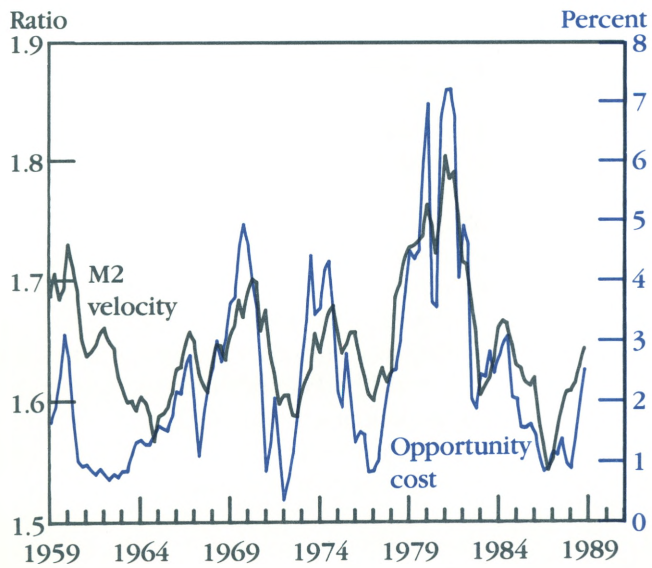
Interest Rate and Opportunity Cost of M2



SOURCE: Board of Governors of the Federal Reserve System.

FIGURE 3

M2 Velocity and Opportunity Cost



SOURCE: Board of Governors of the Federal Reserve System.

However, not all deposits in M2 adjust point-for-point to changes in interest rates. Reserve requirements assure some wedge preventing complete adjustment. Also, since currency pays no explicit yield, its opportunity cost is essentially equal to the interest rate. Thus, if the level of interest rates exhibits drift, the opportunity costs of these components of M2 will also exhibit drift in the same direction. M2 velocity would not be independent of the level of interest rates.

In practice, however, the drift in the opportunity cost of M2 has been highly muted relative to the drift in interest rates (see figure 2). The wedge created by reserve requirements is in fact small—12 percent or less. Moreover, the share of currency and reservable deposits amounts to less than 20 percent of M2; thus, the nonstationary component of the opportunity cost would be small and perhaps negligible. Interest-rate trends, then, would not affect M2 velocity substantially in the long run.

Some evidence indicates that M2 velocity is, in the long run, independent of interest rates. Engle and Granger (1987) conclude that nominal income and M2 are cointegrated, implying that M2 velocity is a stationary process and hence is unaffected by interest-rate trends. Thus, it would appear that M2 velocity is immune to changing inflationary expectations in the long run. This explains why the M2 velocity trend, unlike that of M1, was unaffected by the rise and fall of inflation in the postwar period. In the short run, however, changes in the opportunity cost of M2 are driven largely by changes in market interest rates; and, as figure 3 illustrates, M2 velocity is quite closely related to the opportunity cost of M2.

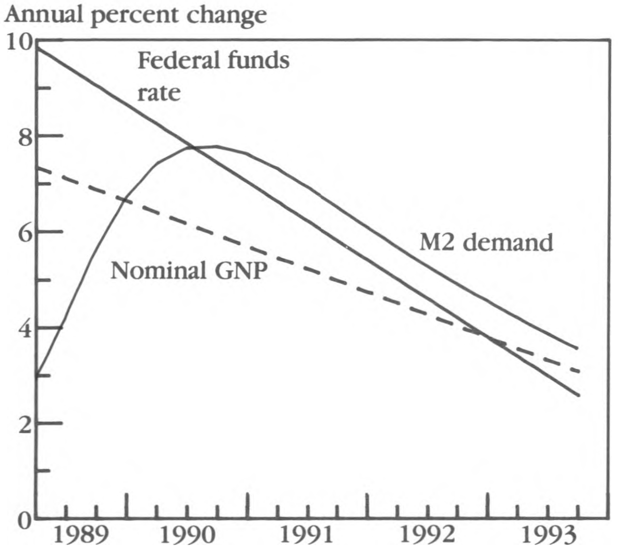
VI. Money as a Policy Guide During Disinflation

Recent evidence indicating that money demand is substantially interest sensitive has important implications for monetary policy. Interest sensitivity of money demand poses serious problems for policies that seek to achieve disinflation. Poole (1988) concludes, "There is a serious and probably insurmountable problem to designing a predetermined money growth path to reduce inflation..." (p. 97).

Poole offers a clear description of the problem: If policymakers embark on a credible policy of disinflation, they should expect that nominal interest rates will ultimately fall as inflationary expectations subside. Consequently, they should expect velocity growth to decline, and perhaps even become negative, if the policy becomes successful. Under these circumstances, inflation

FIGURE 4

Hypothetical M2 Demand: Credible Disinflation



SOURCE: Author's calculations.

could be reduced without a decline in money growth, at least initially. Indeed, a decline in money growth might have a significant depressing effect on the economy. He concludes that the gradualist prescription of predetermined reductions in money growth would not be politically sustainable, as it would likely be associated with unnecessary weakness in economic activity.

Poole further argues that this situation poses a serious dilemma for policymakers. How do they convince markets of their commitment to disinflation without a reduction in money growth rates? Is it not irrational to bet on lower inflation on the basis of a central bank's promises, with no evidence that the central bank is reducing money growth? Poole concludes that a recession may be necessary to convince markets that the central bank is committed to a disinflationary policy.

The problem of targeting money is easy to appreciate in the context of M1. After all, few analysts anticipated the magnitude of the shift in the drift of M1 velocity. Another reduction in inflation would likely result in another shift in the trend in M1 velocity. Moreover, no specification for short-run M1 demand seems acceptably stable at present. On the other hand, there is no evidence that the trend of M2 velocity has been affected by the transition to lower inflation in the 1980s. The recent specification by Moore et al. suggests that the short-run demand for M2 may be reasonably stable.

A hypothetical example illustrates how the problem applies to a disinflation policy specified as a target path for M2. First, assume that on the basis of a promise alone, markets could be convinced of a central bank's commitment to gradual disinflation from current levels to zero inflation in 1993. To the extent that disinflation was perfectly anticipated, we might expect that nominal magnitudes such as interest rates, personal consumption expenditure growth, and nominal GNP growth would decline smoothly to noninflationary trend paths.[15]

If the parameters of the M2 demand function estimated by Moore et al. are approximately structural, then we would expect M2 demand to accelerate initially to growth rates above the equilibrium rate of nominal GNP growth and then begin to slow (see figure 4). The additional money growth would not be for the purpose of financing future spending, but would reflect a pure portfolio decision to hold a greater proportion of wealth as bank deposits in response to a sharply falling opportunity cost; hence, the monetary acceleration could still be associated with a slowing in nominal spending.

The pattern of M2 growth reflects two key features of the M2 demand model. First, own rates on deposits adjust slowly enough to changes in market rates that the opportunity cost in the short run is directly related to changes in the level of interest rates.[16] Second, M2 demand is substantially sensitive to changes in opportunity cost. Thus, as interest rates fall with disinflation, so does the opportunity cost of M2. It is this decline in M2's opportunity cost that induces investors to hold additional bank deposits relative to their spending needs.

This example is hypothetical, of course. If markets were to maintain an expectation of gradual disinflation, they would need to understand the consequences of a falling opportunity cost and have confidence that the estimated *short-run* M2 demand function was reliable. Only then might markets reconcile an accelerating money-growth path with a disinflation policy.

■ **15** We assume here that in noninflationary equilibrium, growth in nominal GNP and personal consumption expenditures equals 3 percent, as does the Treasury bill rate, but that the federal funds rate equals 2½ percent.

■ **16** This, of course, presumes that banks have a rational basis for adjusting some deposits more sluggishly than others. Thus, although market interest rates fully anticipate disinflation, bank deposits would respond with some delay.

The 22-year estimation period for M2 demand is relatively short, however, and it is not evident that deposit-rate pricing has stabilized since deregulation. It would seem doubtful that markets could be convinced of such a strategy.

Nevertheless, the evidence of substantial interest sensitivity of velocity in the short run suggests that policymakers might sometimes prefer to accommodate the effects of interest-rate changes on money demand. During periods of disinflation, one might then expect wide swings in money growth. Once a disinflation strategy becomes credible, velocity could fall substantially, if only temporarily, and it would be appropriate for policymakers to accommodate the consequent surge in money demand.

VII. Interest Sensitivity and Monetary Rules

Apart from the problems that arise during disinflation, the evidence that M2 is more interest-rate sensitive than previously thought raises some interesting issues concerning monetary rules. On the one hand, shocks to money demand would have smaller real consequences under a constant-money-growth rule than previously thought. Consider a positive shock to money demand. Given an inelastic money supply, interest rates would need to rise and output would need to fall. In conventional macroeconomic models, interest rates would respond initially. Higher interest rates would, in turn, tend to slow economic activity. When the interest elasticity of money demand is high, smaller interest-rate changes are required to offset demand shocks, implying smaller adjustments in output.

On the other hand, the consequences of nonmonetary shocks under a constant-money-growth rule are less clear when the demand for money (and hence velocity) is highly interest-elastic. This longstanding issue is illustrated simply in a debate between Johnson (1965) and Friedman (1966). Johnson argued that interest-sensitive money demand militated against a constant-monetary-growth rule "...because variations in interest rates generated by the real sector would make such a rule automatically destabilizing..." (p. 397). Implicitly, Johnson assumed that variations in interest rates would be a natural by-product of stable output growth; in turn, these variations would cause procyclical variations in velocity, which, under the assumption of constant money growth, would produce fluctuations in the rate of nominal income growth.

Friedman acknowledged this potential outcome, but argued that the conditions assumed by Johnson were highly special. Essentially, Friedman contended that while velocity would tend to move with nominal output, a constant-money-growth rule would nevertheless dampen output fluctuations relative to "discretionary" policies. Thus, Friedman was not comparing his rule to an ideal rule, but to the existing practice of the central bank.

It is useful to separate this debate into two issues. The first is the general issue of rules versus discretion. The second is the question of whether monetary rules (or targets) should allow for some kind of systematic (that is, automatic) feedback to account for interest-rate changes and, hence, shifts in velocity. More specifically, should a rule or targeting procedure anticipate changes in interest rates? This first issue is only indirectly relevant to the question of interest-rate sensitivity and therefore is not dealt with here.[17] The question of feedback, on the other hand, is relevant whether a policy admits some discretion or not.

The feedback issue depends on the kinds of shocks that occur and on the poorly understood dynamics of adjustment in the economy. Specifically, it depends on where shocks arise in the economy, what their relative magnitudes are, and how they are propagated through the economy. The answers to these questions depend on the particular model one believes is appropriate for characterizing the economy. Unfortunately, no consensus exists or even seems imminent.

One large and influential class of empirical models, sharing a common propagation mechanism, casts some doubt on the efficacy of *constant* monetary-growth rules. In these models, the inflation process is characterized by an output-gap accelerationist mechanism:

$$\dot{p}_t - \dot{p}_{t-1} = \alpha_0 + \alpha_1 (q_t - q_t^*) + \alpha z_t ,$$

where $\dot{p}$ is the inflation rate, q is the level of output, q^* is full-employment output, and z represents other factors. If z is constant, a change in the inflation rate depends on the output gap. When output exceeds full-employment output (that is, when unemployment is below its natural rate), inflation accelerates. When output is below full-employment output, inflation decelerates. Anderson and Enzler (1987) explain the consequences of such a mechanism for a monetary rule:

■ **17** For a discussion of the general issue of rules versus discretion, see Carlson (1988).

> It is easy to see why holding the money growth rate constant might not result in a stable simulation path for a macromodel containing this mechanism. The fixed money growth path predetermines both the rate of inflation and the price level consistent with the economy's steady-state path at each point of time. Consider what happens if the price level is disturbed upward from the steady-state growth path. The demand for money is increased and interest rates rise. This depresses output and increases unemployment. The increased unemployment, in turn, depresses the rate of change of prices. As long as the price *level* remains too high, a force is created that tends to keep unemployment above its natural rate and the *rate of inflation* continues to fall. The declining rate of inflation eventually returns the price level to its steady-state value, and this in turn allows the unemployment rate to return to the natural rate, but at this point *inflation* is too low to be consistent with the fixed money growth path and the price level falls through the steady-state level. This reduces the demand for money, causing interest rates to fall until unemployment is below the natural rate. Inflation then accelerates until at some point it reaches its steady-state value. But now the *level* of prices is too low. The mirror image of the previous events takes place and overshooting occurs again. (p. 297)

While the estimated parameters of these models suggest that the cycle described above eventually converges, the process is generally only slightly dampened.[18]

Because the estimated interest elasticity of output in these models is typically relatively small, it is likely that a higher interest elasticity of money demand would only attenuate the cycles of such models. To illustrate this point, consider again the propagation of the upward price disturbance. The higher the interest elasticity of money demand, the lower the rise in the level of the interest rate that would result as an effect of the price shock on money demanded, given an inelastic supply. However, because the interest-rate elasticity of output is low, the consequent effect on output would be even smaller, and would hence slow the process that dampens the shock to inflation.[19]

Evidence of a potential for long macroeconomic cycles is not a unique consequence for models with an output-gap mechanism. Indeed, some simple models linking money and prices also exhibit long cycles. One example is a recent single-equation model estimated by Hallman, Porter, and Small (1989). Theirs is a reduced-form model of the relationship between inflation and M2 that does not explicitly include either the current level of output or employment as a variable.[20] While they find rather lengthy adjustments to simulated shocks (for example, more than 100 years), the cycles of their model are more damped than those of many large macroeconomic models.

From a deterministic point of view, the Hallman et al. results suggest that there is a *nonconstant* money-growth path consistent with a relatively smooth *transition* to equilibrium. As they note, inflation, *in equilibrium,* could be controlled at any constant rate with constant growth of M2.

Notwithstanding the well-known critique of Lucas (1976), the use of deterministic simulations as evidence in the debate about an appropriate policy rule is of only limited value. A critical issue in this debate is how a rule performs in a stochastic framework, one that approximates the *distribution* of disturbances that have historically affected the various sectors of the economy. In this context, the issue is not the selection of an appropriate policy response to a particular shock, but the robustness of a contractual commitment to a policy rule in responding to a *series* of likely outcomes arising from a typical distribution.

One sense of robustness has been stressed by McCallum (1988): that a rule perform well for a variety of models, preferably ones incorporating alternative views of macroeconomic relationships. It is important to establish robustness (in this sense) because no structural model of the economy enjoys sufficiently wide acceptance; nor does any consensus seem to be evolving. Thus, to gain acceptance for a proposed rule, the rule advocate must demonstrate that the rule would lead to reasonably good outcomes for variables of interest *and* for a variety of models.[21]

■ **18** It should be noted that these models typically do not result in a trade-off between inflation and unemployment in the long run.

■ **19** It is perhaps ironic that these models suggest that a constant-money-growth rule would result in an interest-rate path that is too smooth to substantially dampen shocks to inflation over reasonably short horizons. Indeed, these models suggest that rather large and sustained increases in interest rates would be required to substantially affect the output gap and hence the inflation rate. However, it is uncommon to find antagonists of the money-growth rule who cite this evidence and also publicly advocate the kind of interest-rate variation that large models suggest is required to stabilize the inflation rate.

■ **20** Nevertheless, the model incorporates estimates of full employment output and equilibrium velocity as determinants of the equilibrium price level. In this model, inflation is a function of the gap between the current price level and its equilibrium level.

■ **21** One method of simulation designed to address this issue is suggested by Tinsley and von zur Muehlen (1983). They essentially offer a technique to generate unplanned disturbances consistent with the error structure observed in historical experience. The robustness of a policy rule is tested by multiple simulations of the performance of the rule over multiyear periods, where each simulation draws a different series or "history" of unplanned disturbances. The horizons are chosen to be long enough to allow significant differences to emerge among the alternative policies and to assure that policies ultimately stabilize outcomes.

The sum of simulation results provides distributions of outcomes for each of the model's variables. For instance, one policy may be associated with a wide

Stochastic simulations, however, are costly to obtain. Moreover, a test for robustness is an open-ended search, encompassing an endless variety of both rules and models. As a consequence, evidence from this analysis is in only an embryonic state. Preliminary results by Tinsley and von zur Muehlen (1983) and Anderson and Enzler (1987) suggest, however, that monetary rules do not perform as well as alternative rules or intermediate targeting procedures. Nevertheless, the monetary rules and targeting procedures examined were based on older, less interest-sensitive estimates of money demand.

The ongoing debate over the efficacy of a constant-money-growth rule, when the interest elasticity of money demand is large, is not likely to be resolved without some convincing empirical basis. Thus, it would seem appropriate for policymakers to take account of the consequences of expected interest-rate changes on velocity when choosing target ranges for M2 over a period of a year or less. That is, it may be appropriate for M2 growth to slow substantially when interest rates are rising and expected to rise further, or to accelerate substantially when interest rates fall.

VIII. Concluding Comments

One common finding of recent empirical research in monetary economics is that the interest elasticity of money demand is estimated to be substantial, and higher than many economists previously thought. The evidence seems strongest for M1 demand in the long run. While interest rates have little long-term effect on M2, the short-run elasticity seems to be greater than previously thought.

When the interest elasticity of money demand is high, velocity can vary widely. This creates a problem for using money as a policy guide. Monetary targets should take into account the consequences of expected changes in interest rates on money demand. This problem is perhaps most difficult during periods of disinflation, when changing expectations about inflation result in large swings in interest rates and hence in velocity.

Finding that a money-demand function is stable is not a sufficient basis for adopting a constant-money-growth rule. The rule advocate has the burden of convincing others that the stabilizing effects of the monetary rule would outweigh the potentially destabilizing effects of maintaining constant money growth when velocity varies systematically with interest rates. Because no consensus exists about the best model for the economy, the rule advocate must argue his case in the context of a variety of models.

The challenge of examining rule robustness has been recognized and addressed by McCallum (1988). It is hoped that others will follow his lead. Recent developments in simulation methods offer promising approaches for examining the robustness of alternative policy rules.

range of outcomes for output and interest rates, but with a small range for prices and money for any given simulation horizon. Another policy may be associated with small ranges for interest rates and money, but with large ranges for prices and output, or vice versa. Tinsley and von zur Muehlen note, "...the essential contribution of stochastic simulation analysis is the empirical premise that while individual unplanned disturbances cannot be predicted (by definition), their ranges of probable outcomes are unlikely to differ significantly from the dispersions observed in historical experience ..." (p.16).

References

Anderson, Robert and Enzler, Jared J., "Toward Realistic Policy Design: Policy Reaction Functions that Rely on Economic Forecasts," in Rudiger Dornbusch, Stanley Fischer, and John Bossons, eds., *Macroeconomics and Finance: Essays in Honor of Franco Modigliani,* Cambridge, Mass.: MIT Press, 1987, 291–330.

Baumol, William J., "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quarterly Journal of Economics,* November 1952, *66,* 545–56.

Carlson, John B., "Rules Versus Discretion: Making a Monetary Rule Operational," *Economic Review,* Federal Reserve Bank of Cleveland, Quarter 3, 1988, 2–13.

Chow, Gregory C., "On the Long-Run and Short-Run Demand for Money," *Journal of Political Economy,* April 1966, *74,* 111–31.

Engle, Robert F. and Granger, C.W.J., "Cointegration and Error Correction: Representation, Estimation, and Testing," *Econometrica,* March 1987, *55,* 251–76.

Friedman, Milton, "The Quantity Theory of Money—A Restatement," in Milton Friedman, ed., *Studies in the Quantity Theory of Money,* Chicago: University of Chicago Press, 1956, 3–21.

_______ , "Interest Rates and the Demand for Money," *Journal of Law and Economics,* 1966, 71–85.

Goldfeld, Steven M., "The Demand for Money Revisited," in Arthur M. Okun and George L. Perry, eds., *Brookings Papers on Economic Activity,* 1973, *3,* 577–638.

Hallman, Jeffrey J., Porter, Richard D., and Small, David H., "M2 per Unit of Potential GNP as an Anchor for the Price Level," *Staff Study No. 157,* Board of Governors of the Federal Reserve System, April 1989.

Havrilesky, Thomas and Boorman, John, *Monetary Macroeconomics,* Arlington Heights, Ill.: AMH Publishing Co., 1978.

Hetzel, Robert L. and Mehra, Yesh, "The Behavior of Money Demand in the 1980s," mimeo, October 1987.

Hoffman, Dennis and Rasche, Robert H., "Long-Run Income and Interest Elasticities of Money Demand in the United States," National Bureau of Economic Research Working Paper No. 2949, April 1989.

Johnson, Harry G., "A Quantity Theorist's Monetary History of the United States," *Economic Journal,* June 1965, *75,* 388–96.

Laidler, David, "The Rate of Interest and the Demand for Money—Some Empirical Evidence," *Journal of Political Economy,* December 1966, *74,* 545–55.

_______ , "The Buffer Stock Notion in Monetary Economics," *Economic Journal,* Supplement, 1984, *94,* 17–33.

Lucas, Robert E., "Econometric Policy Evaluation: A Critique," *Journal of Monetary Economics,* Supplementary Series 1, 1976, 19–46.

McCallum, Bennett T., "Robustness Properties of a Rule for Monetary Policy," *Carnegie-Rochester Conference Series on Public Policy,* Carnegie Mellon University and National Bureau of Economic Research, February 1988.

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

Moore, George R., Porter, Richard D., and Small, David H., "Modeling the Disaggregated Demands for M2 and M1 in the 1980s: The U.S. Experience," mimeo, Board of Governors of the Federal Reserve System, May 1988.

Poole, William, "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quarterly Journal of Economics,* May 1970, *84,* 197–216.

_______ , "Monetary Policy Lessons of Recent Inflation and Disinflation," *The Journal of Economic Perspectives,* Summer 1988, *2,* 73–100.

Rasche, Robert H., "Some Evidence on the Elusive 1982 Shift in Velocity Drift," Shadow Open Market Committee Policy Statement and Position Papers, No. 89-01, 1989, 37–40.

Tinsley, P.A. and von zur Muehlen, P., "The Reliability of Alternative Intermediate Targets," mimeo, Board of Governors of the Federal Reserve System, November 8, 1983.

Tobin, James, "Liquidity Preference as Behavior Towards Risk," *Review of Economic Studies,* 1958, *25,* 65–86.

Accounting for the Recent Divergence in Regional Wage Differentials

by Randall W. Eberts

Randall W. Eberts is an assistant vice president and economist at the Federal Reserve Bank of Cleveland. Comments and suggestions by Brian Cromwell, Erica Groshen, and Sharon Smith, and computer assistance by Ralph Day and John Swinton, are gratefully acknowledged.

Introduction

Convergence of regional income differentials is commonly perceived as the natural result of the gradual development and maturation of regional economies. One expects that factors such as improved transportation and communication, enhanced mobility of capital and labor, and the shift away from resource-based activities would lead regions, and their incomes, to look more and more alike. Indeed, since the 1880s, the general trend has been toward convergence of regional per capita income in the United States.

Recently, this trend appears to be reversing. Browne (1989) shows that since 1979, regional disparities in per capita personal income have been on the rise. Furthermore, she concludes, "...the key to both the converging per capita incomes of the 1970s and the diverging incomes of the 1980s was changes in industry earnings" (p. 38).

According to Nourse (1968), regional income divergence has happened only once in the last century, between 1920 and 1940. After 1940, regional incomes returned to their longer-run path of convergence. Easterlin (1958) concluded from that 20-year disturbance in the longer-run trend that "...it is by no means certain that convergence of regional income levels is an inevitable outcome of the process of development. For while migration and trade do appear to exert significant pressure towards convergence, they operate within such a rapidly changing environment that dynamic factors may possibly offset their influence" (p. 325). It appears that the conclusion Easterlin drew 30 years ago may be relevant in today's situation.

This recent deviation from the general tendency toward convergence raises several questions. Why the relatively sudden shift in the direction of regional income differentials after so many years of convergence? What are the sources of this change in regional per capita income? Have the fundamental forces that shape the nation's economy changed direction during the 1980s, or is this merely a temporary digression from the longer-run trend of convergence?

This paper begins with the observation by Browne that earnings account for most of the shift from income convergence to income divergence among regions. We identify two basic sources of regional wage differentials and examine which of them is more responsible for the shift in wage patterns. The two sources are 1) regional differences in the return on various worker attributes and the wage differentials among industries and occupations, and 2) regional differences in the level of worker attributes and the distribution of workers among industries and occupations.

These two sources can be distinguished by asking whether earnings per worker differ among regions because of differences in the attributes of workers, or because of differences in the value of worker attributes as determined by the regional labor markets. Explaining convergence or divergence of regional wages, therefore, rests with the ability to explain convergence or divergence of characteristic prices, levels of characteristics, or both.

Several studies have explored the relative size of these two components of wage differentials between regions, primarily in an attempt to explain the difference in wages between the South and other regions of the country. Sahling and Smith (1983) were among the first to look at the wages and attributes of individual workers to examine regional wage differentials over time. They compared the South with four other regions in the country: the Northeast, the North Central, the New York metropolitan area, and the West. They estimated separate real and nominal wage equations using a sample of residents from 29 of the largest standard metropolitan statistical areas (SMSAs) found in these five regions. The worker-attribute variables included measures of schooling, experience, race, occupation, sex, industry, job status, and union membership. Using two cross sections of data, from the May 1973 and May 1978 Current Population Surveys, they concluded that much of the variation in wages between the South and the other regions examined is a result of substantial variation in the real and nominal rates of return to worker characteristics.

Farber and Newman (1987) extended Sahling and Smith's analysis to look explicitly at changes in characteristic prices over time. In addition to looking at regional wage differentials in two different years, 1973 and 1979, they estimated the changes in the differentials between the two years for various pairs of regions. They found that more than half of the predicted changes in South/non-South wage ratios can be accounted for by changing relative returns to worker characteristics between the two areas (p. 223).

Other studies, using similar techniques and micro-level data, do not necessarily agree with the conclusion that characteristic prices account for regional wage differentials. Bellante (1979) and Gerking and Weirick (1983) find that regional wage differences are due primarily to differences in the levels of worker characteristics. These results leave open the possibility that both prices and levels are likely sources of regional wage differentials.[1]

This paper extends Farber and Newman's work in two directions. First, it includes three time periods in order to examine the sources of the switch in wage patterns that apparently occurred at the beginning of the 1980s. Each time period is constructed by pooling three years of data: the first period includes the years 1973-75, the second includes 1979-81, and the third includes 1985-87. The interval between the first and second periods is characterized by regional wage convergence, as documented by Farber and Newman (1987) and Browne (1989). The interval between the second and third periods exhibits regional wage divergence, as shown by Browne. The second direction is to look at all nine U.S. regions as defined by the U.S. Bureau of the Census, relative to the national average, instead of comparing pairs of selected regions.

However, unlike the studies by Sahling and Smith and Farber and Newman, which were concerned with comparing wage differentials across different regions, our purpose is to see whether the structure that caused a particular region to converge toward the national average during the early periods can also account for the divergence of wages in that same region during the latter periods. Therefore, it appears that using nominal wages is sufficient for an initial look at the sources of the shift in wage patterns.[2]

I. Explanations of Regional Wage Differentials

One of the longstanding tenets of economics is that efficient markets result in equal prices across regions. Indeed, economists have observed for decades the slow convergence of average wages among the regions of the United States, where goods and factors can flow freely. How, then, can one explain the apparent divergence of wages in recent years?

■ **1** Dickie and Gerking (1988) provide a very comprehensive and insightful critique of the literature.

■ **2** Work by Roback (1982) and Beeson and Eberts (1989) shows that considering nominal wages can be viewed as only a partial-equilibrium analysis. Household spatial equilibrium includes not only wages, but also the price of housing and nontraded local goods. Therefore, focusing only on nominal wages may introduce estimation bias, especially in the prices of worker characteristics, for regions in which housing and other local-goods prices have changed significantly from the national trend.

International trade theory offers useful insights into conditions that lead to regional wage convergence and divergence. Much of the relevant literature discusses wage equalization: average wages across regions are equal if both the prices of worker characteristics and the composition of worker characteristics are the same. If the first condition holds, then wages of identical workers will be the same across regions. However, unless the second condition also holds, the average wages of regions will be unequal.

Within a regional context, conditions for equalization of characteristic prices are less stringent than those for equalization of characteristic levels.[3] A well-known theorem in trade theory, the factor-equalization theorem, states that trade in commodities and factor movements are substitutes. According to this theorem, free trade of goods leads to equal factor prices among regions, even when factors of production are immobile. Therefore, within the United States, which does not limit trade between regions, one would expect the unimpaired flow of goods to tend to equalize wages. It has been this line of thinking, based on the notion that regions trade because of differences in factor endowments, that has led to expectations of regional wage convergence.

Several assumptions, which may or may not be met, are necessary to reach this conclusion, however:

a) relative factor endowments are not identical across regions,
b) regions have identical technologies,
c) regions have identical homothetic demand,
d) production is characterized by constant returns to scale,
e) production is characterized by perfect competition, and
f) there are no domestic distortions in either region.

Markusen (1983) demonstrates that the relationship between commodity trade and factor trade varies depending on the specific assumptions that are retained. By relaxing each of the assumptions one at a time, he shows that the initial trading equilibrium is not characterized by factor-price equalization. In each case, factor prices cannot be equalized between regions until at least one region is specialized. He concludes that the notion that trade in goods and factors are substitutes may be a rather special result, which is generally true only when differences in relative factor endowments are the basis for trade and when no market imperfections exist.

Regions may trade goods for reasons other than initial differences in factor endowments. Markusen considers various other bases for trade between regions in which the initial trade equilibrium is not characterized by factor-price equalization. These conditions include

a) differences in production technologies,
b) production taxes,
c) monopolies,
d) external economies of scale (increasing returns to scale), and
e) factor-market distortions.

If these characteristics hold for regions, then factor prices will not be equalized, even though goods may still flow freely among regions. It is easy to envision regional differences in technology, taxes, market share, agglomeration economies, and unions—all of which would satisfy one or more of the above conditions.

Factor-price equalization can be achieved in these less-specialized cases if factors are mobile. Factors will flow to the region with the higher price, until interregional price differentials disappear. When trade is based on factors such as those listed above, factor prices will differ in such a way that the price will be higher for the factor that is used intensively in the production of the export good of that region. Consequently, the region will be relatively well endowed with the factor that is more intensively used in the production of the region's export good. However, factor flows, particularly labor migration, are impeded by imperfect information, by moving costs (both monetary and psychic) and, in the case of labor, by imperfect matches between labor skills and job requirements.

What does this mean for the second component of wage changes—the level or composition of factors? When trade is based on differences in factor endowments, there will be no migration based on wage differentials, for the simple reason that wages will not differ between regions because of interregional trade in goods.[4] When trade is based on differences in production technologies, taxes, or factor-market distortions, factor-price differentials lead to factor flows, but these flows will result in different proportions of factors. Therefore, these models suggest that average wage levels are very unlikely to be the same across regions.[5] Even though interregional

■ 3 Dickie and Gerking (1988) use trade theory to provide a comprehensive assessment of the necessary and sufficient conditions for regional wage equalization.

■ 4 Of course, individuals may find regions to be attractive for reasons other than higher wages. Site-specific amenities may also influence an individual's preferences.

■ 5 Wages will also differ across regions because of compensating differentials for site-specific characteristics, as discussed by Beeson and Eberts (1989).

prices may be equal, as predicted by both models, it is most likely that the composition of the characteristics will differ among regions.

Dickie and Gerking (1988) summarize the outcomes of trade theory as they pertain to interregional wage differentials. First, equalization of labor-characteristic prices does not depend on geographic mobility of the entire labor force. Rather, equalization occurs if enough markets for goods and factors exist and if those markets are allowed to clear. Second, when a combination of commodity trade and factor mobility guarantees factor-price equalization, then relative factor supplies end up unequal and regions tend not to become homogeneous in factor composition. Third, when labor is heterogeneous, economic efficiency, as evidenced by equal factor prices, does not lead to interregional equality of average wages (pp. 10-11).[6]

Therefore, it appears that a systematic change in characteristic prices is a likely source of the switch from regional wage convergence before 1980 to regional wage divergence after 1980. The subsequent analysis estimates the two basic components of regional wage changes and examines which of them contributes more to these observed changes.

II. Accounting for Regional Wage Differentials

Consider the standard hedonic wage equation in which the wage (W_{ij}) of individual i living in region j is a function of the individual's attributes (H_{ij}) and job or workplace characteristics (C_{ij}):

$$(1) \quad W_{ij} = w(H_{it}, C_{ij}).$$

Assuming perfectly operating labor markets, prices of each attribute are determined by supply and demand conditions. Under the assumptions of perfect information, costless spatial labor mobility, and zero transactions costs, characteristic prices will be the same across regions. Consequently, workers with the same characteristics will be paid the same wage regardless of location.

The technique used to account for the two sources of wage differentials follows the approach of Oaxaca (1973), with modifications made by Sahling and Smith (1983) and Farber and Newman (1987). Writing equation (1) in log-linear form, dropping the individual subscript, and adding a time subscript yields

$$(2) \quad lnw_{jt} = b_{jt} X_{jt},$$

where $j = 1,...,R$ regions, and
$t = 1,...,T$ time periods.

The parameter vector b_{jt} represents the characteristic price and vector X_{jt} represents the levels of characteristics, both of which can differ among regions and over time. Using y for lnw, we can write the percentage change in wages between two regions (S and N) during one time period as

$$(3) \quad (y_{St} - y_{Nt}) =$$
$$(b_{St} - b_{Nt})X_{St} + (X_{St} - X_{Nt})b_{Nt}, \quad t = 1,...,T.$$

The first term on the right-hand side accounts for the change in characteristic prices between regions S and N. For our purposes, N denotes the national average. The second term denotes the change in levels of worker characteristics between the two regions.[7] It is clear from equation (3) that wage differences between regions result either from differences in prices or from differences in levels. One can use this framework to assess which of the two components accounts for the larger share of the regional wage difference.[8]

The issue of wage convergence or divergence requires examining how these regional wage differences change over time. For wages to converge toward the national average, the distance between the regional and national wage level must narrow over time. Consequently, if the region starts out with a wage above the national average, convergence requires that the difference, $(y_{St} - y_{Nt}) - (y_{St-1} - y_{Nt-1})$, must be greater than zero. The same relationship must be negative if the region starts out with a wage below the national average. The condition for divergence,

■ 6 Dickie and Gerking also stress a fourth and important point: if data do not adequately distinguish between workers with particular characteristics, then estimated returns will be averages and tests of the interregional wage equality hypothesis would be biased toward rejection.

■ 7 A residual term, $(b_S - b_N)(X_S - X_N)$, is omitted for simplicity. Furthermore, there is an index problem associated with this technique. Changing the base to one region or the other will change the values of the components. Some studies, such as Sahling and Smith (1983), have attempted to avoid the problem by using averages of the two region's characteristic levels or prices. We instead choose to follow the technique of Farber and Newman (1987), which chooses one region as the base. In this way, we are better able to compare our results with theirs.

■ 8 As Farber and Newman point out, the accounting framework relies on the unbiasedness and consistency properties of OLS estimators, and has avoided the pre-test biases of imposing implicit restrictions on coefficients found to be statistically insignificant (p. 219).

obviously, would require the opposite signs.[9]

The relative change over time in regional wage differentials can be divided into several components using a variation of the same accounting scheme adopted in equation (3) for the static case. Following Farber and Newman, one can specify equation (3) for two different time periods (in this case, periods 1 and 2) and then subtract one from the other. This technique yields the following accounting framework:

$$(4)\quad (y_{S2} - y_{N2}) - (y_{S1} - y_{N1}) = [(X_{S2} - X_{N2}) - (X_{S1} - X_{N1})]\,b_{N2} + (X_{S2} - X_{S1})(b_{N2} - b_{S2}) + (X_{S1} - X_{N1})(b_{N2} - b_{N1}) + X_{S1}[(b_{S2} - b_{N2}) - (b_{S1} - b_{N1})].$$

The four components can be interpreted in the following way. The first term, referred to as the main effect, reflects how much of the change in the wage differential is due to changes in the differences of wage-determining characteristics between the two regions, evaluated at the national average characteristic prices. Notice that this term may be zero even when characteristic levels differ between region *S* and the national average in each time period, as long as these differences are not the same in each time period. The second term is the price-interaction term and reflects the effects of absolute changes in characteristics of workers in region *S* over time. The third term is the price-interaction effect, which allows for characteristic prices to change over time. The last component, the region-time interaction effect, represents the possibility that the characteristic prices in the two regions may change over time at different rates.

These four components of regional wage changes provide the basis for identifying the relative contributions of intertemporal changes in characteristic prices and levels to the regional wage differentials. To construct these wage-change components, separate hedonic wage equations are estimated for each region in each time period. For nine census regions and three time periods, this requires 27 separate regressions. The coefficient estimates and the means of the levels of characteristics are then combined according to equation (4).[10]

Comparing changes in regional wages relative to the national average partially adjusts for the general nominal wage increases observed over the 15-year period between 1973 and 1987. However, any deviations of regional price trends from the national average will be imbedded in the various components, particularly in those related to differences in characteristic prices. Instead of relying on the national trends to capture regional price differentials, it would be ideal to adjust regional wages for differences in the cost of living. Unfortunately, regional indexes are available only for metropolitan areas, and even then, there are no current indexes that can be used to compare cost-of-living differences across metropolitan areas.

III. Empirical Results

Data

The data used to estimate the wage differentials are obtained for various years from the Current Population Surveys (CPS) compiled by the U.S. Department of Labor, Bureau of Labor Statistics. The CPS surveys individual workers periodically regarding hours worked, earnings, worker characteristics, employment status, and so forth. Each time period considered in the analysis consists of a pooled sample of three years.[11] The first

■ **9** We have chosen to compare each region with the national average, which we feel provides the most clarity when so many regions are being compared. This approach may introduce two sources of bias, however. The first is because the national sample is not a region separate from the others, but is made up of individuals in each region. The second source arises from the finding that the characteristic prices of each region are significantly different. Consequently, the characteristic prices estimated for the nation may not represent prices for the national market, but rather the average of prices from each distinctly different regional market.

■ **10** One drawback of this approach, as discussed by Dickie and Gerking (1988), is the lack of a confidence interval estimate around these various components, leaving it unclear how the results generalize to the population.

■ **11** Various features of the CPS files have changed over the years covered in this analysis, which introduces several problems when using these data to derive a consistent time series of regional wages. First, the method of collecting wage and worker characteristics has changed. For the years 1973-78, questions regarding worker wages and characteristics were asked only in the month of May. This poses two problems. First, the sample contains only those individuals who were in the second rotation, which, in addition to being less representative, reduces the number of respondents. Second, annual wage estimates will reflect wages obtained for only one month of the year.

Starting in 1979, the wage questions were asked of one-quarter of the individuals in each of the 12 monthly surveys conducted each year. Because of the difference in the way in which information is gathered, the total number of workers with sufficiently complete records for analysis is much smaller before 1979 than afterward. Pooling the individual years will ameliorate these problems to some extent.

TABLE 1

Regression Estimates for the National Sample of Workers

Variables	1973-75 Mean	1973-75 Coefficient	1979-81 Mean	1979-81 Coefficient	1985-87 Mean	1985-87 Coefficient
Full-time (= 1)	.80	.101	.81	.141	.80	.187
Race (nonwhite = 1)	.10	-.040	.12	-.032	.13	-.048
Sex (female = 1)	.42	-.192	.46	-.168	.48	-.133
Experience	18.72	.026	17.76	.026	17.59	.029
Experience squared	570.60	-.0004	521.57	-.0004	488.80	-.0004
Schooling	1.82	.130	1.98	.131	2.09	.162
Schooling squared	4.03	.007	4.57	.002	5.00	-.001
(exp) x (sex)	7.77	-.004	7.90	-.004	8.32	-.005
Non-SMSA (= 1)	.30	-.159	.42	-.083	.27	-.133
(13 occupation dummy variables)						
(12 industry dummy variables)						
Dependent variable:						
log (earnings/hours)	1.29		1.74		2.02	
R^2	.91		.95		.96	
Number of observations	116,298		554,864		491,510	

NOTE: All coefficients are statistically significant at the 99 percent level.
SOURCE: Estimates are derived from the Current Population Surveys. See text for details.

period combines the responses from the May survey for the years 1973, 1974, and 1975. The second period pools responses from one-quarter of the individuals in each of the 12 monthly surveys for the years 1979, 1980, and 1981. The third period is derived similarly, except that it includes the years 1985, 1986, and 1987.

These time periods were chosen because they correspond to the switch from regional wage convergence to regional wage divergence as documented by Browne (1989). In addition, years were pooled in order that each region contained enough workers to ensure reliable estimates. The size of the samples ranges from 7,203 workers for the New England census region in 1973-75 to 84,641 workers for the East North Central region in 1979-81.

Following the human-capital specification of Hanoch (1967) and Mincer (1974), individual wages (expressed in logarithms) are specified as a function of various worker attributes. We include education level (entered as a quadratic), potential experience (age, minus years of education, minus six, also entered as a quadratic), and the interaction between experience and female. We also include binary dummy variables indicating whether or not the worker is a full-time employee, female, and nonwhite. Dummy variables are also used to denote a worker's occupation, the industry in which he or she is employed, and whether the worker resides in an SMSA. Hourly earnings were computed by dividing average weekly earnings by average weekly hours.[12]

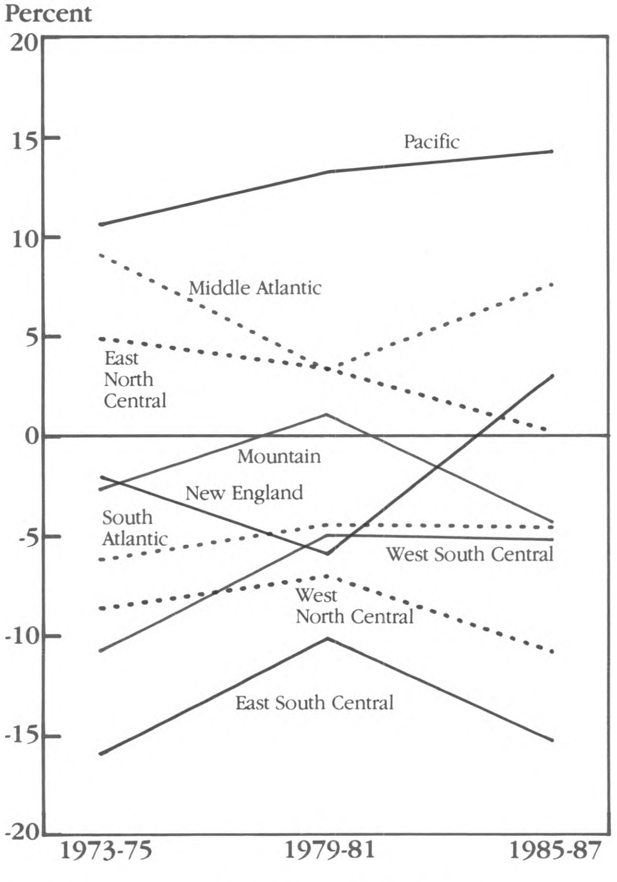
Including the industry-dummy variables is somewhat inconsistent with the notion that the human-capital specification captures supply-side aspects of the labor market. These variables are included, as they have been in other studies, to test the popular notion that industrial restructuring is a primary source of regional wage changes. The changing composition of union membership has also been offered as an explanation for regional wage changes.[13] Unfortunately, the CPS did not ask about union affiliation in the 1979-81 surveys.

■ **12** An interesting extension of the analysis would be to estimate separate regressions for males and females and for whites and nonwhites. Sahling and Smith (1983) found differences in wages between males and females in the South compared with other regions. Changing norms for women and minorities in the workplace may lead to regional differences in the characteristic prices of these groups.

■ **13** However, Farber and Newman (1987) conclude that while unionization is an important contributor to the change in the wage differential attributable to changes in regional differences in worker characteristics, it is not an important variable in explaining changes in wage ratios between regions (p. 222).

FIGURE 1

Regional Nominal Wage Differentials Relative to the National Average



SOURCE: Author's calculations from Current Population Surveys.

Regression Estimates

Separate estimates were obtained for each of the nine census regions for each time period using ordinary least squares. F-tests were performed to test the null hypothesis that the coefficients for each region are equal to the coefficients for the national sample. The null hypothesis was rejected at the 1 percent confidence level for each time period. Even though coefficients differ among regions, estimates from the national sample are displayed and discussed in order to provide an overall perspective of the results. As shown in table 1, all worker-characteristic variables are statistically significant at the 1 percent level and enter with the expected signs. Full-time workers (who work 35 hours or more a week) receive higher wages than part-time workers, everything else being the same. The full-time wage premium has risen from 10 percent in the first period to 19 percent in the most recent period. This fairly sizable increase has occurred even though the percentage of full-time workers in the sample has remained constant.

The nonwhite wage gap appears to have narrowed slightly from 4 percent in 1973-75 to 3.2 percent in 1979-81. However, since that time, the gap has widened, increasing to 4.8 percent in 1985-87. The female wage gap, on the other hand, has steadily narrowed, from 19.2 percent in the first period to 13.3 percent in the most recent period. The wage premium placed on additional hours of work experience has risen steadily for both men and women over the three time periods. Taking into account the interaction terms and evaluating at the mean level of experience, the elasticity of wages with respect to experience for men, for example, rose from 20.6 percent in 1973-75 to 26.4 percent in 1985-87. The net effect of schooling on wages fell between the first two periods and then rose in the third period.

Patterns of Regional Wage Differentials

Nominal earnings estimates, using the CPS sample of workers within nine census regions, reveal a pattern of regional wage convergence followed by divergence, similar to that found by Browne. Figure 1 shows the pattern of regional nominal wage changes relative to the national average. Nominal wages in all regions, except the New England and the Pacific regions, converged toward the national average between 1973-75 and 1979-81.

The standard deviation of the relative wage differentials fell from 0.086 to 0.068 during this period. Wages of workers in the Pacific region increased 2.6 percentage points faster than the national average between the first two periods, which raised the region's wage premium to 13.3 percent. New England, on the other hand, started out below the national average in 1973-75 and continued to lose even more ground by 1979-81, falling from 2.1 percent to 5.9 percent below the national average over this time span.

Between 1979-81 and 1985-87, wages in most of the regions diverged from the national average. The two exceptions were the New England and East North Central regions. Wages in the New England region jumped dramatically during this period, outpacing the national average by 9.1 percentage points. This spurt in wage growth closed New England's wage gap from the previous period and placed its wages 3.1 percent above the national average in 1985-87. Wages in the East North Central region also came closer to the national average, but this was achieved by growing slower than the nation by 3.0 percentage points.

Of the seven regions in which wages diverged from the national average between 1979-81 and 1985-87, five were below the national average. The two regions that lost the most ground were West North Central and East South Central. Wages in the West North Central region fell from 7.0 percent below the national average in 1979-81 to 10.9 percent below in 1985-87. Wages in the East South Central region, which in the first two periods were the lowest in the country, fell even further, to 15.2 percent below the national average.

Wages in the Pacific and Middle Atlantic regions, on the other hand, increased relative to the national average. Overall, six of the nine census regions followed the pattern of wage convergence before 1979-81 and wage divergence after that period. The relative wage gains and losses across the nine regions combined to increase the standard deviation from 0.068 in 1979-81 to 0.086 in 1985-87, which is roughly the same level of dispersion found for the first period.

Components of Regional Wage Differentials

Which of the two components accounts for the switch from convergence to divergence? One way to address this question is to consider the number of cases in which one component or the other dominated the regional wage differential for all three periods. This could be interpreted as indicating that the same "structure" that led to wage convergence also led to wage divergence.

Looking only at the cross-sectional results, as shown in table 2, provides a mixed answer. For the six regions that followed the convergence/divergence pattern, differences in characteristic prices dominated the regional/national wage differential for three regions for all three periods, differences in characteristic levels dominated one region, and the effect was split for the remaining two regions. Tallying up the total number of cases in which differences in characteristic prices dominated the regional wage differentials results in about the same percentage of cases—about 60 percent.

Another way to evaluate the importance of each source is to determine the wage patterns generated if only one of the components varied. For instance, as shown in table 2, if workers were identical in all regions (or, at least, if the composition of worker attributes was the same) and only characteristic prices varied, four of the nine regions would exhibit a convergence/divergence wage pattern. These four cases are consistent with the actual wage patterns of convergence and divergence. The two regions in which price differentials did not yield the desired pattern, even though the actual wage pattern did, were the West South Central and South Atlantic regions. In both cases, differences in the characteristic levels were consistent with the actual wage patterns and were large enough to bring these patterns into line.

Which of the worker characteristics appears to contribute most to these patterns? Three categories of variables were considered: human capital variables, industry variables, and occupation variables. The most striking result (which is not shown in the tables) is that regional differences in the wage premiums paid in various industries virtually never emerged as the dominant category. Rather, human capital dominated in most cases, being the largest contributor in 16 of the 28 cases for the price component, and in 17 of the 28 cases for the level component.

Components of Intertemporal Regional Wage Changes

The previous examination of the sources of regional wage differentials looked at three separate cross sections from different time periods. The next step is to examine how these regional wage differentials changed over time. As mentioned earlier, equation (4) provides a framework to account for the various components of this wage change.

Table 3 displays the components of nominal wage changes for each region between the three time periods. For example, the 1.6 percent reduction between the first two periods in the wages of the East North Central region relative to the national average can be attributed to primarily two effects. The first is the main effect (column 1), which is the change over time in characteristic levels for the region relative to the nation. If all other effects were zero, then these changes in worker characteristics would cause the regional wages to diverge from the national average rather than to converge, as they actually do. The positive sign for this component indicates that the difference in the characteristic levels that favored this region over the nation was greater in the second period than in the first.

TABLE 2

Components of Regional Wage Differentials Relative to the National Average

Census Region	Year	(1) $(b^R-b^N)X^N$	(2) $(X^R-X^N)b^N$	(3) Actual Wage Difference
New England	1973-75	-.0064	-.0096*	-.0208
	1979-81	-.0469*	-.0123	-.0592
	1985-87	.0198*	.0138	.0305
Middle Atlantic	1973-75	.0547*	.0429	.0904
	1979-81	-.0005	-.0363*	.0327
	1985-87	.0221	.0504*	.0758
East North Central	1973-75	.0453	.0064*	.0490
	1979-81	.0117	.0133*	.0337
	1985-87	-.0010*	.0007	.0025
West North Central	1973-75	-.0289	-.0461*	-.0861
	1979-81	-.0137	-.0483*	-.0699
	1985-87	-.0447	-.0605*	-.1089
South Atlantic	1973-75	-.0332*	-.0303	-.0622
	1979-81	-.0459*	-.0027	-.0448
	1985-87	-.0396*	-.0076	-.0463
East South Central	1973-75	-.0861*	-.0695	-.1589
	1979-81	-.0714*	-.0264	-.1011
	1985-87	-.1047*	-.0544	-.1524
West South Central	1973-75	-.0915*	-.0176	-.1078
	1979-81	-.0496*	-.0006	-.0502
	1985-87	-.0471*	-.0020	-.0524
Mountain	1973-75	-.0316*	-.0021	-.0270
	1979-81	.0101	-.0220*	.0106
	1985-87	-.0158	-.0446*	-.0436
Pacific	1973-75	.0782*	.0495	.1063
	1979-81	.1280*	.0256	.1326
	1985-87	.1253*	.0333	.1427

NOTE: Column 1 is the effect of differences in characteristic price between the region and the nation; column 2 is the effect of differences in characteristic levels between the region and the nation. Columns 1 and 2 do not add up to column 3 because of a residual component not shown in the table. Asterisks denote the dominant component for each time period and region.
SOURCE: Author's calculations.

Offsetting the effect of changes in characteristic levels are the changes over the time periods in characteristic prices (column 4). If everything else remained the same, these changes in intertemporal prices would result in East North Central wages converging to the national average by 2.8 percent.

In determining which components contribute most to wage changes, two criteria were used. First, the signs of the components must be consistent with wage convergence between the first and second periods and with wage divergence between the second and third periods. Second, the components should account for a large share of the total wage change.

The asterisks in table 3 indicate the pairs of components that are consistent with the convergence/divergence wage pattern. For the two components that are based on the intertemporal change in characteristic prices (columns 3 and 4), 12 of the possible 18 pairs of estimates are consistent with the convergence/divergence wage pattern. The components related to intertemporal changes in characteristic levels (columns 1 and 2) contain only five pairs. Furthermore, the components related to changes in

TABLE 3

Components of Intertemporal Changes in Regional Wage Differentials

Census Region	Time Span	Components					
		(1)	(2)	(3)	(4)	(5)	(6)
New England	2-1	.002	.002*	–.003	–.034	–.036	–.039
	3-2	.030	–.004*	–.004	.061	.082	.091
Middle Atlantic	2-1	.002	–.001	–.008*	–.052*	–.059*	–.058*
	3-2	–.003	–.005	.017*	.025*	.034*	.044*
East North Central	2-1	.010	.0027	–.003*	–.020	–.010	–.016
	3-2	–.020	.005	.007*	–.014	–.021	–.030
West North Central	2-1	–.015	.010	.013*	.028*	.035*	.016*
	3-2	–.0003	.003	–.012*	–.023*	–.033*	–.038*
South Atlantic	2-1	.020*	–.001	.007*	–.011	.016	.017*
	3-2	–.004*	.006	–.001*	.009	.010	–.001*
East South Central	2-1	.023*	.005	.020*	.020*	.068*	.057*
	3-2	–.017*	.008	–.011*	–.016*	–.036*	–.050*
West South Central	2-1	.011*	.003	.006*	.043	.063	.057*
	3-2	–.001*	.002	–.0003*	.001	.001	–.001*
Mountain	2-1	–.028	–.020	.008*	.037*	–.003	.067*
	3-2	–.003	.016	–.020*	–.015*	–.022	–.053*
Pacific	2-1	–.008	–.018*	–.016*	.032	–.010*	.026
	3-2	–.001	.014*	.009*	.016	.038*	.011

NOTE: Time spans are denoted as 1 (1973-75), 2 (1979-81), and 3 (1985-87). The notation 2-1 represents the difference between the first two periods, and 3-2 represents the difference between the latter two periods. The components are (1) main effect, (2) interaction effects, (3) time-interaction effects, (4) regional time-interaction effects, (5) the sum of the four effects, and (6) the actual change in the regional wage differential (relative to the national average) between the two time periods. The asterisks indicate the components that are consistent with the convergence/divergence wage pattern.
SOURCE: Author's calculations.

prices (again columns 3 and 4) claim the largest share, on average, of the total wage changes. Consequently, it appears (as the trade theory suggests) that differences in characteristic prices account for the larger share of nominal regional wage changes over the three time periods.[14]

Therefore, this simple nonparametric test of counting the number of consistent results suggests that intertemporal changes in worker characteristic prices account for much of the convergence as well as the divergence of wages.

Consequently, basic changes in the way that worker characteristics were valued by the regional markets must have occurred around the turn of the decade. Trade theory suggests various types of market imperfections as possible candidates, including differences in production technologies and factor-market distortions. The back-to-back recessions in 1980-82 and the collapse in oil prices shortly thereafter certainly have taken their toll on regions such as the West South Central, while having little effect on others, such as the Pacific and New England regions. The relative effects of these events among regions can be partially explained by the slow adjustment of labor markets and the differential impact of oil prices between energy-using and energy-producing regions.

Considering the three categories of worker characteristics defined in the previous section offers further insight. As before, the industry variables played very little role in accounting for intertemporal changes in the regional wage differentials (these results are not shown in the

■ **14** Dickie and Gerking (1988) point out that omitted variables, particularly the lack of detailed human-capital variables, could bias the accounting method toward attributing too much importance to characteristic price differences. They find, using another data set that contained unusually detailed measures of worker and workplace characteristics, that they could not reject the hypothesis of equal coefficients across regions. This omission seems less critical for this study, since we look at the change over time in coefficients of the same set of variables within the same regions. It would seem that in order for omitted-variable bias to be significant, the relative contributions of each variable would have to vary substantially over time, which is not supported by the results from the previous section.

tables). However, unlike the cross-section analysis, occupation variables clearly dominated. For example, with respect to component four (differences in prices), occupation variables were the dominant category in 13 of the 18 cases.[15]

This result supports some of the speculation made by various authors about possible reasons why worker characteristic prices may not be equal across regions. Farber and Newman (1987) conjecture that characteristic prices may not necessarily converge because of occupation-specific demand disturbances. Topel (1986) shows that disequilibrium in local labor markets results primarily from stochastic disturbances in labor demand.

IV. Conclusion

After converging for almost half a century, nominal regional wages have diverged since 1980. This paper attempts to isolate the source of this switch in direction either as an intertemporal change in the market prices for worker attributes or as an intertemporal change in the levels of worker attributes. For nine census regions between the periods 1973-75, 1979-81, and 1985-87, results using individual workers from the CPS show that differences in characteristic prices account for a major share of the change in regional wages relative to the national average. Furthermore, virtually all of this intertemporal change in characteristic prices is found in the occupation coefficients; industry and worker characteristic variables account for very little.

Theory suggests that the prices of worker characteristics will converge in the presence of free commodity trade and in the absence of market imperfections. Various types of market imperfections were suggested as possible sources of the divergence of regional wages. For example, incomplete information, a mismatch between worker skills and job requirements, and institutional barriers to mobility can lead to incomplete adjustments to recent changes in the structural demand for labor. A recent study estimates that it takes as much as a decade for local labor markets to adjust fully to such shocks (Eberts and Stone [1989]).

Another possibility for nominal wage divergence is changes in the regional prices of housing and other nontraded goods that deviate from the national average. Because this study did not adjust for regional cost-of-living differences, it may be possible that wage differentials simply compensate workers for higher housing costs. However, this argument runs counter to the predicted results of free trade among regions, once equilibrium has been established. If goods are freely traded, then firms would be hard pressed to pay higher wages in some regions than in others, unless employers were compensated by differences in production technologies and worker productivity. Therefore, for cost-of-living differences to explain the results, workers in areas with higher labor costs coincidentally would have to be more productive. There are no compelling reasons why high living costs and high worker productivity should exist concurrently in equilibrium.

Two exceptions to this general statement are possible: First, site-specific attributes could enhance firms' productivity. Firms would move into the more productive region, bidding up the price of land and the price of labor, everything else being equal. The second possibility is that with the slow adjustment to shocks, we are simply observing these effects in disequilibrium.

The findings that differences over time in characteristic prices account for a majority of the changes in regional wage differentials does not necessarily diminish the importance of migration in explaining differences in regional growth. Rather, the analysis suggests that these flows have not changed the composition of regional labor forces significantly enough to make them the dominant factor in explaining changes in regional wage differentials. The traditional migration patterns of South to North and East to West are less pronounced now than in the past. Formerly, the primary migration pattern was toward the West, particularly for college graduates looking for job opportunities. More recently, the South is receiving many younger persons from the West and North.

If stochastic disturbances have changed the course of regional wage differentials, then it is interesting to speculate why these shocks have had such an impact in a relatively short period of time, when for so many decades the workings of efficient markets and equalizing migration flows seemed to prevail in forcing regional wages to converge. Several possibilities come to mind: increased foreign competition, the collapse of oil prices in the early 1980s, and the severe back-to-back recessions of 1980-82.

These recessions hit some regions harder than others, producing different patterns of change in regional wage differentials. The West South Central states of Texas and Louisiana were particularly

■ **15** Farber and Newman (1987) also find that the worker characteristics that accounted for much of the cross-sectional accounting of regional wage differences were different from the worker characteristics that accounted for the majority of the intertemporal changes in regional wage differentials.

hurt as the bottom dropped out of oil prices. This downturn thwarted the sizable gains that region had made in previous years in narrowing its wage gap.

The farming states of the West North Central region were also severely affected by the recession and the ensuing farm crisis of the early 1980s. After converging toward the national average throughout the 1970s, wages in this region diverged significantly, falling from 7.0 percent below the national average at the beginning of the 1980s to 10.9 percent below the average toward the end of the decade. Wages in some regions continued to grow faster than the national average in spite of the recession. For example, the Pacific region, especially California, was only mildly affected, with its regional wage differential expanding by a percentage point between 1979-81 and 1985-87.

Factors other than economic shocks could also contribute to the wage divergence. One possibility is state tax policies. The late 1970s and early 1980s saw the phasing out of substantial federal grant programs to states and municipalities. Many of these programs were designed to help equalize the fiscal burden across regions. As these funds dried up, many state and local governments found it necessary to raise tax rates to fund the existing programs, while others decided to curtail the programs. These different responses could lead to an increase in regional tax rates, which in turn could affect the location of firms and ultimately the demand for labor.

Will these factors persist? If history is any guide, the answer is no. The long-run trend of regional wage convergence has been interrupted only once in the last century. That episode lasted 20 years, embracing a postwar period and a much deeper and protracted recession than the one that greeted this decade. Consequently, it appears that shocks of this kind will eventually dissipate as the regions' economies regain a more equal footing.

However, many states and localities are not content to wait the decade or so that it takes for these forces to play themselves out. Many areas have pursued vigorous economic development efforts to help quicken the pace of adjustment. As long as these efforts attempt to remove market inefficiencies and strengthen the region's comparative advantage, they are socially desirable. One would expect that as regions continue to develop and mature—and barring further shocks of recent magnitude—the long-run trend of regional wage convergence will return.

References

Beeson, Patricia and Eberts, Randall W., "Identifying Productivity and Amenity Effects in Interurban Wage Differentials," *Review of Economics and Statistics,* August 1989 (forthcoming).

Bellante, Don, "The North-South Differential and the Migration of Heterogeneous Labor," *American Economic Review,* March 1979, *69,* 166–175.

Browne, Lynn E., "Shifting Regional Fortunes: The Wheel Turns," *New England Economic Review,* Federal Reserve Bank of Boston, May/June 1989, 27-40.

Dickie, Mark and Gerking, Shelby, "Interregional Wage Differentials in the United States: A Survey," in Alan Schlottmann, et al., eds., *Migration and Labor Market Adjustment,* Boston: Kluwer Academic Publishers, 1988.

Easterlin, Richard A., "Long Term Regional Income Changes: Some Suggested Factors," in Gerald A.P. Carrothers, ed., *Papers and Proceedings of the Regional Science Association,* 1958, *4,* 313-25.

Eberts, Randall W. and Stone, Joe A., "Wage and Employment Response to Shocks in Local Labor Markets," presented at the Western Economic Association Meetings, Lake Tahoe, Nev., July 1989.

Farber, Stephen C. and Newman, Robert J., "Accounting for South/Non-South Real Wage Differentials and for Changes in Those Differentials Over Time," *Review of Economics and Statistics,* May 1987, *69,* 215–23.

Gerking, Shelby and Weirick, William, "Compensating Differences and Interregional Wage Differentials," *Review of Economics and Statistics,* August 1983, *65,* 483–87.

Hanoch, Giora, "An Economic Analysis of Earnings and Schooling," *Journal of Human Resources,* Summer 1967, *2,* 310–29.

Markusen, James R., "Factor Movements and Commodity Trade as Complements," *Journal of International Economics,* May 1983, *14,* 341–56.

Mincer, Jacob, *Schooling, Experience, and Earnings,* New York: National Bureau of Economic Research, distributed by Columbia University Press, 1974.

Nourse, Hugh O., *Regional Economics: A Study in the Economic Structure, Stability, and Growth of Regions,* New York: McGraw-Hill, 1968.

Oaxaca, Ronald, "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review,* October 1973, *14,* 693–709.

Roback, Jennifer, "Wages, Rents and Quality of Life," *Journal of Political Economy,* December 1982, *90,* 1257–78.

Sahling, Leonard G. and Smith, Sharon P., "Regional Wage Differentials: Has the South Risen Again?" *Review of Economics and Statistics,* February 1983, *65,* 131–35.

Topel, Robert H., "Local Labor Markets," *Journal of Political Economy,* 1986, *94,* S111–43.

Why We Don't Know Whether Money Causes Output

by Charles T. Carlstrom
and Edward N. Gamber

Charles T. Carlstrom is an economist at the Federal Reserve Bank of Cleveland. Edward N. Gamber is an assistant professor of economics at the University of Missouri–St. Louis and is a visiting economist at the Federal Reserve Bank of Cleveland. The authors wish to thank John Carlson, Randall Eberts, William Gavin, William Osterberg, and Alan Stockman for helpful comments. In addition, we wish to thank Susan Black for valuable research assistance.

Introduction

Macroeconomics has undergone a revolution in the past 20 years, in which significant challenges have been made to supposedly well-established theories and facts. Among the most important of these prevailing theories is the positive correlation between money and real output.

Traditionally, most economists and policymakers have interpreted this correlation to imply that Federal Reserve open market operations could affect real output. This interpretation has persisted in spite of weak and sometimes contradictory empirical evidence. Unfortunately, we cannot attempt to examine all of the existing evidence on the direction of causality between money and output. Instead, this paper examines whether Granger-causality is a valid test for causality and what can be inferred from existing tests of Granger-causality. The answers to these questions are of paramount importance, since most policymakers assume that money causes output in a consistent and reliable way. This correlation is illustrated in figures 1, 2, and 3 using three measures of money: base, M1, and M2.[1]

The usual method of distinguishing among competing economic theories involves econometric testing. However, as is well known (see, for example, Black [1982]), econometric models indicate correlation, but not causality. Even the econometric technique of Granger (1969) does not necessarily identify causality as the term is commonly understood. We will show in the following section that the concept of Granger-causality is not robust to changes in the underlying model of the economy.[2] In other words, it is impossible to interpret Granger-causality independent of theory. Given this, sections II through IV examine models that try to explain the correlation between money and output.

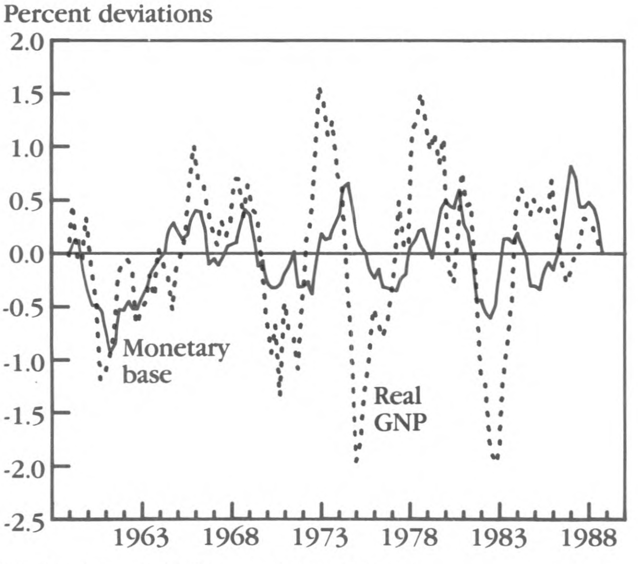
Traditionally, this correlation was explained by assuming some type of nominal rigidity (either prices or wages). Tobin (1970), however, showed that the correlation between money and output could be a result of the Federal Reserve's operating procedure and that it did not necessarily imply that changes in money caused output changes. Section III shows that if the Federal Reserve accommodates increases in output with a corresponding increase in the money supply,

■ 1 The series is detrended using a Hodrick and Prescott (1980) filter. Figure 4 illustrates this method as it is applied to real output (GNP).

■ 2 See also Cooley and LeRoy (1985).

FIGURE 1

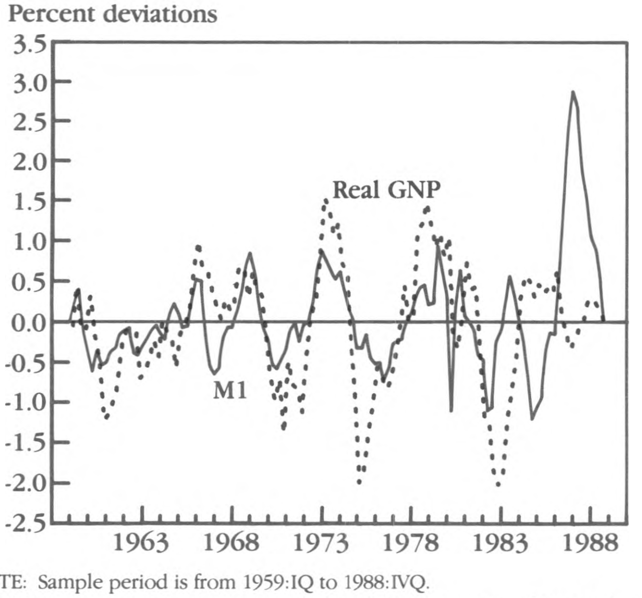
Real GNP and Monetary Base



NOTE: Sample period is from 1959:IQ to 1988:IVQ.
SOURCES: Data Resources, Inc., and Board of Governors of the Federal Reserve System.

FIGURE 2

Real GNP and M1



NOTE: Sample period is from 1959:IQ to 1988:IVQ.
SOURCES: Data Resources, Inc., and Board of Governors of the Federal Reserve System.

then one would expect to observe a positive correlation between output and money even though money is not causing output.

Real business cycle theorists have recently argued that the correlation between money and output could be due to reverse causality; that is, output can cause money independent of the Federal Reserve's reaction function. Section IV examines a model by King and Plosser (1984) showing that M1 and output are correlated because increases in real output cause increases in the demand for financial intermediation. This increased demand leads to the expansion of broader monetary measures, such as M1 and M2, even though changes in money have no influence on real output.

Section V reviews the empirical evidence uncovered in these theories to help ascertain the direction of causality in the money-output correlation. Section VI concludes with a discussion of policy implications.

I. Granger-Causality

Causality is a very elusive concept. In practice, most people define x causing y to mean that a change in x leads to a change in y. As an analogy, we would implicitly assume that if we could cause a low-pressure system to appear over a city (all else remaining constant), then there would be a high probability that rain would fall. This causality usually means that if low-pressure systems cause rain, then low-pressure systems must precede rain.

As can be seen in figure 3, M2 appears to lead GNP. Does this chronology imply that M2 causes GNP? The Granger definition of causality requires two assumptions. As stated by Granger and Newbold (1986, p. 220):

a) The future cannot cause the past. Causality can only occur with the past causing the present or future.

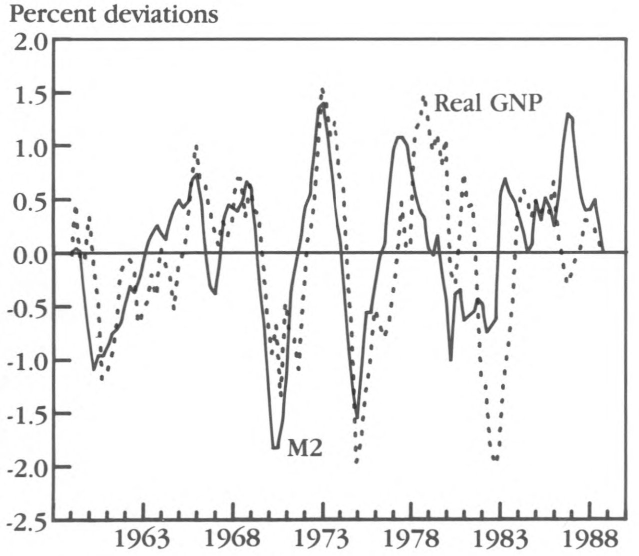
b) A cause contains unique information about an effect that is not available elsewhere.

According to the first assumption, then, if M2 always leads changes in GNP, we can logically infer that GNP does not cause M2. Does this mean that we can conclude the alternative, that M2 causes GNP? Consider the following example.

Suppose that a group of individuals always listens to weather forecasts and that these forecasts are always accurate. Further, suppose that these people decide to carry umbrellas on days that rain is forecasted. Clearly, carrying an umbrella and rain will be correlated, and carrying an umbrella will precede a rainstorm.

FIGURE 3

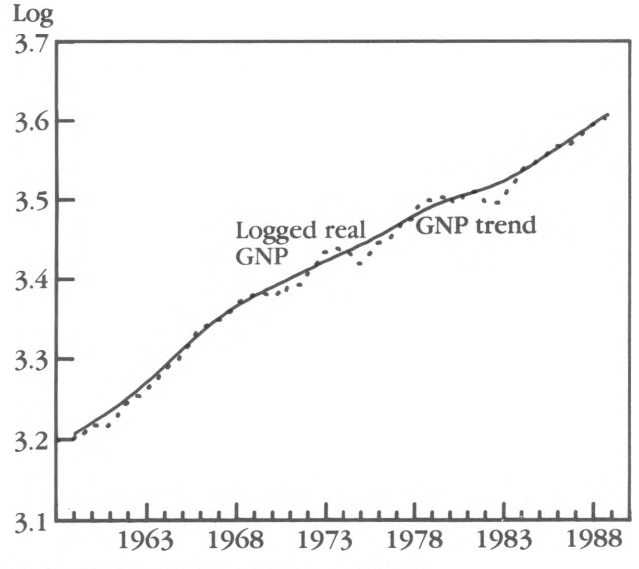
Real GNP and M2



NOTE: Sample period is from 1959:IQ to 1988:IVQ.
SOURCES: Data Resources, Inc., and Board of Governors of the Federal Reserve System.

FIGURE 4

Logged Real GNP and GNP Trend



NOTE: Sample period is from 1959:IQ to 1988:IVQ.
SOURCES: Data Resources, Inc., and Board of Governors of the Federal Reserve System.

According to the first assumption of Granger-causality, rain cannot cause umbrella-carrying. Yet, clearly, meteorologists would reject the conclusion that umbrellas cause rain.

The problem with our umbrella and rain example is that assumption a) is violated. This assumption is also frequently violated in many econometric tests. A third variable that uniquely causes people to carry umbrellas is omitted. Strictly speaking, rain does not cause umbrellas, but the *expectation* that rain may occur causes people to carry umbrellas. Expectations are not formed in a vacuum, however; low-pressure systems in this example could be shown to cause both umbrella-carrying and rain. Neglecting this third variable would cause one to conclude that carrying an umbrella Granger-causes rain.

Because of the importance of expectations in economics, a variable, x, that precedes another variable, y, will frequently not cause y. Variable x may depend on the expected value of y, causing x and y to be correlated. Since expectations depend on numerous variables that are, in principle, observable by the econometrician, one could conceivably conduct a Granger-causality test by including all relevant variables. The econometrician, however, would need to have a well-defined model of how expectations are formed. It is therefore extremely important that Granger-causality tests be interpreted in light of the theory that one is trying to test.

Consider the formal definition of Granger-causality. Let Ω_t be all the information available in the universe at time t. Let x_t and y_t be two random variables within this universe. Granger says that x causes (does not cause) y if

$$F(y_{t+k}|\Omega_t) \neq (=)\ F(y_{t+k}|\Omega_t - x_t)$$

for $k > 1$, where $F(.|.)$ is the conditional probability density function of y_{t+k} given Ω_t or $\Omega_t - x_t$, and $\Omega_t - x_t$ is defined to be the universe less x_t.

Suppose that these conditional distribution functions are equal. If x and y are correlated, it follows that there must exist a third variable in Ω_t that causes both x and y. For example, let y denote the occurrence of rain and let x denote the occurrence of umbrella-carrying. Leaving umbrella-carrying out of the information set does not affect the conditional distribution of rain or, in other words, weathermen can accurately predict rain without seeing whether people are carrying umbrellas. Because the entire universe, including low-pressure systems, is assumed to be in the information set, this example correctly predicts that umbrella-carrying does not Granger-cause rain.

Sims (1972) showed that Granger-causality is identical to the concept of exogeneity. In other words, x Granger-causes y if x is exogenous to y and y is not exogenous to x. A variable x is exogenous to y if the occurrence of x is independent of the occurrence of y. Similarly, a variable y is not exogenous to x if the occurrence of y is dependent on x occurring. Thus, the occurrence of rain is exogenous to whether people carry umbrellas: rain will fall regardless of whether people carry umbrellas. The converse is not true, however; if it starts to rain, people will tend to carry umbrellas.

At first glance, Granger-causality or exogeneity seems to be a reasonable definition of causality. However, it ignores the case of bivariate causality, where two variables cause each other. For example, rain causes puddles, and the evaporation of puddles causes rain to fall at a later date. To make Granger-causality operational, the universe of information must be restricted and the moments of the conditional distribution functions must be tested for equality. The universe of information is restricted by theory. In practice, the distribution functions are said to be equal if their first moments (the means) are equal. Testing for Granger-causality usually involves the following: A variable x is said to Granger-cause (not Granger-cause) y with respect to the information set I_t if

$$E(y_{t+\mathrm{k}} | I_t) \neq (=) \; E(y_{t+\mathrm{k}} | I_t - x_t) \text{ for k} \geq 1.$$

Because we do not consider all moments of the distribution, and we do not use all of the information set, Granger-causality as practiced is neither a necessary nor a sufficient condition to determine the direction of causation between x and y.

Consider the case where all the relevant information in the universe is included in a Granger-causality test, but only the means are tested to see if they are equal. If the means were found to be unequal, then one could logically infer that x must cause y. If the means were found to be equal, however, then one could not infer that x did not cause y.

Now consider the second assumption in the case where all the moments can be tested, but the universe of information is restricted in an ad hoc manner and an important determinant of y is accidentally omitted. Equality between the conditional distribution functions necessarily implies that x does not cause y. However, if the conditional distributions are not equal, then we cannot infer that x causes y. This is the case in our example: umbrellas help to predict rain and thus Granger-cause rain if low-pressure systems are excluded from the information set.

Since any operational test of causality involves restricting both the moments of the distribution functions to be tested and the information set in the universe relevant to the problem, employing a Granger-causality test exposes one to the risk of incorrectly rejecting causality when it is present and incorrectly rejecting the assumption of no causality when causality is not present. The econometrician can seek the direction of causality using a Granger-causality test only by using theory to determine which variables are helpful in predicting y_{t+k}. However, even after choosing variables based on some theory, a specification test should be conducted to help ensure that important variables have not been omitted.

It should be clear from this discussion that Granger-causality is neither a necessary nor a sufficient test for the existence of true causality. First, if bidirectional causality exists, then Granger-causality cannot indicate the presence of causality. Second, even when bidirectional causality is not present, the Granger-causality test may fail to identify whether causality is present if the information set excludes relevant variables or if all moments of the conditional distributions are not tested for equality. In addition, Granger-causality is not a useful test for showing the presence of contemporaneous causality.

Sections II and III present representative theories that have been developed to explain the money-output correlation. Section IV then interprets the econometric evidence that has been uncovered in light of these theories and the problems discussed above.

II. Money Causes Output

Most economists currently favor the interpretation that money causes output. They believe that some nominal rigidities, or price/wage sluggishness, allow changes in nominal variables, like money, to have real effects. These rigidities can be motivated by nominal wage contracts (Fischer [1977], Gray [1976]), or by incomplete information (Lucas [1972, 1977]).

For expositional ease, we consider the nominal wage contracting model as exemplified by Fischer. In his model, agents in the economy have rational (model-consistent) expectations, but wages are "sticky" because of the existence of long-term nominal wage contracts. Further, Fischer assumes that employment is demand-determined; that is, employment is always chosen so that the real wage is equal to the marginal productivity of labor. Thus, changes in the money supply that were unexpected at the time the contract was signed will have real effects. Unanticipated increases in the money supply will cause prices to be higher than expected and will cause the real wage to be lower than expected. The decline in the real wage lowers the marginal cost for firms to hire additional workers, leading to an expansion of employment and thus output.

Consider a scaled-down version of the model analyzed by Hoehn (1988). In this example, contracts will not be overlapping, and the only source of uncertainty will be from the money-supply process. Assume that the aggregate production function is Cobb-Douglas, that is, $Y_t = N_t^{\gamma}$, where Y_t and N_t are real output and the labor supply, respectively. Because wages are assumed to be demand-determined, we set the real wage equal to the marginal productivity of capital. Taking logarithms gives

$$(1) \quad w_t - p_t = ln(\gamma) - (1 - \gamma)n_t ,$$

where w_t, p_t, and n_t are the natural logarithms of wages, prices, and employment. Labor supply is assumed to be of the following form:

$$(2) \quad n^s_t = \beta_0 + \beta_1(w_t - p_t) \text{ for } \beta_0, \beta_1 > 0.$$

Setting labor supply equal to labor demand, one can solve for the real wage rate that clears the market. From this equation, it is assumed that wages are chosen so that the labor market clears on average.[3] This gives the following equation for nominal wages:

$$(3) \quad w^*_t = E_{t-1}p_t + [ln(\gamma) - (1 - \gamma)\beta_0]J ,$$

where $J = [1 + \beta_1(1 - \gamma)]^{-1}$.

To close the model, we must posit a form for money demand and money supply. Money demand is taken to be the simple quantity equation, that is, $M^d = KP_tY_t$. In logarithmic form, it is

$$(4) \quad m^d_t = p_t + y_t + k.$$

For our purposes, this year's log of money supply is equal to last year's money supply plus a random shock. That is, $m^s_t = m_{t-1} + \epsilon_t$, where the shock ϵ_t is assumed to be an independently, identically distributed random variable over time. With these assumptions, output equals

$$(5) \quad y_t = A + \gamma\epsilon_t ,$$

where $A = \gamma[\beta_0 + B_1 ln(\gamma)]J$.

For this simple case, in which contracts do not overlap and there are no shocks other than those to the money supply, changes in output depend only on the shock to this period's money, ϵ_t. If one were to randomly determine different realizations of ϵ_t, and were then to graph money supply and output against time (different realizations of ϵ_t), one would obtain a picture very similar to that given in figure 1. In this case, money causes changes in output. However, because changes in money and output occur contemporaneously, money does not Granger-cause output.

Equation (5) is also the output equation that results from a simple linearized version of the Lucas (1972, 1977) model. Here, workers confuse nominal and real shocks. Unanticipated increases in money result in higher nominal wages, which workers confuse with higher real wages. They do not know the extent to which higher wages reflect an increase in the relative price of their product or an increase in the general price level. Unanticipated changes in the money supply will cause increases in output as workers rationally mistake this nominal shock for a change in their real wage.

Models of the type discussed above were originally developed in response to the lack of empirical and theoretical support for traditional Keynesian and monetarist models. Both the Lucas and the Fischer models have recently come under attack. Barro (1977) shows that contracting models such as Fischer's are inconsistent with maximizing behavior. He argues that there is no a priori reason why labor should be demand-determined in these models.

In addition, economists question why firms have not indexed their wages, because sticky wages result in alleged output swings at both the firm and the macro level. Ahmed (1987) also presents empirical evidence showing that nominal wage contracting is not important for explaining output movements in Canadian data. Although Lucas' model is consistent with maximizing behavior, it also lacks empirical support. Mishkin (1983) and Boschen and Grossman (1982), for example, find evidence against the equilibrium monetary explanation of the business cycle.

The following section shows why the Federal Reserve's operating procedure may cause money and output to be correlated.

■ **3** Actually, this assumption is not quite true. Wages in Hoehn's model are chosen not so that $EN^d = N^S$, but so that $Eln(N^d) = En^d = lnN^S = n^S$.

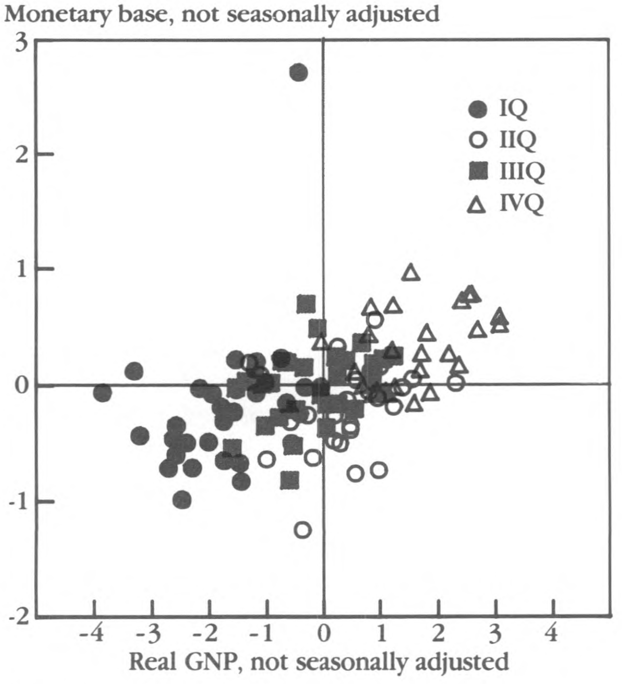
III. Post Hoc: Does The Federal Reserve Cause Christmas?

Figure 5 plots a scatter diagram of quarterly changes in the monetary base versus quarterly changes in output. Fourth-quarter points generally lie to the northeast of the first- through third-quarter points. Therefore, money and output are both higher on average in the fourth quarter, or around Christmastime. One could erroneously conclude that Federal Reserve policy causes holiday spending.

Clearly, causality in this case goes the other way. Output increases in the fourth quarter because of holiday spending, and the Federal Reserve, attempting to remove the seasonality from the interest-rate series, accommodates this higher output by increasing the money supply. This is an example of a point given by Tobin (1970) in his seminal article, "Money and Income: Post Hoc Ergo Propter Hoc?" meaning "after this therefore because of it." Tobin's argument was that a positive correlation between money and output may be the result of the Federal Reserve's operating procedure and not a reflection of the common belief that money causes output.

FIGURE 5

Real GNP and Monetary Base



NOTE: Sample period is from 1959:IQ to 1988:IVQ.
SOURCES: Data Resources, Inc., and Board of Governors of the Federal Reserve System.

Instead of presenting Tobin's model, we show how the operating procedure of the Federal Reserve can cause one to incorrectly conclude that the Federal Reserve causes, or at least influences, business cycles. Consider the following variation of the model presented in the previous section: Let output be Cobb-Douglas, so that the log of real wages will again be given by equation (1). Further, assume that the log of the labor supply is given by the following equation:

$$(6) \quad n^s_t = \beta_0 + \beta_1(w_t - p_t) + \beta_2 r_t$$

for $\beta_0, \beta_1, \beta_2 > 0$.

This equation differs from equation (2) because the labor supply is also assumed to be influenced by the real interest rate, r_t. Equation (6) assumes that the labor supply depends positively on the real interest rate, because of the intertemporal substitution effect. That is, when interest rates are high, workers transfer consumption from today until tomorrow to take advantage of the high real rate. Consumption is reduced, thus increasing the marginal utility of consumption in the current period. This, in turn, increases the incentive for agents in the economy to work additional hours in order to consume more today.

Instead of assuming that there are long-term nominal wage contracts, this model assumes that wages vary to clear the market continuously so that money does not influence output. By equating the real wage in equations (1) and (5), we solve for the equilibrium amount of labor supplied (demanded) in this economy:

$$(7) \quad n_t = [\beta_0 + \beta_1 ln(\gamma)]J + J\beta_2 r_t.$$

Real interest rates in the economy are assumed to fluctuate randomly around a constant mean r:

$$(8) \quad r_t = r + \eta_t.$$

Temporary changes in interest rates, η_t, can result because of either shifting tastes or temporary changes in government expenditures. Incorporating this variable into equation (7), we see that output depends positively on the innovation in real interest rates today.

$$(9) \quad y_t = \gamma[\beta_0 + \beta_1 ln(\gamma) + \beta_2 r]J + \gamma J \beta_2 \eta_t.$$

To close the model, we assume that money demand is given by equation (4) and that the Federal Reserve follows a nominal interest rate rule:

(10) $m^s_t = b + \lambda(R_t - r)$, and $\lambda > 0$,

where $R_t = r_t + E_t p_{t+1} - p_t$.

Nominal interest rates are assumed to be the sum of the real rate plus expected inflation over the next period. Using equations (4), (8), (9), and (10), the reduced form for the nominal interest rate is given by the following equation:

$$(11)\quad R_t = r + \eta_t[(1/(1+\lambda)) + \gamma\beta_2 J/(1+\lambda)].$$

Innovations in the real interest rate are assumed to be temporary. An increase in the real interest rate causes policymakers to expand the money supply in order to stabilize nominal interest rates. Prices are then temporarily high and deflation is expected over the next period, which will offset the increase in the real interest rate. When λ approaches infinity, the nominal interest rate approaches the long-term real interest rate, r. That is, when λ approaches infinity, the Federal Reserve is following an interest-rate peg.

From equation (11), the reduced form of the money-supply equation is given by

$$(12)\quad m^s_t = b + \lambda[(1/(1+\lambda)) + \gamma\beta_2 J/(1+\lambda)]\eta_t.$$

If one were to randomly determine different realizations of η_t, and were then to graph money supply and output against time (different realizations of η_t), one would again obtain a picture very similar to that given in figure 1. A temporary increase in interest rates causes people to supply more labor today. This occurs since high real interest rates imply that, on the margin, individuals greatly value consumption today, causing them to work longer hours today. The increase in interest rates also causes the Federal Reserve to expand the money supply in order to smooth nominal interest rates, which causes a temporary rise in prices.

This example implies that, on average, prices will fall over the next period, leading to a decline in the nominal interest rate. Unlike the example given in the previous section, interest rates in this model cause changes in both output and money. Thus, money and output are positively correlated. Like the example given in section III, however, interest rates do not Granger-cause output, because interest rates and output occur contemporaneously.

The above model illustrates how an interest-rate target can produce a positive correlation between money and output. The example was extremely simple and predicted that money and output would move contemporaneously. One could likewise construct examples in which money leads changes in output and would thus appear to cause changes in output.

For example, consider an economy in which money has no real effects, but in which agents are able to predict future output. The prospect of higher future output will cause agents to borrow (or save less) in an attempt to smooth their consumption stream over time. This increased borrowing will boost interest rates. If the effect on output today from an increase in interest rates is negligible, then changes in money will occur before changes in output when the Federal Reserve pursues an interest-rate peg. In this economy, money leads, but does not cause, output.

The next section discusses another mechanism in which output can cause changes in money. Unlike the model presented in this section, the mechanism will not come from the Federal Reserve's operating procedure, but will result from the public's willingness to hold currency versus either demand or time deposits.

IV. Output Causes Money

Real business cycle theorists typically assume that the cause of business cycles is either a shock to consumer preferences or a shock to real productivity.[4] Because an indirect measure of these shocks can be obtained through the use of Solow residuals (see Solow [1956]), theorists have tended to concentrate on technology shocks as a source of business cycle fluctuations.

Real business cycle theory has been successful in explaining the quantitative aspects of business cycles. These include the standard deviations of—and comovements among—real variables such as output, investment, consumption, and hours worked. In contrast, monetary-driven business cycle models have concentrated on explaining the qualitative aspects of the correlation between money and output.[5]

Because real business cycle models do not include a role for money, they have been criticized for not explaining the comovements

■ 4 For a thoughtful exposition of real business cycles, see Prescott (1986) or Stockman (1988).

■ 5 As noted by Stockman (1988, p. 35), "The large-scale econometric models do not qualify because they are not true structural models in the sense of the Lucas critique of econometric policy evaluation...."

TABLE 1

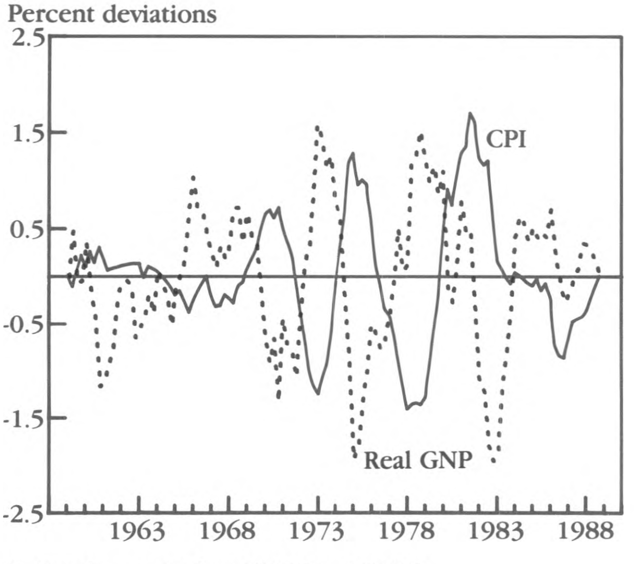
Cross-Correlations of Output

Variable x	$x(t-4)$	$x(t-3)$	$x(t-2)$	$x(t-1)$	$x(t)$	$x(t+1)$	$x(t+2)$	$x(t+3)$	$x(t+4)$
Real GNP	.25	.45	.68	.87	1.00	.87	.68	.46	.26
Monetary base	.08	.20	.29	.39	.44	.44	.42	.39	.37
M1	.12	.25	.35	.37	.34	.26	.21	.16	.15
M2	.62	.68	.69	.62	.48	.29	.10	–.08	–.25
Interest rates	–.55	–.36	–.15	.10	.34	.47	.51	.49	.45
Real interest rates	–.57	–.41	–.20	.07	.30	.40	.44	.43	.41
CPI	–.70	–.75	–.75	–.69	–.56	–.40	–.23	–.05	.13

NOTE: Sample period is from 1959:IQ to 1988:IVQ.
SOURCES: Data Resources, Inc., and Board of Governors of the Federal Reserve System.

FIGURE 6

Real GNP and Consumer Price Index



NOTE: Sample period is from 1959:IQ to 1988:IVQ.
SOURCES: Data Resources, Inc., and Board of Governors of the Federal Reserve System.

among nominal variables such as the price level, wages, and money (see Summers [1986]). However, as figure 6 and table 1 illustrate, the comovements among interest rates, prices, and real output are qualitatively consistent with real business cycle theory. In particular, interest rates have been contemporaneously procyclical and prices have been countercyclical since 1959.[6]

Procyclical interest rates arise in real business cycle models generated by temporary productivity shocks. A temporary increase in productivity today, which is expected to lead to higher output in the future, causes individuals to borrow money in order to smooth consumption. Countercyclical prices arise in these models because the demand for real money balances increases when output increases. Assuming that the Federal Reserve does not fully accommodate the increases in interest rates and output, it follows that prices must fall.

Table 1 provides further evidence that the Federal Reserve may accommodate increases in output. Note that the strongest correlations between the monetary base and output occur contemporaneously and with money lagging output by one quarter. Real business cycle theorists argue that the correlation between the monetary base and output is the result of the Federal Reserve's operating procedure. They point out that this correlation is small relative to the correlation between output and broader measures of money, such as M1 and M2.

Table 1 shows that while the contemporaneous correlation between the monetary base (percent deviations from trend) and real GNP is only .44, the correlation between M2 (percent deviations from trend) and real GNP two quarters later is .68. Although table 1 indicates that the correlation between M1 and real GNP is similar to the correlation between the monetary base and real GNP, the correlation between M1 (percent deviations from trend) and real GNP is .59 if one ignores the tremendous increase in M1 during 1986.

While the monetary base is determined solely by the Federal Reserve, components of M1 and M2, such as checking accounts, short-term time deposits, money market accounts, and mutual

■ 6 Prior to 1953, prices seem to be more procyclical.

funds, are determined by commercial banks and the public.[7] This suggests an important role for reverse causality. The public appears to respond endogenously to future output changes by shifting its portfolio from currency to demand and time deposits. Some mechanism must therefore serve to link output and deposits.

King and Plosser (1984) develop a model in which individuals demand both currency and financial services (demand deposits). In their model, demand deposits, like other goods, are produced with capital and labor. They derive a demand curve for both inside money (financial services) and outside money (currency). They assume that the cost of making a transaction depends negatively on the real amount of inside and outside money that a person holds. The demand for both financial services and currency increases with real output in this model, explaining why empirically both real currency and real demand deposits are correlated with real output.

However, King and Plosser also show empirically that there is a positive correlation between nominal demand deposits and currency with real output. If one restricts their cost of transactions and assumes that with larger purchases (higher output) there is an extra cost associated with currency over demand deposits, one can also generate a positive correlation between nominal demand deposits and output. This assumption seems natural because the demand for high-ticket durable goods is much more procyclical than for less-expensive purchases such as services. A model like this can explain the positive correlation between nominal bank deposits and real GNP.

An example of reverse causality occurred during the Great Depression. The monetary base grew slightly through the period, while the money supply, defined by M1, declined substantially as depositors shifted out of demand deposits and into currency. The result was a decline in the currency/deposit ratio as output fell and banks failed. The ensuing bank failures were probably both a cause and an effect of the Great Depression. The decline in the money supply, therefore, was partly the effect of factors that caused the Great Depression, although it may also have been a contributing factor in causing the financial collapse.[8] Empirical work has not been able to distinguish this causation.

Real business cycle models have generated a resurgence in interest to test for the direction of causality between money and output. The next section reviews this literature in light of the theories presented in sections I through IV.

V. Tests of the Money-Output Relationship

To determine the direction of causality between money and output, economists since Sims (1972) have employed Granger-causality tests. The results of these tests are not robust to changes in the sample period, to changes in the variables included in the test, or to whether the data are in log-level or first-differenced form.

Sims finds that money Granger-causes output in a simple bivariate setting. In a later paper, Sims (1980) determines that money fails to Granger-cause output when the commercial paper rate is included in the test. Litterman and Weiss (1985) replicate this result and also show that the nominal commercial paper rate Granger-causes both money and output. They find that the real interest rate, however, does not Granger-cause either output or money.

Eichenbaum and Singleton (1986) replace the commercial paper rate with the real rate of return on stocks and the real rate of return on Treasury bills in their Granger-causality tests. They find that while the real rate of return on Treasury bills does not Granger-cause output, the real rate of return on stocks does. Their model allows no explanatory power for money once these variables are included.

Stock and Watson (1989) find that money Granger-causes output if the rate of return on stocks is omitted and the nominal rate of return on Treasury bills is included. Friedman and Kuttner (1989), however, find that this result is sensitive to the sample period chosen. They also determine that money fails to Granger-cause output (except for one subsample) when the nominal commercial paper rate is replaced by the spread between the commercial paper rate and the Treasury bill rate.

What do these results tell us about the direction of causality between money and output? First, the inclusion of interest rates seems to weaken the explanatory power of money. This seems to be inconsistent with a money-driven business cycle. McCallum (1983), however, argues (but does not show) that if the Federal Reserve attempts to peg the interest rate, then interest-rate innovations are a better indication of the influence of money on output than are monetary innovations. This result is obtained

■ **7** The Federal Reserve currently can control the nonborrowed monetary base with a fair amount of precision. However, to control total monetary base, the Federal Reserve would need to alter the current administrative practices of the discount window and reserve accounting practices. See Laurent (1979).

■ **8** See Friedman and Schwartz (1963).

because monetary innovations that affect output also cause interest rates to change. There are also nonmonetary shocks that cause interest rates to change, leading to changes in output.

Second, different measures of the rate of return yield drastically different results. The reason is probably that some rates of return are a better proxy for future changes in output than others. As Friedman and Kuttner indicate, the primary determinant of the spread between the Treasury bill rate and the commercial paper rate is the default risk on corporate securities. The primary determinant of the default risk of corporate securities is probably the anticipation of future business conditions, that is, future changes in output. The real rate of return on stocks in Eichenbaum and Singleton's study is probably also a proxy for future changes in output.

The issue of whether money is significant in its ability to predict future output when the spread or return on stocks is included in the causality test tells us little about the actual direction of causality between money and output. Money will Granger-cause output whether money actually causes output or whether future output causes money, whenever the spread (or the return on stocks) is a proxy, but an imperfect proxy for future output. Money would appear to be significant for both models because it helps to eliminate some of the noise present in the spread. Similarly, money will not Granger-cause output if the spread (or the return on stocks) is a perfect proxy for future output. The two models, money causing output and output causing money, are thus observationally equivalent in their predictions concerning whether money Granger-causes output.

This analysis indicates that inferences about the direction of causality between money and output cannot be made from the existing Granger-causality tests. One of the major problems with the existing empirical studies is that they use M1 as their measure of money. As indicated in the previous sections, broader measures of money respond to future business conditions more than narrow measures of money, such as the monetary base. It appears that it would be difficult to distinguish between money causing output or output causing money when measures of money containing endogenous components are used. The same caveat holds for narrow measures of money like the monetary base. These measures, however, do not seem to respond to future business conditions to the same degree as M1 or M2.

These results suggest that the use of causality tests should proceed along the lines indicated by Sims (1989). He urges that researchers should concentrate on combining the theoretical techniques developed by real business cycle theorists and the empirical technique of vector autoregressions. That is, researchers should proceed along the lines of Prescott (1986), but should compare more than simple correlations when matching simulated data to actual data. Sims recommends that they compare the results of Granger-causality tests run on both simulated data and actual data. This requires models to pass stricter empirical tests before being judged as either successful or unsuccessful. Applying this technique to help determine the direction of causality between money and output would require building a real business cycle model with money and then comparing the vector autoregressions run on simulated data from both models with actual data.

VI. Conclusion and Policy Implications

This paper has shown that Granger-causality tests alone cannot settle the debate about the direction of causality between money and output. One reason is the ever-present problem of a potentially missing third variable. In section I, we showed how umbrellas could Granger-cause rain when a variable proxying for the expectation of rain, low-pressure systems, is excluded from the tests. The above studies seem to affirm the notion that leaving out variables that proxy for the expectation of future output could leave money with explanatory power when no causality is actually present. It should be clear that this debate is not likely to be settled on the basis of Granger-causality tests alone. Unfortunately, the issue can probably never be completely settled without having the Federal Reserve conduct controlled experiments with monetary policy that would be infeasible.

Causality tests are not necessarily useless, however. They may provide some information about the direction of causality, as long as they are interpreted within the confines of a model. That is, we must start with the null hypothesis that a specific model is correct and attempt to test whether or not we can reject this hypothesis. This approach is in the spirit of Eichenbaum and Singleton (1986); however, the suggestions made by Sims (1989) seem more appropriate.

Many policymakers currently assume that money causes output in a consistent and reliable way. Economists have been unable to demonstrate this relationship, however. If money does *not* cause output, are policies predicated on such causation benign or harmful? At first glance, it would seem that the effects of current policy would be benign if money does not cause output.

However, by not being able to pin down the direction of causality, we cannot rule out other possibilities. For instance, it may be possible that inflation or monetary growth decreases output. Support for this proposition comes from Kormendi and Meguire (1985). Using cross-country data, they find a negative correlation between inflation and the growth rate of real output. The possibility that inflation may lower output should not be too surprising, given that inflation is a tax on real cash balances. As is the case with any other tax, we would expect increases in this tax to depress output. For example, higher rates of inflation cause people to engage in wasteful activities in order to economize on money holdings, thus serving to lower output.

Because researchers cannot tell whether increases in money cause output to increase—and there is some evidence that increases in the growth rate of money actually depress output—how should policymakers proceed? Policy actions should be analyzed in light of their potential costs and benefits. Traditional Keynesian analysis assumes that all output fluctuations are inefficient and that policy could improve economic welfare by stabilizing output. However, as Lucas (1987) points out, the welfare gains associated with smoothing business-cycle fluctuations are small and are dwarfed by the potential gains associated with increasing long-run economic growth.

The costs associated with stabilizing output may not be small. If unanticipated money increases output as described by Lucas (1972, 1977), then the real output effects from money are welfare-reducing. The reason is that the output effects of money are generated by misperceptions on the part of the public. As Lucas points out, this analysis prescribes that the Federal Reserve should follow a rule when conducting monetary policy. In Lucas's model, any output changes induced by money are inefficient. Even if his reasons for why money affects output are incorrect, it still may be best for policymakers to follow a rule.

Stockman (1988) also makes the point that conducting policy as if output fluctuations are inefficient can be damaging. If the true explanation of business cycles turns out to require both Keynesian and real business cycle elements, then there may be substantial welfare losses associated with output stabilization. As argued by real business cycle theorists, some output changes are efficient. In addition, it is presently impossible to distinguish inefficient from efficient movements in output. Using monetary policy to offset these shocks could very well leave us worse off. Therefore, even if money has real effects, it is not clear how aggressively, if at all, monetary policy should try to stabilize output.

Policymakers should accept the possibility that money does not cause output. Instead of conducting policy as if money does cause output, they could base monetary policy on what we currently know about its costs and benefits. The preceding analysis leads us to believe that policymakers should be more reluctant to fine-tune the economy without understanding the inefficiencies present in the economy. Because the costs of economic stabilization are thought to be large, while the potential benefits have been shown to be fairly small, we recommend that monetary policy be predicated on a rule that is easy for policymakers to implement and even easier for the public to monitor.

References

Ahmed, Shaghil, "Wage Stickiness and the Non-Neutrality of Money: A Cross-Industry Analysis," *Journal of Monetary Economics,* July 1987, *20,* 25-50.

Barro, Robert J., "Long-Term Contracting, Sticky Prices and Monetary Policy," *Journal of Monetary Economics,* July 1977, *3,* 305-16.

Black, Fischer, "The Trouble with Econometric Models," *Financial Analysts Journal,* March/April 1982, *38,* 29-38.

Boschen, John F. and Grossman, Herschel I., "Tests of Equilibrium Macroeconomics Using Contemporaneous Monetary Data," *Journal of Monetary Economics,* November 1982, *10,* 309-34.

Cooley, Thomas F. and LeRoy, Stephen F., "Atheoretical Macroeconomics: A Critique," *Journal of Monetary Economics,* November 1985, *16,* 283-308.

Eichenbaum, Martin and Singleton, Kenneth J., "Do Equilibrium Real Business Cycle Theories Explain Postwar U.S. Business Cycles?" *NBER Macroeconomics Annual,* Cambridge, Mass.: MIT Press, 1986, 91-135.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy,* February 1977, *85,* 191-206.

Friedman, Benjamin M. and Kuttner, Kenneth N., "Money, Income and Prices after the 1980s," National Bureau of Economic Research Working Paper No. 2852, February 1989.

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

Granger, C.W.J., "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica,* July 1969, *37,* 424-38.

________ **and Newbold, Paul,** *Forecasting Economic Time Series,* second edition, Orlando, Fla.: Academic Press, 1986.

Gray, Jo Anna, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics,* April 1976, *2,* 221-35.

Hodrick, Robert and Prescott, Edward, "Postwar U.S. Business Cycles: An Empirical Investigation," Working Paper, Carnegie-Mellon University, 1980.

Hoehn, James G., "Procyclical Real Wages Under Nominal-Wage Contracts with Productivity Variations," *Economic Review,* Federal Reserve Bank of Cleveland, Quarter 4, 1988, 11-23.

King, Robert G. and Plosser, Charles I., "Money, Credit, and Prices in a Real Business Cycle," *American Economic Review,* June 1984, *74,* 363-80.

Kormendi, Roger C. and Meguire, Philip G., "Macroeconomic Determinants of Growth: Cross-Country Evidence," *Journal of Monetary Economics,* September 1985, *16,* 141-64.

Laurent, Robert D., "Reserve Requirements: Are They Lagged in the Wrong Direction?" *Journal of Money, Credit, and Banking,* August 1979, *11,* 301-10.

Litterman, Robert B. and Weiss, Laurence, "Money, Real Interest Rates, and Output: A Reinterpretation of Postwar U.S. Data," *Econometrica,* January 1985, *53,* 129-56.

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

________ , "Understanding Business Cycles," in *Stabilization of the Domestic and International Economy,* edited by Karl Brunner and Allan H. Meltzer, Carnegie-Rochester Conference Series on Public Policy, Amsterdam: North Holland Publishing Co., 1977, *5,* 7-30.

________ , *Models of Business Cycles,* New York: Basil Blackwell, Inc., 1987.

Mankiw, Gregory, "Comment: Do Equilibrium Real Business Cycle Theories Explain Postwar U.S. Business Cycles?" *NBER Macroeconomics Annual,* Cambridge, Mass.: MIT Press, 1986, 139-45.

McCallum, Bennett T., "A Reconsideration of Sims' Evidence Concerning Monetarism," *Economic Letters,* 1983, *13,* 167-71.

Mishkin, Frederic S., *A Rational Expectations Approach to Macroeconometrics: Testing Policy Ineffectiveness and Efficient-Markets Models,* Chicago: University of Chicago Press, 1983.

Prescott, Edward C., "Theory Ahead of Business Cycle Measurement," *Quarterly Review,* Federal Reserve Bank of Minneapolis, Fall 1986, 9-22.

Sims, Christopher A., "Money, Income and Causality," *American Economic Review,* September 1972, *62,* 540-52.

———, "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered," *American Economic Review,* May 1980, *70,* 250-57.

———, "Models and Their Uses," Institute for Empirical Macroeconomics, Federal Reserve Bank of Minneapolis, Discussion Paper No. 11, April 1989.

Solow, Robert, "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics,* February 1956, *70,* 65-94.

Stock, James H. and Watson, Mark W., "Interpreting the Evidence on Money-Income Causality," *Journal of Econometrics,* 1989 (forthcoming).

Stockman, Alan C., "Real Business Cycle Theory: A Guide, an Evaluation, and New Directions," *Economic Review,* Federal Reserve Bank of Cleveland, Quarter 4, 1988, 24-47.

Summers, Lawrence H., "Some Skeptical Observations on Real Business Cycle Theory," *Quarterly Review,* Federal Reserve Bank of Minneapolis, Fall 1986, 23-27.

Tobin, James, "Money and Income: Post Hoc Ergo Propter Hoc?" *Quarterly Journal of Economics,* May 1970, *84,* 301-17.

Third Quarter Working Papers

Working Paper Notice

Current *Working Papers* of the Cleveland Federal Reserve Bank are listed in each quarterly issue of the *Economic Review.* Copies of specific papers may be requested by completing and mailing the attached form below.

Single copies of individual papers will be sent free of charge to those who request them. A mailing list service for personal subscribers, however, is not available.

Institutional subscribers, such as libraries and other organizations, will be placed on a mailing list upon request and will automatically receive *Working Papers* as they are published.

■ **8908**
Intervention and the Risk Premium in Foreign Exchange Rates
by William P. Osterberg

■ **8909**
Public Infrastructure and Regional Economic Development: A Simultaneous Equations Approach
by Kevin T. Duffy-Deno and Randall W. Eberts

■ **8910**
Structure, Conduct, and Performance in the Local Public Sector
by Randall W. Eberts and Timothy J. Gronberg

■ **8911**
Factor-Adjustment Costs at the Industry Level
by Mary E. Deily and Dennis W. Jansen

■ **8912**
Enforcement of Pollution Regulations in a Declining Industry
by Mary E. Deily and Wayne B. Gray

Please complete and detach the form below and mail to:

Federal Reserve Bank of Cleveland
Research Department
P.O. Box 6387
Cleveland, Ohio 44101

Check item(s) requested

Please send the following Working Paper(s):

☐ **8908** ☐ **8910** ☐ **8912**

☐ **8909** ☐ **8911**

Send to: Please print

Name

Address

City State Zip