Economic Review

Federal Reserve Bank of San Francisco

Fall 1988 Number 4

Table of Contents

The Federal Reserve Bank of San Francisco's Economic Review is published quarterly by the Bank's Research Department under the supervision of Jack H. Beebe, Senior Vice President and Director of Research. The publication is edited by Barbara A. Bennett. Design, production, and distribution are handled by the Public Information Department, with the assistance of Karen Rusk and William Rosenthal.

For free copies of this and other Federal Reserve publications, write or phone the Public Information Department, Federal Reserve Bank of San Francisco, P.O. Box 7702, San Francisco, California 94120. Phone (415) 974-2163.

Saving, Investment, and the U.S. External Balance

Ramon Moreno

Economist, Federal Reserve Bank of San Francisco. Editorial committee members were Michael Keeley, Reuven Glick, and Ronald Schmidt.

The unprecedented rise in the U.S. external deficit in the 1980s was not only the result of large government budget deficits. A significant decline in the extent to which the private saving-investment balance adjusted to finance government budget deficits also contributed to the U.S. external deficit. It is hypothesized that the shift in U.S. monetary policy after 1979 reduced domestic financing of U.S. budget deficits in the 1980s by encouraging foreign capital inflows.

The size and persistence of the U.S. external deficit in recent years is unprecedented in this century, and has prompted extensive discussion and research on its underlying causes. Many observers have argued that large government budget deficits are primarily responsible for the U.S. external deficits. However, external deficits depend not only on government budget deficits, but on the private saving-investment balance, as well. This paper discusses the role of the private saving-investment balance in the growth of U.S. external deficits of the 1980s.

Prior to the 1980s, the private saving-investment balance varied negatively with the government budget balance, almost fully offsetting budget deficits. Thus, until the 1980s, budget deficits largely were not associated with external deficits. The extent of this offset decreased significantly in the eighties, thereby increasing the impact of budget deficits on the U.S. external position. This change in the behavior of the private saving-investment balance helps to explain why in the 1980s the external deficit rose in response to the increase in the government budget deficit.

This study argues that the change in the behavior of the private saving-investment balance may have been caused by the change in inflationary expectations associated with the shift in monetary policy that took place at the end of 1979. Specifically, after 1979, the change in monetary policy meant that higher budget deficits no longer would cause money growth, inflation, and inflationary expectations to rise automatically, thereby increasing the willingness of foreigners to finance such deficits.

The paper is organized as follows. Section I discusses the behavior of budget deficits, the private saving-investment balance, and the external balance between 1960 and 1987. It reviews the findings of recent empirical and simulation studies on the response of the private saving-investment balance to fiscal and monetary policy and identifies certain developments that may have changed this response over time. Section II investigates the empirical relationship between fiscal and monetary policy and the private saving-investment balance, and tests for changes in this relationship after 1974 and after 1980. Section III discusses some factors that may have contributed to the change in the response of the private saving-investment balance to fiscal policy. Section IV summarizes the findings of this paper and highlights some policy implications.

I. Saving, Investment, and the External Deficit—An Overview

Internal and External Balances

To set the context for the discussion that follows, consider the national income accounting identity:

$$Y = C + I + G + B = C + S + T \qquad (1)$$

where all variables are real and:

Y = gross national product;
C = domestic consumption;
I = domestic gross private investment;
G = domestic government expenditure;
B = exports minus imports of goods and services
= external balance;
S = private domestic saving; and
T = government receipts.

Dropping C from both sides, and re-arranging yields the following:

$$(S - I) + (T - G) = B = \text{Net capital flow} \qquad (2)$$

where the net capital flow is the difference between U.S. investment abroad and foreign investment in the United States.

Equation (2) describes the external balance of an economy as the sum of the saving of its private sector, or the private saving-investment balance (the difference between gross private saving and gross private investment), and of its public sector, or the government budget balance. The left hand side corresponds to the internal balance of the economy, the right hand side to the external balance.

Equation (2) also illustrates why the external balance may be interpreted as the saving of the economy as a whole. A country experiencing an external surplus is producing more than it spends, and its saving is used to finance excess foreign spending. Conversely, a country experiencing an external deficit is purchasing more goods than it produces, and foreign capital inflows finance excess domestic spending. There must be a correspondence between a country's external balance and the balance in its capital account.

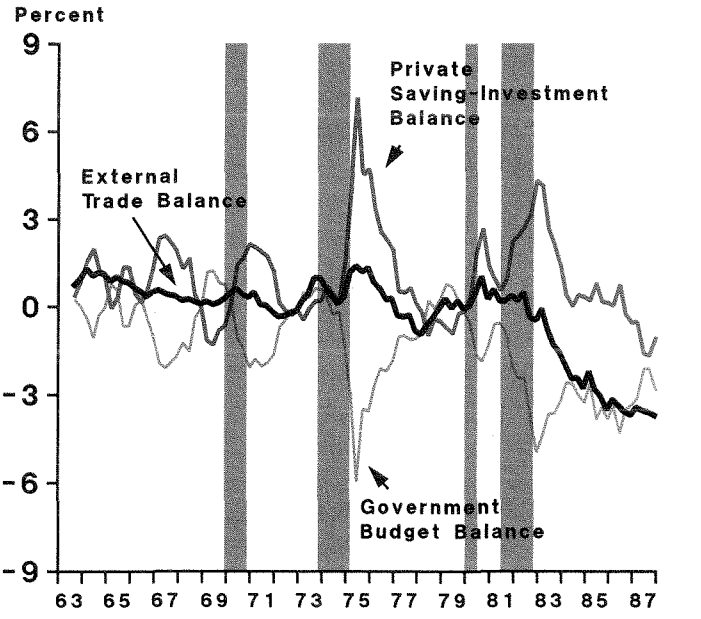
Chart 1 illustrates the path of the U.S. external balance since 1960.[1] The series is nominal (not adjusted for inflation), and shown as a proportion of the middle expansion trend of GNP.[2] As a net exporter of capital, the U.S. maintained a trade surplus averaging over two-fifths of a percent of GNP up to 1980. However, the U.S. external balance began falling sharply at the end of 1982. Between the last quarter of 1982 and the last quarter of 1987, the nominal U.S. external deficit averaged 2.5 percent of GNP, and peaked at 3.7 percent of GNP in the last quarter of 1987. The magnitude of the U.S. external deficit since 1982 is unprecedented in the twentieth century.

The duration of this external deficit also is unprecedented. Over five years have passed since the U.S. external balance went into deficit in the third quarter of 1982. In contrast, from the end of World War II until 1980, the U.S. experienced external deficits for more than one quarter on only four occasions, and the average duration was less than a year and a half.

Chart 1 also illustrates the two components of internal balance that together equal the external balance—the government balance and the private saving-investment balance, both as proportions of the middle expansion trend of GNP.[3] The chart shows that during contractions, the government balance tends to fall as tax revenues fall, and the private saving-investment balance tends to rise as investment declines. The reverse is true during expansions.[4]

However, the most recent economic expansion, which began in 1982, has not been accompanied by the typical reduction in the government deficit that has characterized earlier expansions. Given the typical cyclical reduction in

Chart 1
U.S. External and Internal Balance*



*Shaded areas indicate recessions as defined by the National Bureau of Economic Research.

the private saving-investment surplus as the recovery progressed, analysts point to these budget deficits as the primary cause of the unprecedented external deficits observed to date.

But a closer examination of Chart 1 also suggests that prior to the 1980s, the private saving-investment balance tended to vary opposite to budget deficits *even apart from cyclical influences,* thereby producing no discernible cyclical or secular trend in the trade balance until the 1980s. Thus a change in the behavior of the private saving-investment balance, as well as rising budget deficits, apparently has contributed to the external deficits of the 1980s.

The question whether external deficits in the 1980s are the result of budget deficits, a change in the behavior of the private saving-investment balance, and/or other factors cannot be resolved simply by looking at the accounting relationships embodied in Equation 2. These variables respond to exogenous changes in fiscal and monetary policy as well as to other autonomous factors. Thus, to determine the effect of any of these variables on the external balance of the U.S., it is important to examine how they have behaved in response to exogenous changes in policy. Because changes in the private saving-investment balance have received relatively less attention, this article focuses on the implications of this relationship for the U.S. external deficit.

Abstracting from cyclical effects, the response of the private saving-investment balance to fiscal and monetary policy may be expressed as follows:

$$S - I = a_0 + a_1(T - G) + a_2 M \qquad (3)$$

Where (T − G) and M now refer to the exogenous behavior of the budget balance and the money supply, and a_0 contains all other factors.[5] (Note that the observed budget balance in Chart 1 is the sum of the exogenous budget balance and the endogenous response of the budget balance to all exogenous disturbances.) Equation (3) may be interpreted as a reduced form. The underlying structure may be motivated in terms of a standard Keynesian macroeconomic model of an open economy.

With this framework in mind, consider first the effects of an expansionary fiscal policy. Such a policy tends to raise income and interest rates; the rise in income tends to increase domestic saving, and the rise in interest rates tends to discourage domestic investment. Consequently, budget deficits tend to produce an offsetting rise in the private saving-investment balance, suggesting that a_1 is likely to be negative.

An expansionary monetary policy tends to lower interest rates, stimulating investment and income. The increase in income, in turn, stimulates saving. (In an open economy with floating exchange rates, lower interest rates tend to cause the currency to depreciate, stimulating net exports and causing a further increase in income.) Since both saving and investment tend to rise, the net impact of an expansionary monetary policy on the private saving-investment balance (the sign of a_2) is ambiguous.

The Response of Private Saving-Investment

Equations (2) and (3) imply that it is the interaction of fiscal and monetary policy and the private saving-investment balance that determines the external balance. Thus, the magnitude of the response of the private saving-investment balance to fiscal policy—that is, the magnitude of a_1—determines whether fiscal policy affects the trade balance. If $a_1 = -1$, fiscal deficits generally will not be associated with external deficits; however, fiscal deficits will be reflected in external deficits if $a_1 > -1$.[6] Unfortunately, the literature provides conflicting evidence on the magnitude of a_1.

Two well-known structural simulation models (Taylor, and Sachs and Roubini), and a recent study by Benjamin Friedman that estimates a reduced form model suggest that the private saving-investment balance does not fully offset budget deficits (that is, $a_1 > -1$). Taylor (1987) estimates a multi-country version of the Mundell-Fleming model[7] and finds that five years after the start of a simulated cut in government purchases "virtually all of the cut generates a rise in [national] saving, and about 3/4 of this rise in saving [is reflected in] an increase in net exports."[8] This suggests that a cut in the government deficit does not produce a fully offsetting reduction in the private saving-investment balance and is thus reflected largely in a reduction in the external deficit.

Similarly, using a dynamic general equilibrium simulation model of a six-region world economy, Sachs and Roubini (1987) argue that the combination of sharply higher budget deficits in the U.S. and sharply reduced deficits in Japan goes far to explain the movements of the external balance and exchange rates of the two economies.[9] Friedman's (1986) reduced-form estimates also suggest that budget deficits largely are reflected in external deficits.

In contrast to these three studies, a well-known study by Feldstein and Horioka (1980) found that national saving (T − G + S) was positively associated with gross private investment (full crowding out) in a cross-section sample of industrial countries. Subsequent time series analysis by

Obtsfeld (1986) and Frankel (1985) found a similar positive correlation between national saving and gross private investment in the U.S.[10] The results of the studies by Feldstein and Horioka, Obtsfeld, and Frankel suggest that $a_1 = -1$, or close to it.

There is also no agreement on the direct impact of changes in the money supply on the private saving-investment balance. Friedman finds that an increase in the ratio of money to GNP increases the ratio of private saving to GNP more than it increases the ratio of private investment to GNP,[11] thus reducing the external deficit (this suggests that $a_2 > 0$). In contrast, Darby, Gillingham, and Greenless (1987) find that a rise in the real money supply in the 1980s has tended to *increase* the external deficit[12] through its negative impact on private saving (that is, $a_2 < 0$). Similarly, Taylor finds that in the short run, an expansionary monetary policy tends to increase the external deficit, and, by implication, to reduce the private saving-investment balance.[13]

Several reasons may be offered for the conflicting results, including different specifications for models, variables, and econometric methods. Omitted variables may explain the differences in some cases and simultaneous equations bias in others. An alternative explanation is a change in the response of the private saving-investment balance to fiscal and monetary policy. This possibility has received relatively little attention, although studies reported by Darby (1987) and Darby, Gillingham, and Greenless (1987) suggest that changes in the behavior of the private saving-investment balance may have contributed to the external deficits of the 1980s.[14]

A change in the relationship between the private saving-investment balance and budget deficits might be expected, in view of two major developments that occurred in the 1970s. First, industrial countries shifted to floating exchange rates[15] and liberalized capital controls[16] in the first half of the 1970s. This process largely was completed by 1974, although restrictions on capital movements in the U.K. and Japan were not removed until 1979. As discussed more fully below, increased capital mobility and floating exchange rates could be expected to lower the offsetting response of the private saving-investment balance to budget deficits.

The second major development was the decision by the Federal Reserve in October 1979 to change its operating procedures for implementing monetary policy from reliance on an interest-rate instrument to the use of an aggregates instrument. Dewald (1982) finds evidence that during the earlier period, monetary policy tended to "accommodate" fiscal policy, in the sense that there was a positive relationship between money growth and fiscal deficits in the U.S. In particular, the acceleration in money growth and inflation in the 1970s appears to have been directly related to the near tripling of fiscal deficits to over one percent of GNP in the 1970s.[17]

As a result, rising budget deficits in the 1970s may have produced rising inflationary expectations. As discussed below, this may have discouraged foreign capital inflows and raised the offsetting response of the private saving-investment balance to budget deficits in the 1970s. However, once monetary policy changed and money growth and inflation apparently ceased to respond to budget deficits in the 1980s, foreign capital was more likely to flow in, thereby lowering the offsetting response of the private saving-investment balance to budget deficits in the 1980s.

II. The Response of the Private Saving-Investment Balance to Fiscal and Monetary Policy

To determine whether the response of the private saving-investment balance to fiscal and monetary policy has changed, regressions of the following form were run using seasonally-adjusted quarterly data:

$$S-I = b_0 + b_1 \cdot (T-G)_t + \sum_{i=0}^{8} b_{2+i} \cdot M2_{t-i} \qquad (4)$$

$$+ b_{11} \cdot GNPGAP_t + b_{12} \cdot INVMET_t$$

$$+ b_{13} \cdot DUM \cdot (T-G)_t + \sum_{i=0}^{8} b_{14+i}$$

$$\cdot DUM \cdot M2_{t-i} + b_{23} \cdot DUM \cdot GAP_t$$

where $S-I$ is the private saving-investment balance, $T-G$ is the government budget balance, GNPGAP is the gap between the middle expansion trend of GNP and actual GNP (a negative gap indicates a strong economy) as defined by the Department of Commerce,[18] and INVMET is the reciprocal of the middle expansion trend of GNP. All variables are scaled by the middle expansion trend of GNP.[19] The variables prefaced by DUM are slope dummy variables, and they correspond to values of $T-G$, M2, and the GNPGAP. Significant coefficients for b_{12}, b_{14+i}, and b_{22} would indicate a change in the response of the private saving-investment balance to fiscal policy, monetary policy, and cyclical fluctuations, respectively.

M2 was selected as the proxy for monetary policy because of the severe instability characterizing the demand for M1 in the 1980s. In view of possible simultaneous equation bias, an instrumental variable was used for contemporaneous M2 as well as for the contemporaneous budget balance.[20] A correction for serial correlation also was performed.[21]

A regression first was run over the period 1963:3–1979:4,[22] with slope dummies beginning in the first quarter of 1974, to ascertain whether there was any change in the relationship between the budget balance and the private saving-investment balance following the liberalization of capital controls and the shift to floating exchange rates. The results are reported in the first column of Table 1.

Table 1

Response of the Private Saving-Investment Balance to Fiscal and Monetary Policy

Explanatory Variables	M2 & dummies lags 0–8 63:3–79:4	M2 & dummies lags 0–4 63:3–87:4	No M2 dummies lags 0–4 63:3–87:4
Government Balance(1)	−0.809***	−0.875***	−0.884***
	(.213)	(.118)	(.119)
M2 (coeff sum)	−0.021**	−0.013***/**(2)	−0.011***/**(2)
	(.009)	(.005)	(.005)
GNP Gap	0.085	0.214***	0.231***
	(.06)	(.049)	(.049)
INVMET	1820.99**	1600.88***	1527.71***
	(524.89)	(337.91)	(330.84)
Slope dummies	74:1 − 79:4	80:1 − 86:4	80:1 − 86:4
Government Balance	0.345	0.550***	0.433**
	(.457)	(.208)	(.126)
M2 (coeff sum)	0.017**/*(2)	0.005	—
	(.009)	(.009)	
GNP Gap	0.208	0.286***	0.261**
	(.167)	(.078)	(.076)
Rho	0.369	0.251	0.264
Adj RSQ	0.884	0.814	0.814
PC	.371	0.496	0.465
Sum of Coefficients Gov't balance plus dummy			−0.451***

*** significant at 1 percent
** significant at 5 percent
* significant at 10 percent
Figures in parentheses are standard errors.

(1) Cannot reject hypothesis that coefficients are equal to −1.
(2) F-test on block of coefficients/t-test on sum of coefficients.

A second regression then was run over the period 1963:3–1987:4, with slope dummies beginning in the first quarter of 1980, to examine whether the response of the private saving-investment balance changed with the shift in monetary policy in the last quarter of 1979 and the further liberalization of capital controls in the U.K. and Japan. (The slope dummies for 1974–1979 were not included in this regression.) The results are reported in the second column of Table 1. Four and eight quarter lags on M2 were tested in both the first and second regressions. To select the best specification, Amemiya's prediction criterion (PC) was used.[23]

As can be seen in Table 1, both the fiscal and monetary policy variables are significant. The extent to which the private saving-investment balance offsets fiscal deficits did not decline following the liberalization of capital flows in 1974, but did decline after 1980. In addition, the slope dummies on M2 are significant after 1974, but not after 1980. To improve the fit of the second regression, the slope dummies for M2 after 1980 were eliminated, and the second regression was re-run. The results are reported in the third column of Table 1.

One potential objection to all these regressions is that there may be a lag in the response of the private saving-investment balance to the government budget balance as well as to monetary policy. The regressions were therefore re-run over the period 1963:3–1987:4, with four and eight quarter lags on M2 and on the government budget balance, respectively. However, in all cases, the third regression of Table 1 was superior to the alternative regressions according to the PC criterion.

The results of the third regression in Table 1 suggest that a one point increase in the budget balance brings about a 0.88 decline in the private saving-investment balance. In addition, the hypothesis that the private saving-investment balance fully offset the government budget balance up to 1980 cannot be rejected. There was no change in this relationship after 1974, and exogenous increases in the fiscal deficit apparently did not translate into external deficits up to the 1980s. The extent of the private saving-investment offset weakened by nearly 50 percent after 1980, and budget deficits came to be reflected in external deficits. Since changes in the response of the private saving-investment balance to fiscal policy appear to account for a significant part of the deterioration of the external balance in the 1980s, these results are interpreted more fully in the next section.

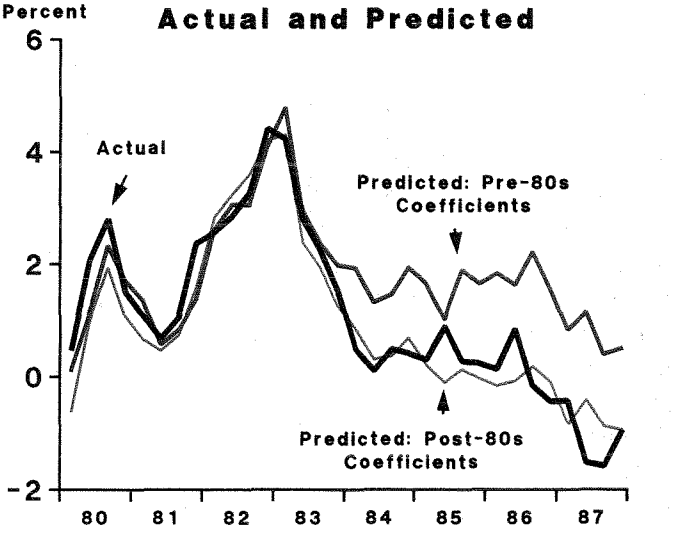
While these results assign a major role to the behavior of the private saving-investment balance in explaining the external deficits of the 1980s, they nevertheless imply that a reduction in budget deficits now would reduce the U.S. external deficit substantially, as long as the relation between the saving-investment and budget balance remains unchanged.

The results also indicate that cyclical effects were amplified in the 1980s. The private saving-investment balance appeared to be higher in the recessions of the 1980s and lower in the most recent expansion, which began in 1982, than was characteristic of earlier cycles. The stronger cyclical response is not easy to explain. It could reflect a stronger response of investment to income (a stronger accelerator effect) in the 1980s or the tendency for wealth effects to reduce private saving in the current expansion. Financial wealth has risen due to the accumulating government debt and the stock market boom of recent years.[24]

Finally, money exerts an independent influence on the private saving-investment balance. The sum of the coefficients on M2 consistently is negative and the t-statistic on this sum is significantly different from zero, indicating that an increase in the money-to-GNP ratio lowers the private saving-investment balance—and by this channel, the external balance. In particular, a rise in the M2/GNP ratio after 1982 has contributed to a lower private saving-investment balance, and therefore has tended to increase the external deficit.

As demonstrated in the Appendix, the significant and positive slope dummies on M2 after 1974 are consistent with the liberalization of capital flows in the early 1970s, although the net effect of the positive shift is apparently very small.[25] A greater degree of capital mobility will bring about a positive shift in the response of the private

Chart 2
Private Saving-Investment Balance: Actual and Predicted



saving-investment balance to monetary policy because it reduces the responsiveness of interest rates and investment to changes in the money supply, and increases the responsiveness of exchange rates, income, and therefore, saving to such changes.

To illustrate the implications of these results, Chart 2 compares the predicted path of the U.S. private saving-investment balance in the 1980s using the pre- and post-1980 coefficients. For reference, the actual path of the U.S. private saving-investment balance also is shown. Given the actual fiscal deficit, which averaged 3.1 percent of GNP between 1982:4 and 1987:4, the model predicts that the private saving-investment balance would have averaged two percent, rather than the 0.7 percent actually observed, had the pre-1980s' relationships prevailed. As a result, the external deficit would have averaged 1.3 percent of GNP over the period, rather than 2.5 percent.[26]

III. Interpreting the Results

The finding that the private saving-investment balance adjusted to fully offset changes in the budget balance until 1980 suggests a very limited degree of net international capital flows, which is in line with the results of the literature inspired by Feldstein and Horioka cited earlier. The absence of net international capital flows up to the early 1970s might be explained by restrictions on international capital movements, since such restrictions require the private saving-investment balance to adjust fully to cover the financing requirements of the public sector.[27] However, it is surprising that there was no change in the relationship between the private saving-investment balance and the budget balance after 1974 (that is, the slope dummy on the budget balance variable was not positive and statistically significant between 1974 and 1979).[28]

Theory suggests that liberalization of capital flows as well as the shift to floating exchange rates in the early 1970s should have reduced the offsetting response of the private saving-investment balance to fiscal policy. The reason is that in an open economy where capital flows freely and exchange rates float, the rise in domestic interest rates associated with government budget deficits should tend to attract foreign capital and limit the required adjustment in the domestic private saving-investment balance.

The adjustment in the private saving-investment balance should be muted by capital mobility for two reasons. First, foreign financing directly limits the extent to which government deficits reduce or "crowd out" investment. Second, under floating exchange rates, capital inflows cause the currency to appreciate. This increases the external deficit, which in turn reduces the stimulus budget deficits provide to income, thereby limiting the rise in saving. With capital mobility and floating rates, the limited response of the private saving-investment balance to budget deficits should ensure that the latter are reflected in external deficits.

However, if budget deficits raise inflationary expectations and thus increase uncertainty about the investment environment, increased capital mobility will not necessarily diminish the response of the private saving-investment balance to budget deficits. The positive correlation between money growth and budget deficits in the 1970s suggests that there also may have been a positive link between inflationary expectations and budget deficits. As demonstrated more formally in the Appendix, such a link has two effects that could influence the response of the private saving-investment balance to budget deficits.

First, an increase in inflationary expectations reduces money demand and could *lower* real interest rates even as budget deficits increase. In the standard analysis, such lower interest rates would discourage capital inflows and cause the dollar to depreciate. Huizinga and Mishkin (1986) provide evidence that rising inflationary expectations were in fact associated with falling *ex ante* real interest rates in the 1970s. The real trade-weighted value of the dollar also depreciated almost continuously between 1971 and 1980.

As in the case of a direct increase in the money supply, however, the currency depreciation stimulates income and saving, while the interest rate decline stimulates investment, so the net impact of excess money demand on the private saving-investment balance must be determined empirically. The negative coefficients for monetary policy in the previous regression suggest that the stimulus to investment from an excess demand for money is stronger than the stimulus to saving. Thus, the tendency for rising inflationary expectations to lower money demand apparently does not explain why the behavior of the private saving-investment balance did not change in the 1970s.[29]

However, rising inflationary expectations could have an additional effect on the behavior of the private saving-investment balance. Specifically, rising inflationary expectations may increase uncertainty about the investment environment, and thus raise the risk premium demanded on U.S. dollar assets. A rise in the risk premium, in turn, would discourage capital inflows, and cause the currency

to depreciate, even if domestic real interest rates do not fall. The currency depreciation would stimulate saving, but not investment spending in this case. Thus, a rise in the risk premium unambiguously would raise the offsetting response of the private saving-investment balance to budget deficits.

This analysis suggests that if inflationary expectations had not been rising in the 1970s, the impact of the liberalization of capital controls and the shift to floating exchange rates on the behavior of the private saving-investment balance would have been felt earlier. Instead, the impact of the liberalization of capital controls and the shift to floating exchange rates was felt only after 1979, when monetary policy changed and budget deficits no longer had the same influence on inflationary expectations. This break in the link between budget deficits and inflationary expectations led to capital inflows and an appreciating currency in the 1980s, as would have been expected, given enhanced international capital mobility.

IV. Conclusions

The unprecedented rise in the U.S. external deficit in the 1980s mainly is the result of the interplay of two factors. First, government deficits remained large in the expansion of the 1980s, rather than tending towards zero as they had during previous expansions. Second, the private saving-investment balance failed to offset the rising budget deficits as it had in the past.

The liberalization of capital movements in industrial countries and the shift to floating exchange rates in the first half of the 1970s had no perceptible effect on the offsetting response of the private saving-investment balance to fiscal policy until a major shift in monetary policy and a further liberalization of international capital movements occurred in 1979.

This paper offers a hypothesis that is theoretically consistent with the timing of the changes in the response of the private saving-investment balance to fiscal policy. By feeding back into inflationary expectations and increasing uncertainty about the investment environment, the tendency toward monetary accommodation of fiscal policy until 1979 (found by Dewald) reduced the willingness of foreigners to finance U.S.deficits and caused a continuing currency depreciation which stimulated a strong offsetting response of the private-saving investment balance to the budget balance. This curtailed the tendency for the external deficit to increase in response to rising budget deficits in the 1970s, notwithstanding the liberalization of capital restrictions and the shift to floating exchange rates in the early 1970s. Once monetary accommodation of fiscal policy ceased in the 1980s, the external balance deteriorated significantly. As the reduced-form specification used in this paper does not permit a direct test of this hypothesis, further research is needed.

If this interpretation is valid, the results presented here have important policy implications. Budget deficits pose a dilemma—a decline in the external balance can be averted by accommodating budget deficits with monetary policy and currency depreciation, but at the cost of high inflation and greater crowding out of domestic investment. Conversely, policymakers can avoid high inflation and reduce crowding out by refusing to accommodate budget deficits, but with a rising external deficit. Thus, if efforts to reduce the U.S. external deficit are to succeed without a resurgence of inflation, they must be accompanied by a reduction in budget deficits.

APPENDIX

The purpose of this Appendix is to demonstrate how the response of the private saving-investment balance to policy may be affected by inflationary expectations, as this mechanism is not usually discussed in the literature. It also shows that the direct response of the private saving-investment balance to monetary policy is more likely to be positive if capital mobility is high.

Rewrite equation (2) in the text as follows:

$$\underset{+}{S(Y)} - \underset{-\ +}{I(r,\ Y)} + T - G = \underset{+\ -\ +}{B(E,\ Y,\ Y^*)} \qquad \text{(A-1)}$$

The signs refer to the partial derivatives and:

Y = income
r = real domestic interest rate
E = domestic currency price of a unit of foreign currency (e.g. dollars per yen)
* = superscript for foreign variables
f = superscript for holdings by foreign residents

Equation (A-1) represents equilibrium in the goods market. The money market equilibrium equates real money supply to demand by domestic residents (assume foreign residents hold no domestic money):

$$M = \underset{+\ -}{L(Y,\ i)} \qquad \text{(A-2)}$$

where M is the real money supply, L is the demand for money. The nominal interest rate, i, is defined as the sum of the real interest rate and inflationary expectations:

$$i = r + \Pi \qquad \text{(A-3)}$$

The domestic bond market is in equilibrium when the real supply of domestic bonds equals the demand by domestic and foreign residents:

$$k^s = \underset{+\ +}{k(Y,\ i,\ i^*} + \underset{-}{\Pi} + \underset{+}{\phi)} + \underset{+\quad +\quad -}{k^f(Y^*, i - \Pi - \phi,\ i^*)} \qquad \text{(A-4)}$$

where k^s is the supply of domestic bonds, k(.) and k^f(.), respectively, are the domestic and foreign demand for domestic bonds, and ϕ is the risk premium. It is assumed that inflationary expectations, Π, also correspond with the expected rate of depreciation. For similar specifications up to this point see Dornbusch (1980) or Marston (1985).

Assume that the risk premium on dollar assets responds to rising inflationary expectations, that is:

$$\phi = \underset{+}{\phi(\Pi)} \qquad \text{(A-5)}$$

Assume further that monetary authorities respond to changes in fiscal policy by changing the *rate of growth* in the money supply.* If agents are aware of this, they will form inflationary expectations as follows:

$$\Pi = \underset{(-)}{\Pi(T - G)} \qquad \text{(A-6)}$$

The extent to which inflationary expectations respond to fiscal deficits will depend on the degree of monetary accommodation to fiscal policy.

The above is a Mundell-Fleming model that allows for imperfect asset substitutability and includes specific assumptions about monetary policy and expectations formation. The following simplifying assumptions have been made in obtaining a solution:

a. The country is small, which means that the ramifications of domestic policies on foreign income and world interest rates are ignored. The impact of these foreign effects on the U.S. economy apparently is small (see Sachs and Roubini and Taylor).

b. Short-term price rigidity is assumed. Except for specifying inflationary expectations, the dynamics of price adjustment are not spelled out.

c. The effect of exchange rate changes on the domestic price level is ignored. This implies that two effects of an exchange rate appreciation are ignored in the present analysis. These are: the improvement in the terms of trade and real income, which tends to *raise* saving; and the increase in real wealth brought about by an exchange rate appreciation, which tends to *reduce* saving. Wealth effects are important in principle, specifically in portfolio demand, but they are ignored to simplify notation. This does not affect our main conclusions.

d. The implications of a country's net creditor or debtor position similarly are ignored. As discussed in Frenkel and Razin (1987), if a country is a net debtor, the slopes of the LM and IS curves may change (the LM curve may slope downward, the IS curve becomes steeper), reversing a number of standard conclusions. For example, a tax increase may be expansionary, rather than contractionary.

Discussion of this type of model usually focuses on the implications of capital mobility and the exchange rate

regime for the ability of fiscal and monetary policy to affect income. Here we will focus instead on how inflationary expectations affect the responsiveness of the private saving-investment balance to fiscal and monetary policies.

The Solution

Substituting (6) into (3) and totally differentiating (2) to (4), we obtain:

$$\begin{bmatrix} s+m-i_y & -B_E & -I_r \\ -L_y & 0 & -L_i \\ -k_y & 0 & -(k_i+k^f_{i\text{-}\Pi\text{-}\phi}) \end{bmatrix} \begin{bmatrix} d_y \\ d_E \\ d_r \end{bmatrix} = \begin{bmatrix} -d(T-G) & 0 \\ L_i\Pi_{(T-G)}d(T-G) & -dM \\ \left(k_{i^*+\Pi+\phi} - k^f_{i\text{-}\Pi\text{-}\phi}\right)\phi_\Pi\Pi_{T\text{-}G}d(T-G) & 0 \end{bmatrix}$$

where s = marginal propensity to save
m = marginal propensity to import

and subscripts refer to partial derivatives.

The solution to the system is

$$\begin{bmatrix} d_y \\ d_E \\ d_r \end{bmatrix} = \frac{1}{\Delta}A' \begin{bmatrix} -d(T-G) & 0 \\ L_i\Pi_{(T-G)}d(T-G) & -dM \\ \left(k_{i^*+\Pi+\phi} - k^f_{i\text{-}\Pi\text{-}\phi}\right)\phi_\Pi\Pi_{T\text{-}G}d(T-G) & 0 \end{bmatrix}$$

where: $\Delta = \left((k_i + k^f_{i\text{-}\Pi\text{-}\phi})L_y - k_yL_i\right)B_E > 0,$

under plausible conditions, and

$$A' = \begin{bmatrix} 0 & -B_E(k_i + k^f_{i\text{-}\Pi\text{-}\phi}) & B_EL_i \\ -[L_y(k_i + k^f_{i\text{-}\Pi\text{-}\phi}) - L_ik_y] & -(s+m-i_y)(k_i + k^f_{i\text{-}\Pi\text{-}\phi}) - I_rk_y & (s+m-i_y)L_i + I_rL_y \\ 0 & B_Ek_y & -B_EL_y \end{bmatrix}$$

Fiscal Policy:

$$\frac{d_y}{d(T-G)} = \frac{-(1+\phi_\Pi)B_EL_i\Pi_{T-G}\left(k_i + k^f_{i-\Pi-\phi}\right)}{\Delta} < 0$$

$$\frac{dE}{d(T-G)} = \frac{L_y\left(k_i + kf_{i-\Pi-\phi}\right) - L_i k_y}{\Delta}$$

$$\frac{\left\{\left[\left(s+m-i_y\right)\left(k_i + kf_{i-\Pi-\phi}\right) + I_i k_y\right]L_i - \left(\left(s+m-i_y\right)L_i + I_i L_y\right)\phi_\Pi\left(k_i + kf_{i-\Pi-\phi}\right)\right\}\ \Pi_{T-G}}{\Delta}$$

$$\frac{dr}{d(T-G)} = \frac{B_E\Pi_{(T-G)}\left(k_y L_i + L_y\phi_\Pi(k_i + kf_{i-\Pi-\phi})\right)}{\Delta} \lessgtr 0$$

Where, to simplify notation, the following relation is used:

$$k_{i^*+\Pi+\phi} - kf_{i-\Pi-\phi} = -\left(k_i + kf_{i-\Pi-\phi}\right)$$

Inflationary expectations (reflected in the term $\Pi_{(T\text{-}G)}$) expand income by making it more likely that the currency will depreciate in response to fiscal deficits (that is, $dE/d(T-G)<0$). Two effects are at work here. First, the rising inflationary expectations in response to fiscal deficits lower real money demand. The resulting excess demand for money lowers real rates in the short run, tending to depreciate the currency, and stimulate net exports and income. Second, in addition to the effect of lower interest rates, the currency depreciates further because inflationary expectations raise the risk premium demanded by foreigners, thereby stimulating net exports and income even more.

The effect of inflationary expectations on money demand tends to lower real interest rates, while the effect of inflationary expectations on the risk premium tends to raise real interest rates. If the impact of inflationary expectations on the risk premium is sufficiently strong, the currency may depreciate even when domestic real interest rates are not falling, or perhaps even when they are rising. In a large economy such as the United States, it is likely that the effects of variations in the risk premium will be reflected largely in the exchange rate rather than in the interest rate.

The effect of an increase in the government surplus on the private saving-investment balance is therefore:

$$\frac{d(S-I)}{d(T-G)^2} = s\ \frac{dy}{d(T-G)} - I_r\ \frac{dr}{d(T-G)} \qquad (A\text{-}7)$$

$$-\left\{\frac{\left(s(k_i + kf_{i\text{-}\Pi\text{-}\phi}) + I_i k_y\right)L_i}{\Delta} + \frac{\phi_\Pi(k_i + kf_{i\text{-}\Pi\text{-}\phi})(sL_i + I_i L_y)}{\Delta}\right\}\ B_E\Pi_{T-G}$$

In the absence of international capital mobility, $d(S-I)/d(T-G) = -1$, because domestic saving must fully finance government deficits. However, in the presence of capital mobility and floating exchange rates, as assumed here, $d(S-I)/d(T-G) = 0$ if $\Pi_{(T-G)} = 0$. This is because neither income nor interest rates will increase in response to fiscal deficits, in the case where fiscal deficits do not affect inflationary expectations. Thus, in the absence of changes in inflationary expectations, capital mobility and floating rates imply that the offsetting response of the private saving-investment balance to budget deficits will decline. The intuition is discussed in the text.

Equation (A-7) shows that if $\Pi_{(T-G)}$ is negative, the response of the private saving-investment balance to fiscal deficits will not necessarily fall to zero even with capital mobility and floating exchange rates. The sign of the first

right hand side term, which reflects the impact of inflationary expectations on money demand, is ambiguous; the private saving-investment balance may rise or fall. In contrast, the second right-hand side term (multiplied by Φ_{Π}) is unambiguously negative. Thus, $d(S-I)/(T-G)$ will remain negative if the second right-hand side term is sufficiently large.

The text argues implicitly that the response of inflationary expectations, and particularly its impact on the risk premium in the 1970s, may have risen by enough to prevent $d(S-I)/d(T-G)$ from falling in absolute value in the 1970s. In the 1980s, $d(S-1)/d(T-G)$ fell because $\Pi_{(T-G)}$ fell to zero.

In the next section, it is shown that the conditions that determine the sign of the impact of monetary policy on the private saving-investment balance determine the sign of the first right hand side term of equation (A–7).

Monetary Policy

$$\frac{dy}{dM} = \frac{B_E(k_i + k^f_{i-\Pi-\Phi})}{\Delta} > 0$$

$$\frac{dE}{dM} = \frac{(s + m - i_y)(k_i + k^f_{i-\Pi-\Phi}) - I_r k_y}{\Delta} > 0$$

$$\frac{dr}{dM} = -\frac{B_E k_y}{\Delta} < 0$$

The effect of an increase in the money supply on the private saving-investment balance is

$$\frac{d(S-I)}{dM} = s\frac{dy}{dM} - I_r\frac{dr}{dM} = \frac{B_E\left[s\left(k_i + k^f_{i-\Pi-\Phi}\right) + I_r k_y\right]}{\Delta} \qquad (A\text{-}8)$$

In the text the impact of an expansionary monetary policy on the private saving-investment balance, and therefore the external balance, is ambiguous. For example, if the interest sensitivity of investment demand (I_r) is large, an expansionary monetary policy will lower the private saving-investment balance.

An increase in capital mobility [$(k_i + k^f_{i-\Pi-\Phi})$ increases in absolute value] means the term $(B_e I_r k_y/\Delta)$ becomes smaller, which implies that a monetary expansion is more likely to improve the external balance.** The reason is that a greater degree of capital mobility will tend to weaken the ability of monetary policy to influence domestic interest rates, and therefore, investment demand and the external balance. The finding that there was an increase in the impact of monetary policy on the private saving-investment balance after 1974 is consistent with the expected effect of liberalization of capital controls.

It has been assumed that an increase in the stock of money does not directly affect inflationary expectations; instead, expectations respond to the *growth* in the money supply associated with fiscal deficits. Inspection of equations (A–7) and (A–8) also confirms that if $d(S-I)/d(T-G)$ in equation (A–8) is negative, as found in the regressions in the text, the impact of inflationary expectations on money demand cannot explain why $d(S-I)/(T-G)$ did not fall in the 1970s.

Notes to Appendix

* *See Dewald (1982), who finds evidence of this type of accommodation between 1948 and 1980.*

** *The effect of a tax cut may differ from that of an increase in government spending in two ways: first, a tax cut will raise disposable income directly as well as indirectly. Second, a tax cut may increase money demand for any level of pretax income. This tends to reduce the expansionary impact of a tax cut on income. These effects are ignored in order to simplify the present discussion.*

ENDNOTES

1. The external balance measure used here is U.S. net foreign investment abroad, the measure which is conceptually most consistent with the use of equation 2. This measure is approximately equal to net exports of goods and services as measured in the national income and product accounts.

2. The middle-expansion trend of GNP is calculated by classifying each quarter into one of four cyclical phases: recession, recovery, middle expansion, and late expansion. The geometric mean of GNP during each middle expansion phase provides one observation of the trend GNP. The middle expansion begins when the level of real GNP passes its pre-recession peak and lasts 12 quarters unless a downturn occurs before 12 quarters have passed. In the latter case, the middle expansion ends at the cyclical peak just before the downturn. The advantage of this approach is that it reflects the path of actual GNP purged of cyclical movements and requires no assumption about potential GNP. See De Leeuw and Holloway (1983).

3. The statistical discrepancy between internal and external balances has been added to gross private saving. It is therefore reflected in the private saving-investment balance.

4. The cyclical patterns disguise certain trends in these variables and may provide a misleading picture of the relationships among the variables. For example, the unadjusted U.S. government budget deficit, illustrated in Chart 1, averaged 0.7 percent between 1976:1 and 1979:4 and turned into a surplus for a brief period. On a cyclically adjusted basis, however, the government budget was consistently in deficit, averaging nearly 2 percent of the middle expansion trend of GNP. The empirical analysis reported later controls for cyclical effects.

5. The observed budget balance, and by the accounting identity of (2), the external balance, are also the consequence of these same exogenous disturbances to fiscal and monetary policy.

6. A coefficient for a_1 of -1 could mean that an expansionary fiscal policy will produce a trade *surplus* because such an expansionary policy will tend to create an offsetting improvement in the fiscal balance.

7. The model assumes perfect capital mobility and perfect asset substitutability. Careful attention is paid to dynamics, and rational expectations in asset and labor markets is assumed. Sticky wages are modelled by staggered wagesetting. An earlier example of a structural analysis of the U.S. external balance as determined by internal balances is provided by Von Furtensberg (1980), who examines the domestic price and quantity determinants of three components of the net national saving rate (government saving, personal saving, and corporate saving) and two components of net domestic investment (fixed domestic investment and the rate of inventory change). A similar approach, which focuses on international as well as domestic determinants, is followed by Turner (1986) for the seven major OECD countries.

8. Taylor (1987) p. 15. Taylor performs a counterfactual experiment in which U.S. government spending grows less rapidly than it actually did starting in the first quarter of 1982, so that by 1986:1 real government purchases are lower than they actually were by an amount equal to 3 percent of real GNP. This roughly would balance the fiscal deficit, and result in a reduction in the outstanding stock of government bonds.

9. The model is related to intertemporal dynamic models of fiscal policy and solves for a full intertemporal equilibrium in which agents have rational expectations of future variables. Attention is given to intertemporal optimization and intertemporal budget constraints. In this respect, it differs from the simple Mundell-Fleming framework utilized in this paper. Obstfeld (1987) provides an analytic (as opposed to simulation) solution to this type of optimization problem.

10. For a study that includes developing countries see Dooley, Frankel, and Mathieson (1987). For a similar approach that treats investment as the exogenous, rather than the endogenous variable, see Sachs (1981).

11. This is consistent with the standard trade literature, recently summarized by Hooper and Mann (1987), who suggest that by bringing about a currency depreciation, a monetary expansion would tend to improve the external balance, presumably by increasing the private saving-investment balance as well as the budget balance. Friedman uses the detrended logarithm of the ratio of M1 to GNP as the monetary policy variable.

12. They argue that four years of erratic upward movements in real per capita M1, which reversed a secular decline, contributed significantly to a decline in saving.

13. To see this, recall equation (2), $S - I + T - G = B$. An expansionary monetary policy will always tend to increase $T - G$, because as income rises, tax revenue increases. If an expansionary monetary policy lowers B, it must be because $S - I$ has fallen. Note that given the neutrality of money in Taylor's model, in the long-run money has no effect on the external balance in his simulations ($a_2 = 0$). Sachs and Roubini also find that monetary policy is of little importance in influencing the external balance.

14. Darby, Gillingham, and Greenless argue that the reduction in the U.S. national saving rate in the 1980s, and the associated deterioration in the U.S. external balance, were caused in large measure by a decline in the personal saving rate. Darby reports preliminary studies that find a significant increase in investment demand in the U.S. over the period 1981–85. In Darby's view, such an increase in demand permitted investment to flourish in the 1980s, even though negative real U.S. interest rates in the 1970s turned positive in the 1980s. Darby argues that this upward shift in investment demand was due to reductions

in anticipated business taxes and greater confidence that the regulatory environment would not arbitrarily turn against business. Note, however, that the empirical approach of the present paper is closer to that of Friedman than that of Darby *et. al.* The theoretical interpretation of the results also differs from those in the studies conducted by Darby.

15. The U.S. shifted to floating exchange rates in March 1973. Except for a brief effort to strengthen a rapidly falling dollar in November 1978, the behavior of exchange rates apparently had little influence on U.S. monetary policy from March 1973 until the Louvre agreement of February 1987.

16. Capital controls, which were widely used in OECD countries after World War II, were liberalized in the first half of the 1970s following the adoption of generalized floating exchange rates and in response to the rapid growth of the Euromarkets in the 1960s, which tended to limit the effectiveness of such controls. Capital mobility probably had increased after the convertibility of European currencies was restored in 1958, but restrictions on capital flows largely remained effective throughout the 1960s. See OECD (1982).
In the case of the U.S., restrictions that were designed to prevent capital outflows largely were eliminated in January 1974. These were the Interest Equalization Tax (IET), the Voluntary Foreign Credit Restraint Program, and controls on direct foreign investment. More stringent controls on capital flows had been imposed from time to time in the 1960s. For example, at the beginning of 1968, President Johnson announced controls on outflows of capital by American businesses, banks, and other financial institutions. This included a requirement that no U.S. capital finance direct investment in other industrial countries. This action was taken in response to the deterioration in the U.S. external position.

17. Dewald finds evidence that money growth was positively related to fiscal deficits between 1948 and 1980. He estimates that over the period a unit rise in the ratio of the fiscal deficit to high employment output was associated with a rise in the growth of M2 of 0.4 percent.

18. See De Leeuw and Holloway (1983).

19. Because the variables are expressed in ratios, a significant constant term (b_0) indicates that the *level* (not the ratio) of the private saving-investment balance is related to the middle expansion trend of GNP. Furthermore, a significant coefficient on INVMET indicates that if the relationship between the private saving-investment balance and the budget deficit were expressed in levels rather than ratios, the constant term would be significant. To see this, assume 8 lags on M2. Suppose the true relationship in levels (not ratios to the middle expansion trend of GNP) is

$$S - I = c_0 + c_1 \cdot (T-G)_t + \sum_{i=0}^{8} c_{2+i} \cdot M2_{t-i} \qquad (5)$$

$$+ c_{11} \cdot GNPGAP_t + c_{12} \cdot GNPMET_t$$

$$+ c_{13} \cdot DUM \cdot (T-G)_t + \sum_{i=0}^{8} c_{14+i}$$

$$\cdot DUM \cdot M2_{t-i} + c_{23} \cdot DUM \cdot GAP_t$$

where GNPMET is the middle expansion trend of GNP. Then the relationship expressed as ratios will be the equation shown in the text. Note that $c_0 = b_{12}$ and $c_{12} = b_0$ in equation (4) in the text.

20. The first stage regression to construct an instrumental variable for the budget balance included the budget balance lagged 1 to 3 quarters and contemporaneous department of defense spending. It produced an adjusted R^2 coefficient of .81 and a D.W. statistic of 1.93. The first stage regression for M2 included a constant, the short-term nominal interest rate in the U.S. lagged 1 to 8 quarters and M2 lagged 1 quarter. The adjusted R-squared was .967, the Durbin-Watson statistic 1.4. Equation (4) in the text resembles one of the reduced form regressions performed by Friedman (1986). However, Friedman did not use an instrumental variables procedure.

21. The correction was implemented by running a regression with quasi-differenced data. The data were quasi-differenced with the rho coefficient estimated from the instrumental variables regression.

22. The sample begins in the first quarter of 1959. However, degrees of freedom were used up by various lag lengths tried in the second stage regressions. Lagged variables were also used in creating instrumental variables.

23. The PC criterion is a better indicator than the adjusted R-squared because it considers the losses associated with choosing an incorrect model. It thus imposes a higher penalty for adding variables than does Theil's adjusted R-squared. See Judge *et. al.* (1985), pp 865–866, 868.

24. However, attempts to introduce a stock market variable as an explanatory variable were not fruitful. The absence of a slowdown in the economy after the stock market decline of October 1987 also suggests that wealth effects are not very strong.

25. The marginal significance level is 10 percent, which is a weak basis for not rejecting the hypothesis that there was a shift.

26. The external balance is estimated by adding the actual government deficit and the predicted private saving-investment balance. Note that the actual government deficit may be seen as the sum of the exogenous contemporaneous government deficit used on the right hand side of the regressions and of the endogenous response of the budget balance to fiscal and monetary policy.

27. In terms of equation (3), $a_1 = -1$ or close to it, as there is no foreign financing of fiscal deficits.

28. It also should have increased the positive response of

the private saving-investment balance to monetary policy, which did occur, as there is a statistically significant and positive slope dummy variable for M2 for 1974–79.

29. This interpretation needs to be qualified because the regression allows for lags in the impact of the M2/GNP ratio. Furthermore, although the demand for M2 has been more stable than the demand for M1 in the 1980s, it is still not perfectly stable. The rise in the M2/GNP ratio may in some cases reflect a rise in money demand, particularly in the most recent expansion.

REFERENCES

Darby, Michael R. "The Shaky Foundations of the Twin Towers: The Current Account Deficit, Capital Account Surplus, and National Investment and Saving." Manuscript, October 2, 1987.

Darby, Michael R., Robert Gillingham and John S. Greenless. "The Impact of Government Deficits on Personal and National Saving Rates." U.S. Treasury Department. The Office of the Assistant Secretary for Economic Policy. Research Paper No. 8702. August 1987.

De Leeuw, Frank and Thomas M. Holloway. "Cyclical Adjustment of the Federal Budget and Federal Debt," *Survey of Current Business*, December 1983.

Dewald, William G. "Disentangling Monetary and Fiscal Policy," *Economic Review,* Federal Reserve Bank of San Francisco, Winter 1982.

Dooley, Michael, Jeffrey Frankel and Donald J. Mathieson. "International Capital Mobility: What Do Saving-Investment Correlations Tell Us?" *Staff Papers,* International Monetary Fund, V. 34, No. 3, September 1987.

Eichengreen, Barry. "Trade Deficits in the Long Run." Prepared for the Conference *The U.S. Trade Deficit—Causes, Consequences and Cures.* Federal Reserve Bank of St. Louis, October 23–24, 1987.

Evans, Paul. "Do Deficits Raise Interest Rates?" *Journal of Monetary Economics*, September 1987.

Feldstein, Martin. "Domestic Saving and International Capital Movements in the Long Run and the Short Run", *European Economic Review*, 21, 1983.

Feldstein, Martin and Charles Horioka. "Domestic Saving and International Capital Flows," *Economic Journal*, 90, 1980.

Feldstein, Martin and Douglas W. Elmendorf. "Taxes, Budget Deficits and Consumer Spending: Some New Evidence." National Bureau of Economic Research Working Paper No. 2355, August 1987.

Frankel, Jeffrey. "International Capital Mobility and Crowding Out in the U.S. Economy: Imperfect Integration of Financial Markets or of Goods Markets?" National Bureau of Economic Research Working Paper No. 1773, December 1985.

———. *The Yen Dollar Agreement: Liberalizing Japanese Capital Markets.* Policy Analyses in International Economics, No. 9. Washington D.C., Institute for International Economics, 1984.

Frenkel, Jacob A. and Assaf Razin. "The Mundell Fleming Model A Quarter of a Century Later: A Unified Exposition." *IMF Staff Papers*. Vol 34, No. 4, December 1987.

Friedman, Benjamin. "Implications of the U.S. Net Capital Inflow," National Bureau of Economic Research Working Paper No. 1804, January 1986.

Hooper, Peter and Catherine L. Mann. "The U.S. External Deficit: Its Causes and Persistence." Prepared for the Conference *The U.S. Trade Deficit–Causes, Consequences and Cures."* Federal Reserve Bank of St. Louis, October 23–24, 1987. Also, *International Finance, Discussion Papers* No. 316, Board of Governors of the Federal Reserve System, November 1987.

Huizinga, John and Frederic S. Mishkin. "Monetary Policy Regime Shifts and the Unusual Behavior of Real Interest Rates." *Carnegie Rochester Conference Series on Public Policy,* Volume 24, pp. 231–274. Amsterdam: Elsevier Science Publishers, 1986.

Judge, George, W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl and Tsoung-Chao Lee. *The Theory and Practice of Econometrics* (2nd edition). New York: John Wiley and Sons, 1985.

McKinnon, Ronald I. "The Exchange Rate and Macroeconomic Policy: Changing Postwar Perceptions," *The Journal of Economic Literature,* Vol XIX. (June 1981).

Marston, Richard C. "Stabilization Policies in Open Economies," in Ronald W. Jones and Peter B. Kenen (eds.) *Handbook of International Economics*, v. 2. Amsterdam: Elsevier Science Publishers, 1985.

Obstfeld, Maurice. "Capital Mobility in the World Economy: Theory and Measurement," *Carnegie Rochester Conference Series on Public Policy,* 24, 55–104. Amsterdam: Elsevier Science Publishers, 1986.

Obstfeld, Maurice. "Fiscal Deficits and Relative Prices in a Growing World Economy." Manuscript. May 1987.

OECD. *Controls on International Capital Movements.* Paris, 1982.

Sachs, Jeffrey. "The Current Account and Macroeconomic Adjustment in the 1970s," *Brookings Papers on Economic Activity.* 1, 1981.

Sachs, Jeffrey and Nouriel Roubini. "Sources of Macroeconomic Imbalances in the World Economy: A Simulation Approach," National Bureau of Economic Research Working Paper No. 2339. August 1987.

Taylor, John B. "The U.S. Trade Deficit, Saving-Investment Imbalance and Macroeconomic Policy: 1982–87." September 1987. Prepared for the Conference *The U.S. Trade Deficit–Causes, Consequences and Cures."* Federal Reserve Bank of St. Louis, October 23–34, 1987.

Turner, Philip P. "Savings, Investment and the Current Account: An Empirical Study of Seven Major Countries 1965–84," *Bank of Japan Monetary and Economic Studies,* Vol. 4., No. 2, October 1986.

Von Furtensberg, George M. "Domestic Determinants of Net U.S. Foreign Investment," *IMF Staff Papers*, December 1980.

Financial Intermediation, Monetary Policy, and Equilibrium Business Cycles

Carl E. Walsh and
Peter R. Hartley

Wide disagreement exists over the exact role that money plays in the economy and why money seems to matter. There is a related disagreement concerning the role played by financial intermediaries. This paper provides a discussion of alternative views of the role played by financial intermediaries in determining the impact of monetary policy. The emphasis is on the macroeconomic impact of intermediaries and the discussion is limited to equilibrium models of the business cycle.

University of California, Santa Cruz, and Federal Reserve Bank of San Francisco; and Rice University and the Centre for Policy Studies, Monash University. Editorial committee members were Ramon Moreno, Fred Furlong, and Ronald Schmidt.

Policymakers charged with responsibility for monetary policy take it as self evident that their policy actions have an impact on the real economy in the short-run. Professional economics journals, on the other hand, are filled with models of equilibrium business cycles that imply systematic monetary policies have no real effects. While most economists would agree that monetary actions can—and do—have real effects on the macro-economy, they disagree on the exact role that money plays in the economy and why money seems to matter.

Business cycle theories in the Keynesian tradition assume that monetary disturbances affect real output because wages and prices adjust slowly in the face of economic shocks. Changes in the nominal quantity of money generate changes in the real quantity of money—the nominal quantity adjusted for the level of prices—since prices are sticky. Fluctuations in the real supply of money then affect interest rates and aggregate spending.

In sharp contrast, equilibrium business cycle theories assume wages and prices adjust continually to ensure that markets are in equilibrium. In most equilibrium models, the real effects of monetary fluctuations are typically either nonexistent or arise only when individuals have incorrect information about the current stock of money.

Economists also disagree on the role played by financial intermediaries. Some economists incline to the view that financial intermediaries are a "veil" in the sense that they re-package financial assets but do not affect real savings or investment behavior. Others emphasize that financial intermediaries can have real effects on economic resource allocation. This divergence of opinion is significant for the bearing it has on the debate about the role of monetary policy. An understanding of the roles of both money and financial intermediaries is necessary for evaluating and designing both macroeconomic monetary policy and bank regulatory policy.

In this article, we discuss how the behavior of financial intermediaries—and that of banks, in particular—may have an influence on real economic activity and how, through its impact on banks, monetary policy influences economic activity.[1] The objective of this article is not to present a complete survey of recent developments in the economics of financial intermediaries. Rather, the article focuses on developments that promise to advance our

understanding of the roles played by both financial intermediaries and monetary policy. The emphasis is almost exclusively on the macroeconomic impact of intermediaries, and the discussion is limited to equilibrium models of the business cycle.[2] Specifically, this article examines some of the channels through which systematic monetary policy will have real effects even when prices adjust quickly.

Sections I and II examine the role played by the liability side of the banking sector's balance sheet. Bank deposits are an important component of the medium of exchange, and variations in the quantity of bank deposits may affect economic activity. Section I discusses one recent approach in the economics literature that allows some role for major disruptions in the banking sector to affect the economy, but in which monetary policy itself has no effect on real activity. Section II discusses other recent work that examines more closely the determinants of the demand for bank deposits and concludes that monetary policy actions may have real effects via their impact on bank liabilities.

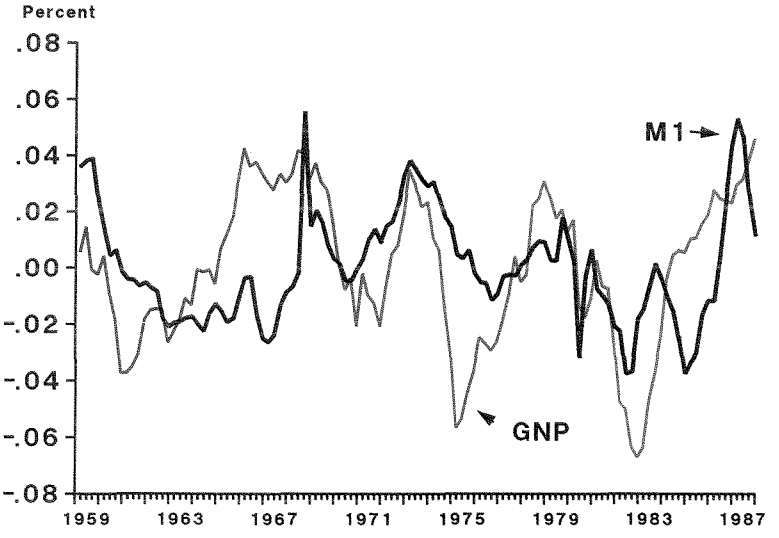
Section III turns to the asset side of the banking sector's balance sheet. Recent work that attempts to account for the economic role played by financial intermediaries is reviewed. One conclusion from this work is that variations in the supply of intermediated credit can affect the level of economic activity. The effect of monetary policy on the supply of bank-intermediated credit is then discussed. Conclusions are summarized in Section IV.

I. Transactions Services in Real Business Cycle Models

Charts 1 and 2 show that fluctuations in the money supply and fluctuations in the general level of real economic activity exhibit a high degree of association. In Chart 1, deviations of real GNP and M1 around trend are plotted using quarterly data for the period 1960.1 to 1987.4. Chart 2 plots the growth rates of M1 and real GNP. It is easy to see why monetary disturbances have played a major role in theories of the business cycle. Disagreements arise over whether this close association should be interpreted as evidence that monetary fluctuations have helped to cause business cycles, or whether *both* output and money supply movements are caused by nonmonetary economic disturbances.

The economics profession recently has seen the development of a body of work that employs stochastic growth models of competitive economies as a stylistic framework within which to study business cycles. For example, Kydland and Prescott (1982) and Long and Plosser (1983) studied business cycles as induced responses to real productivity shocks in models of economies that exclude any role for money. Because they ignore monetary factors as possible sources of cycles, these "real business cycle models" contrast strongly with models that focus on monetary disturbances as the major cause of cyclical fluctuations (for example, Lucas, 1975).

Chart 1
Deviations of Real GNP and M1 from Trend



In an important paper, King and Plosser (1984) introduce money into a real business cycle model. In their model, the sources of business cycles are entirely nonmonetary. Money does not cause cycles. But their model does predict a positive correlation between real output and monetary aggregates, like M1, that incorporate both outside money (the liabilities of the central bank) and inside money (the liabilities of the banking sector).

King and Plosser focus on the financial sector as a producer of transaction services that are used by firms and consumers in the process of production and the purchase of goods and services. Variations in the total output of goods and services generate positively correlated movements in the demand for transactions services. These changes in demand then induce similar movements in the actual supply of transaction services, leading to a positive correlation between measures of transaction services and real output during the course of a business cycle.

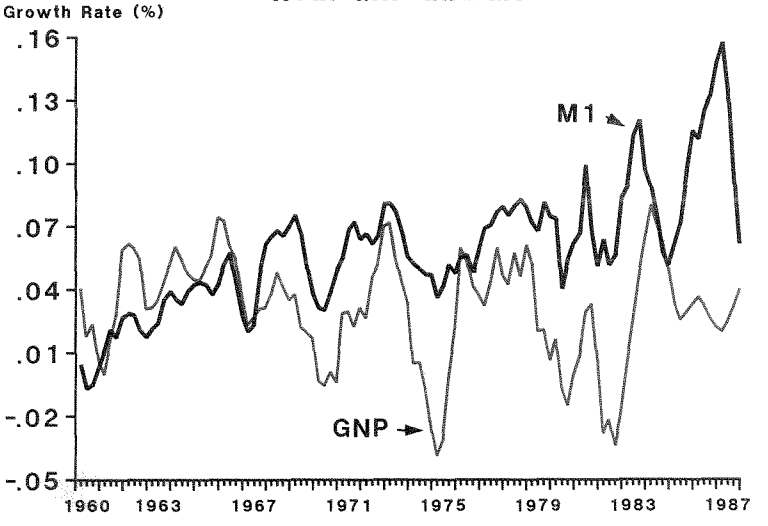
To account for the observed co-movements of real output and the stock of deposits at banks, it is necessary to provide some link between deposits and the quantity of transactions services produced by the financial industry. An economic rationale for such a tie is not straightforward, as Fama (1980) points out. King and Plosser skirt this issue by assuming that transaction services are linked directly to the stock of deposits held by the banking sector. Thus, by construction, the positive co-movement of output and transaction services translates into a positive co-movement between output and bank deposits. Inside money and output move together even though monetary factors have no causal role in generating business cycles.

This reverse causality argument—output causes money and money does not cause output—is not new. See, for example, Tobin (1971). But King and Plosser extend the argument by showing that their model implies measures of inside money, such as real bank deposits, should be more highly correlated with measures of real economic activity than are measures of outside money, and they present some evidence that such is the case in the U.S.

In the King and Plosser model, both individuals and banks care only about the real value of their asset holdings and the real value of the transaction services produced by the payments system. There is no money illusion; individuals do not care about nominal values. Consequently, changes in the nominal quantity of outside money simply result in proportional variations in the aggregate price level, leaving all real variables unaffected. Monetary policy, which varies the path of outside money, has importance only for the price level. Thus, a change in the stock of outside money does not affect the real value of bank assets or liabilities.

The neutrality of money holds in their model even if the variations in the path of outside money induce changes in the expected rate of inflation. An increase in the money supply that is viewed as only temporary, for example, would raise the current price level relative to future prices and generate expectations of future deflation. Such changes will cause nominal interest rates to adjust, but in the model

Chart 2
Growth Rates of Real GNP and M1



of King and Plosser these nominal rate adjustments leave all real rates of return on interest bearing instruments unchanged. Since outside money is non-interest bearing, a rise in the expected rate of inflation reduces its real return. The resulting fall in the demand for outside money would be eliminated by an immediate rise in the price level, thereby reducing the real supply of outside money.[3]

Unfortunately, the empirical evidence that King and Plosser present to support their model is weak. For example, correlations between outside money, inside money, and real economic activity are very dependent on the way monetary policy has been conducted. Outside money has never been used as a policy target by the Federal Reserve, but instead has been allowed to fluctuate in a manner consistent with the Fed's interest rate or monetary aggregate targets. Thus, the relative strength of the correlation of outside money with measures of real activity does not necessarily provide much information on King and Plosser's hypothesis. In fact, King and Plosser actually demonstrate that the exact relationship between the monetary base, the stock of currency, the stock of inside money, and the price level will depend on the presence or absence of reserve requirements and the particular aggregate targeted by the central bank.

Moreover, King and Plosser choose to ignore one channel in their model by which monetary variables could have effects on real economic activity. Because transaction services—the banking sector's output—are used as inputs into the production of other goods and services, a real shock to the banking sector can have real effects on production in other sectors. Thus, a major disruption of the payments system, such as occurred during the episode of massive bank failures in the early 1930s, would contribute to a fall in real output by lowering the quantity of an important input into production.[4] Bernanke's 1983 study of the impact of intermediation during the Great Depression provides some empirical evidence relevant to this issue. Bernanke showed that measures of the liabilities of failing banks enter significantly in real output equations estimated over the 1920s and 1930s, even when measures of the money supply also are included.[5]

If the transaction services provided by bank deposits are an important determinant of the general level of economic activity, then more mild fluctuations in the stock of bank deposits than those accompanying major bank failures also may produce general economic fluctuations. Substitution between bank deposits and other financial assets might then affect economic activity. King and Plosser ignore this channel from monetary variables to real variables on the grounds that it is likely to be unimportant empirically. Whether, in fact, such an effect is small enough to neglect will depend on the nonbank sector's demand for liquid assets.

II. Bank Liabilities and the Medium of Exchange

Until recently, economists studying the microeconomic foundations of the demand for transaction services concentrated on the demand for outside money (the liabilities of the central bank). In some of the earlier literature, such as the Baumol and Tobin transactions costs models, there was an explicit recognition that interest bearing assets were an alternative to non-interest bearing money as a store of value so that the interest yield on those assets represented the opportunity cost of using money to fund transactions. However, the models did not attempt to explain why non-interest bearing money was used to effect transactions, and in that sense, left the microeconomic foundations of the demand for money partly unexplained.

In any event, changes in expected inflation induced by monetary policy will lead market interest rates to adjust and cause substitution between interest bearing bank deposits (inside money) and non-interest bearing outside money. Consequently, the real impact of monetary disturbances may depend on the properties of bank liabilities as medium of exchange substitutes for outside money. This suggests that the liability side of the banking sector's balance sheet may be important in determining the impact of monetary policy.

In recent years, the overlapping generations model of Samuelson (1958) has been the most popular way of explaining the role of money in intermediating transactions. This model suggests that money is needed to facilitate spatially or temporally-separated transactions since any one generation is unable to arrange trades with all successive generations without the use of some "money-like" asset.[6] A difficulty with these models is that they leave unexplained the use of non-interest bearing outside money to finance transactions when interest bearing inside money also is available.

Bryant and Wallace (1984) appeal to legal restrictions on private intermediation to explain the co-existence of currency and interest bearing default-free bonds. Interest bearing default-free bonds are unsuitable for financing many transactions because of the bonds' large denominations. Legal restrictions prevent intermediaries from creating a better medium of exchange by issuing default-free small denomination claims to such bonds.

Bryant and Wallace use their model to examine the interactions between such legal restrictions and monetary policy. In effect, the legal restrictions and the use of both currency and bonds to fund transactions permit the government to levy a non-linear inflation tax. The real equilibrium their model achieves is not independent of either monetary policy or the institutional factors explaining the demand for currency and intermediary liabilities. However, these effects depend solely upon the nature of the legal restrictions on intermediary behavior and may not be intrinsic to all economies in which financial intermediaries issue inside money.

Another strand of the recent literature on the demand for monetary assets has focused on the cash-in-advance constraint model of liquidity. Drawing on a suggestion of Clower (1967), Lucas (1980) developed a formal model of the transaction services provided by money by assuming money balances were required to finance purchases of consumption goods. Goods could be exchanged for money and money for goods, but goods could not be exchanged for goods. Also, purchases were limited by the cash in hand at the time of purchase and could not be paid for by a subsequent exchange of interest bearing assets. In Lucas and Stokey (1983), the model was further elaborated to allow for two categories of goods. Some goods could only be purchased with cash, while other goods could be paid for with a subsequent exchange of interest bearing assets. In Lucas (1984), consumers can hold interest bearing state-contingent "securities" in addition to money. While the securities bear interest, they can be exchanged only at infrequent intervals and therefore are not very useful for financing consumption.

These ideas are developed further in Svensson (1985) and Hartley (1988). In Hartley's model, consumers can use either cash or interest-bearing "deposits" to purchase some goods, while other goods can only be purchased with cash. Consumers also can hold interest-bearing state-contingent securities. Cash and deposits are more liquid than securities in that only the former can be used at any time to effect purchases. Cash is more liquid than deposits because it can effect a wider range of purchases. While the liquidity return is highest for cash and absent for securities, the explicit interest payments are highest on securities and absent for cash. This inverse relationship between liquidity and interest yield suggests that changes in the nominal interest payments on deposits or securities will affect the demand for all three assets. As a result, monetary policy, by changing the anticipated rate of growth of base money, will have real effects by altering the demands for the different categories of financial assets.

Englund and Svensson (1986) examine banking in a general equilibrium model related to the Lucas, Svensson, and Hartley models. They show that changes in the credit multiplier, or "banking sector shocks," will have real effects, but one-time changes in the level of base money will affect only prices. In the macroeconomic model examined in Hartley and Walsh (1986), one-time changes in the level of the base money supply also are neutral, whereas temporary changes in the level of the base, or changes in its rate of growth, will have real effects by inducing substitution between holdings of inside and outside money.

Economists typically rule out the real effects of inflation-induced portfolio substitution (Tobin effects) on the grounds that they are likely to be unimportant empirically. For example, this is the position taken by King and Plosser. This is a reasonable assumption concerning direct substitution between money and capital on the part of the economy's wealth holders. However, the real effects of monetary shocks in Hartley and Walsh are driven by substitution among monetary assets such as currency and bank deposits. The elasticity of substitution among different monetary assets is likely to be large, even if the elasticity of substitution between portfolio holdings of capital and liquid assets as a whole is esentially zero.

While the microeconomic models that allow roles for both outside and inside money as mediums of exchange are far from complete, they all suggest that monetary policy could have significant aggregate real effects in equilibrium models of the business cycle. In general, these effects arise by altering the relative demands for different liquid assets. As consumers substitute between different liquid assets as mediums of exchange, interest rates are affected and the flow of "savings" to finance investment may be altered. Thus, even in equilbrium models of the business cycle—models in which prices freely move in response to perceived monetary disturbances—the role of bank liabilities in the payments system, and their substitutability for outside money plays a role in determining the impact of monetary policy on the real economy.

This discussion has focused on the liability side of the banking sector's balance sheet and on banks as producers of transaction services. Banks, however, also provide portfolio management services. Consumers hold deposits in order to gain access to the payments system and the transaction services banks provide, and these deposits represent claims against the assets held by the bank. Thus, it makes sense to consider the asset side of the banking sector's balance sheet and banks' role as suppliers of credit as a key channel of the impact of monetary policy. To understand this potential channel for monetary policy actions, it is necessary to understand the role of banks as suppliers of credit.

III. Increasing Returns, Financial Intermediaries, and Credit

Variations in the demand for currency and bank deposits may affect the total volume of bank liabilities and, as a result, lead to variations in the volume of bank lending. An increased demand for currency, for example, may produce a decline in bank lending. But while banks are distinguished by the transaction accounts they offer, banks are not the only intermediaries that supply credit, and the impact on the total supply of credit due to a reduction in bank lending will depend on how easily borrowers can replace bank loans with credit from nonbank sources. Changes in the volume of bank loans will have the greatest impact on economic activity when there is something "special" about bank credit.

This section begins with a discussion of an environment in which bank credit has no special characteristics that distinguish it from other sources of credit and in which financial intermediaries are unimportant for the determination of real economic activity. Then, some recent work that focuses on the role played by intermediaries is discussed. This work suggests there may be something special about bank credit. If this is the case, monetary policy is likely to affect economic activity through its impact on bank lending.

Suppose, as in the real business cycle model of King and Plosser, that banks face constant returns to scale as providers of portfolio management services. In this case, banks will care only about the management fees they earn, and not about the composition of the portfolio of assets they hold. Banks "simply cater to the tastes and opportunities of suppliers of securities and demanders of deposits. Thus, the real activity that takes place, the way it is financed, and the prices of securities and goods are not controlled either by individual banks or by the banking sector." (Fama 1980, p.48)

The size of the intermediation industry can undergo proportional expansions and/or contractions without having any effect on the relative prices of different assets. Asset prices and the financing of real economic activity are determined by the behavior of the economy's savers. They ultimately hold the economy's assets, whether they do so directly or indirectly by holding the liabilities of the financial intermediaries. When returns to scale are constant, shifts in the public's demand for the liabilities of intermediaries have no real effects; a reduced demand for these liabilities shrinks the assets held by intermediaries, but the affected assets can instead simply be held directly in the public's portfolio.

Reserve requirements present a potential problem in this framework. Reserve requirements force institutions subject to such regulations to hold some of their assets in the form of non-interest bearing assets or, in some countries, in the form of low interest rate government securities. Reserve requirements impose a tax on the banking sector, and drive a wedge between the return on banks' portfolio of assets and the return paid to depositors. With constant returns to scale, the demand for the portfolio management services of intermediaries subject to a reserve requirement would fall to zero. Individuals would prefer to hold assets directly rather than use the portfolio services of the intermediary. Banks would not be able to pass this reserve requirement tax on to the consumers of their portfolio management services.

King and Plosser argue that such a reserve requirement tax will lead to higher deposit service fees—that is, fees for access to the payments system will rise. As a result, the demand for bank deposits will fall, and the banking sector will contract. However, in the model of King and Plosser, variations in the size of the banking sector have no effect on the real allocation of credit or the financing of real economic activity. Either financial intermediaries not subject to reserve requirements will expand to offset the shrinkage of the banking sector or individuals will hold the liabilities of the economy's ultimate borrowers directly.

Fama (1985) recently has argued, however, that the reserve tax seems to be borne by bank borrowers, not by bank depositors. This implies that there is something "special" about bank loans [see also James (1988)]. Borrowers are willing to pay more to obtain a loan from a bank than from a nonbank source of credit. But this uniqueness of bank loans seems at odds with the view that intermediation is simply a veil behind which real activity is conducted, or that variations in bank-intermediated credit can be offset by the actions of nonbank intermediaries.

If bank credit is special, monetary policy actions that affect the size of the banking sector will have an impact on real economic activity. Thus, an increase in reserve requirements, for example, would shrink bank credit and force firms to switch to less attractive sources of funds. This would raise the net cost of funds in the economy and lead to a fall in aggregate investment activity.

To understand fully the role of banks in determining the effectiveness of monetary policy, it is necessary to examine more closely why bank loans might be special. In King and Plosser's model, asset choices of banks play no real economic role. A very different view of financial intermediaries emerges from another body of recent reserarch which includes papers by Boyd and Prescott (1986), Bernanke (1983), Bernanke and Gertler (1986), Stiglitz and

Weiss (1981), and Williamson (1986a, 1986b). These papers all attempt to provide economic explanations for the endogenous development of such institutions as financial intermediaries (both bank and nonbank).

Two characteristics of economic transactions are sufficient to generate the presence of intermediaries: asymmetric information and increasing returns to scale. The exact manner in which these two characteristics interact has been modelled differently by different authors. For example, Williamson (1986b) develops a model in which entrepreneurs have access to a technology that requires a fixed investment and yields a random real return, the expected value of which is known by both borrowers and lenders. The actual realization of the random return is known (*ex post*) to the entrepreneur, but other individuals can obtain information on the realized return only by incurring a fixed cost to monitor each project. The presence of fixed monitoring costs makes it costly for individual investors to attempt to diversify by lending to many different borrowers. Moreover, the projects are assumed to be sufficiently large relative to individual wealth that entrepreneurs must gain access to the savings of several individuals in order to carry out their investment projects.

This rudimentary framework is sufficient to generate a role for intermediaries. Since the project's actual return is known only by the entrepreneur, in the absence of monitoring, the entrepreneur always has an incentive to report a low return to his creditors and abscond with the profits. To prevent this behavior, each of the individual investors who finances a project must incur the monitoring cost. A large intermediary, on the other hand, can finance a large number of projects and incur the fixed cost of monitoring only once for each project. The intermediary is able to exploit the increasing returns to scale implicit in the fixed-cost monitoring technology. In addition, the intermediary's ability to fund a large number of projects permits diversification of nonsystematic risks. If there is no systematic risk, a large intermediary can offer a certain return to its depositors.[7]

Information asymmetries also give rise to debt contracts between lenders and entrepreneurs. In Williamson's model, the optimal contract can be shown to involve a fixed payment to the lender if the project return exceeds some critical value R*. If the actual return is less than R*, the lender receives the entire return. In other words, the borrower pays a fixed rate of interest on the loan if the return exceeds R*; otherwise the entrepreneur declares bankruptcy and the intermediary recovers the entrepreneur's assets, which will be worth less than R*. This contract minimizes monitoring costs since the entrepreneur has no incentive to lie if the actual return is greater than R*.

The introduction of financial intermediaries in the presence of asymmetric information and monitoring costs leads to increasing returns to scale from intermediation. In contrast to the view of intermediation as a constant-return-to-scale industry, increasing returns imply that the level of intermediation has an impact on real activity, the way that activity is financed, and the prices of securities and goods. This is particularly apparent if the equilibrium involves credit rationing (Williamson, 1986a).

Although these models help to explain why intermediation matters, they do not explain why *banks* might be special and therefore, why monetary policy might have real effects. One reason is that banks are both lenders and providers of transaction services. Banks have informational advantages that result in lower monitoring costs because they simultaneously lend to and maintain the transaction accounts of firms. The firm's transactions account provides the bank with low cost information about the firm. A nonbank intermediary lacking this source of information faces higher monitoring costs. In this case, banks are able to supply credit more efficiently than can other intermediaries. Consequently, if monetary policy affects the size of the banking sector, it also will affect the level of real economic activity.

While the economic role of intermediaries seems more fully developed in the asymmetric information literature, the real business cycle research has, somewhat paradoxically, provided a much more detailed analysis of the impact of monetary policy.[8] One attempt to bridge this gap is developed in Hartley and Walsh (1986), which supplements a conventional *ad hoc* macroeconomic model with a banking sector that makes loans to finance real investment spending. This framework permits the study of the macro implications of intermediation when intermediation matters. They show that monetary policy has real effects when changes in expected inflation induce substitution between bank liabilities and non-interest bearing outside money. Equal changes in all nominal interest rates (in order to restore expected real rates) alter the relative demands for non-interest bearing outside money and interest bearing inside money. Both the market for bank deposits and the market for outside money are affected, and adjustments in the price level cannot restore equilibrium to both markets simultaneously. As a result, real interest rates must adjust. Movements in the rate on bank deposits then lead to changes in the supply of bank loans and bank loan rates that affect the level of real economic activity. Unlike the case considered by King and Plosser in which returns to

scale in intermediation are constant, changes in the quantity of bank credit are not fully offset by changes in credit supplied by nonbank intermediaries or by changes in direct lending by individuals.

Additional channels through which monetary policy can affect real activity arise when bank liabilities are subject to reserve requirements. By reducing the nonbank sector's demand for outside money, an increase in expected inflation increases the supply of reserves available to the banking sector. Since reserves can be viewed as an input in the intermediation process under a fractional reserve system, an increase in expected inflation allows the banking sector to expand the supply of loans. This reduces the equilibrium loan rate and leads to a rise in real investment.

Shocks to the banking sector have effects on the level of real economic activity, then, because they affect the supply of loans. In Williamson's model of intermediation, for example, disturbances work through the asset side of the banking sector's balance sheet. In contrast, disturbances to banks in King and Plosser's model can have real effects only if they influence the provision of transaction services. Real effects arise, not because of variations on the asset side of banks' balance sheets, but because of variations on the liability side due to the role of bank deposits as a means of payment.

Several recent attempts have been made to determine whether it is bank credit (the asset side) or money (bank liabilities) that matters for real economic activity. Bernanke (1983) found that for the 1920s and 1930s money had effects on real output even after controlling for credit. Empirical evidence from post-war data is reported by King (1986) and Bernanke (1986). King finds little support for the role of credit as the transmission mechanism for monetary policy. In vector autoregressions (VARs) that include real GNP, demand deposits, and various measures of bank loans, demand deposits generally account for a much higher fraction of the variance of GNP forecasts errors than do any of the loan variables. Since King's measure of money—demand deposits—is a measure of inside money, these results seem most consistent with the real business cycle view.

Bernanke (1986) obtained somewhat different results when he used a structural model to identify underlying money and credit shocks in a VAR that included, in addition to M1 and a measure of credit, real GNP, real defense spending, and the monetary base. Based on a decomposition of the output forecast error variance, credit shocks appeared to be much more important than shocks to the monetary base (outside money). M1 and credit shocks were of roughly equal importance. These results make it clear that few generally agreed upon empirical regularities exist in this area.

IV. Conclusions

Recent research in monetary economics that has focused on the role of information asymmetries and the costs of monitoring provide an improved understanding of the role of financial intermediaries. This research highlights three characteristics of intermediaries that seem of special importance from the perspective of understanding the role played by monetary policy in equilibrium models of business cycles. First, when intermediation is modeled as a constant-returns-to-scale activity (as in the real business cycle model of King and Plosser), the asset side of the banking sector's balance sheet is irrelevant for real economic activity. Variations in bank-intermediated credit are offset by other intermediaries or by direct portfolio adjustments on the part of individuals.

Second, the characteristics of bank liabilities as a means of payment play a role in determining the impact of monetary policy. Monetary policy can induce individuals to substitute between bank deposits and outside money. These portfolio shifts will affect relative rates of return and real economic activity. The importance of the portfolio adjustments caused by changes in the relative yields of bank deposits and outside money will depend on the transaction properties of currency and bank deposits, and the characteristics of the payments system.

Third, if bank loans are special, perhaps due to the information efficiencies attributed to the banking sector's role as a provider of both credit and transaction services, then variations in the banking sector's aggregate lending will have an impact on real economic activity. Monetary policy will influence the real economy through its influence on the supply of bank loans. Variations in the path of outside money that induce changes in expected inflation will result in nominal interest rate adjustments, but such adjustments generally will affect real rates and will thereby affect the supply of bank loans.

If financial intermediaries essentially form a veil behind which real activity takes place, the resolution of many of the issues faced by economic policymakers is quite simple. If real activity, and the way it is financed, is independent of the actions of financial intermediaries as Fama (1980) and King and Plosser (1984) assume, then there would appear to be no justification on monetary policy grounds for any special regulation of the banking sector.[9] The appropriate conduct of monetary policy in such an environment also is

straightforward. Since variations in the monetary base have no effect on real variables, the monetary authority need concern itself only with achieving price stability.

However, if bank loans or deposits are in some sense special, then the optimal design of policy becomes a more complicated task. From both sides of the banking sector's balance sheet there seem to be good theoretical reasons to believe monetary policy disturbances will not be neutral, even in equilibrium models of the business cycle. Policy analysis requires a better understanding of the role of both bank lending and bank provision of the medium of exchange. Without such an understanding, we are unable to evaluate alternative policy proposals.

ENDNOTES

1. By "banks," we mean financial intermediaries whose liabilities provide transaction services.

2. For a more general summary of the real effects of monetary policy, see Blanchard (1987).

3. This superneutrality result does not strictly hold in a model like King and Plosser's which incorporates an endogenous labor supply decision unless the labor supply decision also depends only on ex-ante real interest rates. However, King and Plosser ignore this potential effect as empirically unimportant.

4. For a model of bank runs, see Diamond and Dybvig (1983).

5. As will be discussed in Section III, Bernanke's evidence also is consistent with the view that it is bank lending that is the key channel through which banking disturbances affect real economic activity.

6. Assets used to carry out intergenerational trades need not, however, bear much resemblance to money. For some examples from nonmonetary economies, see Walsh (1983).

7. Since the large intermediary can earn a certain rate of return through diversification, the type of asymmetric information problem between bank and depositors analyzed by Leland and Pyle (1977) does not arise.

8. A recent paper by Williamson (1987) attempts to incorporate his earlier work on intermediaries into a real business cycle model.

9. For a discussion of banks and regulatory policy, see Furlong and Keeley (1988).

REFERENCES

Bernanke, Ben S. "Non-Monetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review,* 73 (3), June 1983, 257-276.

_____."Alternative Explanations of the Money-Income Correlation," in K. Brunner and A.H. Meltzer (eds), *Real Business Cycles, Real Exchange Rates and Actual Policies*. Carnegie-Rochester Conference Series on Public Policy, 25, Autumn 1986, 49-99.

Bernanke, Ben and Mark Gertler. "Banking and Macroeconomic Equilibrium," *Discussion Paper* #108, Woodrow Wilson School, Princeton University, February 1986.

Blanchard, Oliver. "Why Does Money Affect Output? A Survey," *N.B.E.R. Working Paper* No. 2285, June 1987.

Boyd, John H. and Edward C. Prescott. "Financial Intermediary-Coalitions," *Journal of Economic Theory,* 38 (2), April 1986, 211-232.

Bryant, John and Neil Wallace. "A Price Discrimination Analysis of Monetary Policy," *Review of Economic Studies,* 51, 1984, 279-288.

Clower, Robert W. "A Reconsideration of the Microfoundations of Monetary Theory," *Western Economic Journal,* 6, 1967, 1-9.

Diamond, Douglas W. and Philip H. Dybvig. "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy,* 91 (3), June 1983, 401-419.

Englund, Peter and Lars E.O. Svensson. "Money and Banking in a Cash-In-Advance Economy," *International Economic Review,* forthcoming.

Fama, Eugene F. "Banking in the Theory of Finance," *Journal of Monetary Economics,* 6 (1), January 1980, 39-57.

_____."Financial Intermediation and Price Level Control," *Journal of Monetary Economics,* 12 (1983), 1-25.

_____."What's Different About Banks?," *Journal of Monetary Economics,* 15 (1), January 1985, 29-39.

Furlong, Fred and Michael Keeley. "Bank Regulation and the Public Interest," *Economic Review,* Federal Reserve Bank of San Francisco, September 1986, 55-71.

Hartley, Peter R. "The Liquidity Services of Money," *International Economic Review,* February 1988, 1-24.

Hartley, Peter R. and Carl E. Walsh. "Inside Money and Monetary Neutrality," *N.B.E.R. Working Paper* No. 1890, April 1986.

James, Christopher. "Some Evidence on the Uniqueness of Bank Loans," *Journal of Financial Economics*, forthcoming.

King, Robert G. and Charles I. Plosser. "Money, Credit, and Prices in a Real Business Cycle," *American Economic Review*, 74 (3), June 1984, 363-380.

King, Stephen R. "Monetary Transmission: Through Bank Loans or Bank Liabilities?" *Journal of Money, Credit and Banking*, 18 (3), August 1986, 290-303.

Kydland, Finn E. and Edward C. Prescott. "Time to Build and Aggregate Fluctuations," *Econometrica*, 50 (6), November 1982, 1345-1370.

Leland, Hayne and David Pyle. "Informational Asymmetries, Financial Structure, and Financial Intermediaries," *Journal of Finance*, 32 (2), May 1977, 371-387.

Long, John B., Jr. and Charles I. Plosser. "Real Business Cycles," *Journal of Political Economy*, 91 (1), February 1983, 39-69.

Lucas, Robert E., Jr. "An Equilibrium Model of the Business Cycle," *Journal of Political Economy*, 83 (6), December 1975, 1113-1144.

_____."Equilibrium in a Pure Currency Economy," *Economic Inquiry*, 18, April 1980, 203-220.

_____."Money in a Theory of Finance," *Carnegie-Rochester Conference Series on Public Policy*, 21, 1984, 9-46.

Lucas, Robert E., Jr. and Nancy L. Stokey. "Optimal Fiscal and Monetary Policy in an Economy Without Capital," *Journal of Monetary Economics*, 12, July 1983, 55-93.

Samuelson, Paul A. "An Exact Consumption-Loan Model of Interest With and Without the Social Contrivance of Money," *Journal of Political Economy*, 66, December 1958, 467-482.

Stiglitz, Joseph E. and Andrew Weiss. "Credit Rationing in Markets With Imperfect Information," *American Economic Review*, 71 (3), June 1981, 393-410.

Svensson, Lars E.O. "Money and Asset Prices in a Cash-in-Advance Economy," *Journal of Political Economy*, 93, October 1985, 919-944.

Tobin, James. "Money and Income: Post Hoc Ergo Propter Hoc?," reprinted in *Essays in Economics, Vol. 1: Macroeconomics*, 1971, Chapter 24.

Walsh, Carl E. "Savings in Primitive Economies," *American Anthropologist*, 85 (3), September 1983, 644-650.

Williamson, Stephen D. "Costly Monitoring, Financial Intermediation, and Equilibrium Credit Rationing," *Journal of Monetary Economics*, 18 (2), September 1986a, 159-179.

_____."Increasing Returns to Scale in Financial Intermediation and the Non-Neutrality of Government Policy," *Review of Economic Studies*, 53 (5), October 1986b, 863-875.

_____."Financial Intermediation, Business Failures, and Real Business Cycles," *Journal of Political Economy*, 95 (6), December 1987, 1196-1216.

Banks Affiliated with Bank Holding Companies: A New Look at their Performance

Randall Johnston Pozdena

The bank holding company (BHC) form of organization has a number of advantages for banking firms. It also is the form that many are recommending be used to enforce corporate separation of traditional banking from expanded banking activities. This paper examines the influence of BHC affiliation on bank behavior. The literature on this subject is large, but has ignored an important potential source of bias. The measured effects of BHC affiliation generally are larger when this bias is treated statistically using a technique described in the paper. BHC-affiliated banks do appear to behave differently than their non-affiliated counterparts, a finding that does not augur well for using this organizational form to isolate a bank from the effects of nonbank activities.

Assistant Vice President, Banking and Regional Studies, Federal Reserve Bank of San Francisco. The author wishes to thank William M. Robertson and Rachel A. Long for their excellent research assistance. Editorial committee members were Michael Keeley, Reuven Glick, and Ronald Schmidt.

The bank holding company (BHC) form of organization has a number of advantages for banking firms, but analysts have argued that such an organizational form also may lead to changes in the behavior of banks affiliated with BHCs. In the 1960s and 1970s, the rate of formation of BHCs was very rapid, leading to increased concern that this organizational form would have adverse effects on bank performance.

This interest in the effects of BHC affiliation on bank performance has been revived recently as part of the debate over the expansion of bank powers. Many are recommending that expanded powers be placed in nonbank subsidiaries of bank holding companies as a means of insulating the bank from any risks arising from those new activities. The assumption is that if the new, nonbanking activities are "corporately" separate from banking activities, the behavior and financial soundness of the bank will be unaffected. An examination of the effect of BHC affiliation on the performance of banks may shed some light on this debate. To the extent that affiliation with a bank holding company affects bank behavior, expectations about the effectiveness of the BHC structure in insulating banks from other activities in the BHC may be too sanguine.

The methodology for examining the influence of BHC affiliation has been quite straightforward. Analysts have compared the income and portfolio characteristics of affiliated banks with those of banks that are not affiliated with a BHC. To ensure that other characteristics of the banking organizations do not bias the comparisons, various statistical control methods have been used. Most commonly, each affiliated bank is "matched" with an unaffiliated bank in size, location, or other attributes. Any differences in performance are then attributed to the affiliation status of the banks. Alternatively, econometric techniques have been employed to control for the diverse characteristics of affiliated and non-affiliated banks. Both types of studies have found important differences in the behavior of BHC-affiliated and non-affiliated banks.

These analyses implicitly assume that, except for their organizational form, affiliated and non-affiliated banks are identical. If they really are identical, however, why are some banks part of BHCs and others not? It seems likely that there is some tendency for self-selection processes to bias simple comparisons of the behavior of affiliated and

non-affiliated banks. Banks that choose to become part of holding companies may have more aggressive management, for example. This may influence observed performance, and simple comparisons with non-affiliated banks will detect the differences. In this case, it may be incorrect to attribute the cause of these differences to the affiliation status. With the renewed importance of understanding how banks behave in different organizational contexts, it would be useful to reexamine the behavior of affiliated banks and to correct, if possible, for the effects of self-selection processes.

The purpose of this paper is to explore the possible influence of self-selection bias on the typical findings regarding the behavior of BHC-affiliated banks. By using simple techniques to control for self-selection in the affiliation decision, I obtain results that differ from many traditional findings. Some of my findings are more consistent with the theory of why banks affiliate with BHCs in the first place.

In the first section, the reasons why banks choose to affiliate with a BHC and the theoretical implications for bank behavior are reviewed briefly. In Section II, the conventional techniques for evaluating the effect of BHC affiliation are examined, along with the findings that such studies have produced. This section also presents the concept of self-selection bias and discusses the various methods to control for such a bias. In Section III, a statistical control technique is applied to data from a sample of western banking organizations. The paper concludes with a summary and discussion of the policy implications of this research.

I. Bank Holding Company Affiliation: The Economic Implications

To understand why affiliation of a bank with a BHC might affect the bank's behavior, it is important to discuss the motivations for BHC affiliation. These motivations involve both operational and tax advantages of the BHC form of organization and explain formations of both one-bank and multi-bank holding companies.

Motives for BHC Formation

The numerous operational advantages of BHC affiliation derive largely from distortions introduced by regulation and law. First, the activities of non-affiliated banks traditionally have been restricted by regulation to endeavors related to conventional banking business. One way a banking organization may expand its range of activities is to affiliate with a bank holding company.[1] A bank holding company may engage in a variety of activities through the nonbank affiliates of the bank; the affiliated bank thus may gain advantages from joint marketing or production of services with these subsidiaries.

Second, banking organizations structured as holding companies also can avoid some of the laws that restrict branching in certain states. By acquiring individual banks and maintaining them as separate subsidiaries of a BHC, such a banking organization and its bank subsidiaries may be able to enjoy geographical portfolio diversification, economies of scale, and other benefits that accrue to branch bank structures. In fact, the multi-bank holding company structure is common in states with laws that restrict branching by individual banks.

Third, bank holding company affiliation affords a banking organization greater flexibility in financing its activities. For example, shares of equity in holding companies are more liquid than shares in individual banks. Banks generally may not repurchase their own stock, though a BHC may. By forming a one-bank holding company, therefore, the shareholders of an existing bank effectively obtain increased marketability of their assets. The debt securities of the BHC also enjoy favorable reserves treatment compared to the same securities of a bank. Specifically, BHC non-deposit debt is not subject to reserve requirements, whereas similar obligations of a bank are.

Affiliation with a BHC, particularly prior to 1982, also was used as a mechanism to avoid capital regulation at the bank level. Specifically, the minimum proportion of equity (vs. debt) used to finance a bank's assets is regulated as a means of limiting bankruptcy risk in the banking system. Through means of a bank holding company structure, the affiliate bank's equity (capital) can be provided partly by the BHC. To the extent that the BHC funds its equity investment in the bank via issuance of debt at the parent level, the organization is able, in effect, to avoid leverage constraints imposed at the bank level. This practice, known as "double leverage," is defined by the Federal Reserve as the ratio of the equity in banks and nonbank subsidiaries to the equity in the parent BHC. The ability to double lever equity investments in a bank thus is a potential advantage of the BHC form of organization.

Since 1982, bank regulators have taken steps to limit this practice by coordinating capital regulation of banks and BHCs. Both the bank and the consolidated BHC (that is, the bank, the parent BHC, and the nonbank subsidiaries, with their interorganization obligations netted out) must

maintain the same ratio of capital to total assets. This does not completely eliminate opportunities for double leverage of the bank's capital, however, since capital in nonbank subsidiaries can be manipulated to create the appearance of more capital in the bank (although regulators try to monitor nonbank capital).[2] In addition, there are differences in the treatment of goodwill and certain types of debt on the books of the BHC versus those of the bank that tend to have the effect of creating double leverage opportunities. Also, for banks under $150 million in assets, up to 300 percent double leverage is permitted by regulation. Thus, the BHC form of organization still may be perceived as offering opportunities to loosen binding bank capital regulation.

Double leverage also creates a tax-related incentive for using the bank holding company form of organization. It involves the tax treatment of dividends generated by a banking organization. If the bank is owned directly by private shareholders, dividends paid to these shareholders are non-deductible expenses of the bank, taxable to the shareholder at his personal tax rate. If the bank, instead, is owned by a bank holding company, 85 to 100 percent of the dividends are deductible at the bank level (that is, they may be passed essentially tax-free to the parent BHC). To the extent that the parent BHC can use debt to finance its activities, these "upstreamed" dividends can be converted, in effect, to deductible interest expenses. Thus, an investor desiring to finance banking activities with a given proportion of debt and equity enjoys better tax treatment if he does so through a holding company rather than through direct, private ownership of the bank.

There are potential disadvantages to BHC affiliation as well, since the holding company form of organization is more complex in a legal sense, and regulation of BHC-affiliated banks differs from that of non-affiliated banks. Specifically, BHC affiliation brings the banking organization under the regulatory *aegis* of the Federal Reserve System, which is charged with implementing bank holding company law. In general, the range of activities permitted BHCs by federal law and regulation is not as broad as that permitted by some state bank charters. Nonetheless, the advantages appear to have outweighed the disadvantages historically, as the proportion of banks affiliated with BHCs has increased steadily.

Implications for Behavior

The advantages of BHC affiliation discussed above have fairly straightforward implications for the *consolidated* organization. Since shareholders have an interest in financing and operating the banking organization as a whole in a value-maximizing manner, the influence of affiliation should be reflected in enhancement of aggregate shareholder wealth and in the superior ability of BHC organizations to compete in providing financial services. There is some evidence for this. For example, the BHC organizational form does seem to be competitively superior because it has come to dominate American banking market structure.[3] Nonetheless, it would be desirable to observe the improved (or degraded) value of the affected entity directly. Most research on the effects of BHC affiliation has not focussed on the consolidated entities, however, but rather on the affiliate banks. There are two reasons.

First, studies of the consolidated enterprises are inherently difficult to conduct. To study directly the effects of affiliation on shareholder wealth, good estimates of the market value of equity must be available. Since most banking organizations are relatively small, closely held companies, estimates of the market value of equity (with or without affiliation) are not easily derived. It is possible to narrow one's focus to the larger banking organizations whose shares are actively traded, but the number of such institutions is small and the effective sample size in such studies compromises these efforts.[4]

The second reason that research has not focussed on the consolidated entity also is a pragmatic one. Policy interest in the bank holding company movement has been focussed on the implications of affiliation on *bank affiliate* behavior. This is natural, since lawmakers and regulators view the subsidiary bank as the entity delivering banking services; a different corporate structure and method of corporate control might well influence such an affiliate's behavior. However, the link between the motives for BHC formation and the likely behavior of bank affiliates generally does not follow in an obvious way from the motives for BHC formation.

Consider, for example, the potential influence of BHC affiliation on the profitability of the subsidiary bank. The underlying motivations for affiliation suggest only that the profitability of the *consolidated enterprise* would be higher with affiliation. The profits of the affiliated bank *may* be higher if the affiliation produces scale or scope economies for the bank affiliate. These may not appear at the bank level, however, if the way in which the bank funds or compensates other units in the holding company is through payment of fees (implicitly or explicitly) rather than through upstreaming of dividends to the parent holding company. Indeed, if inter-affiliate fees are high enough, it would be consistent with theory to find measured net bank income *lower* in affiliated banks (even though consolidated company earnings are improved by affiliation).

Similarly, the effects of affiliation on the capital position

of the bank also are ambiguous. If, for example, the desired use of debt is greater than is permitted at the bank level and the parent funds subsidiary bank equity with debt to relieve the regulatory constraint, then affiliation might result in increased capital at the bank level. If, on the other hand, regulatory capital constraints are not binding, a bank affiliated with a BHC might *reduce* its capital—redeploying it to fund the sister affiliates of the bank. This would be consistent with a view that the benefits of affiliation flow not from the double leverage opportunities afforded the BHC, but rather from the economies offered by the expanded scope of activities.

The influence of affiliation on the portfolio composition of banks also has been a concern of policy makers. One likely possibility is that bank portfolios become less diversified or otherwise riskier because diversification opportunities exist elsewhere in the holding company. On the other hand, the BHC form of organization avoids branching constraints and thus permits greater geographical diversification of lending activity and reduced portfolio risk.

These theoretical ambiguities, coupled with the focus of policy-makers on banks, rather than on consolidated banking organizations, makes the effect of BHC affiliation on bank behavior an empirical matter of some importance. It is a matter of increasing policy relevance, too, as lawmakers debate the appropriate organizational form in which to vest expanded powers.

II. Studying the Impact of Affiliation

The effect of BHC affiliation on bank behavior has received considerable attention from banking analysts. Over 50 studies published since the late 1960s have examined the effect of BHC affiliation on the performance of the subsidiary bank.[5]

Both simple means and frequency comparisons, as well as more sophisticated econometric techniques, are employed in this type of BHC research. Both types of studies employ techniques to control at least partially for the wide variation observed in bank characteristics and market conditions. In the simple statistical studies, the variation in bank characteristics is controlled by comparing the behavior of a bank after affiliation with its own behavior before affiliation. To control for changes in overall banking market conditions, the changes in the affected banks' performance are compared with the changes in performance observed in a "paired" sample of unaffiliated banks. The "pairing" involves identification of a non-affiliated bank of approximately the same size as the affiliated bank, located in the same (or a similar) banking market.[6]

In other studies, variation in bank characteristics and market conditions is controlled partially by entering attributes of the bank and the banking market as independent variables in regressions on bank performance measures. An estimate of the effect of BHC affiliation in a cross-section of affiliated and non-affiliated banks can then be observed with a dummy variable indexing the affiliation status of the banks in the sample.

Both types of studies have obtained similar estimates of the effect of BHC affiliation on bank behavior. Specifically, affiliation is found to (1) increase the proportion of loan assets in bank portfolios; (2) increase the proportion of state and local obligations; (3) increase loan fees and interest charges; (4) reduce holdings of cash and U.S. Treasury securities; and (5) with less regularity, increase deposit rates.[7] There has been variation in all of these findings across studies, as might be expected given the variation in models, samples, and statistical techniques. But the BHC affiliation studies have been striking in their tendency to find significant differences in the behavior of affiliated and non-affiliated banks.

In general, however, the findings have been particularly weak regarding the effects of affiliation on profitability and capital ratios—effects crucial to formulating regulatory implications. By using either paired comparisons or econometric models, little change is found in regulatory capital measures or profitability measures such as return on equity (ROE) or return on assets (ROA).[8]

Problems with BHC Studies

Bank holding company research has been subject to a variety of criticisms. One is that available statistical controls are insufficient to correct for the great variation in circumstances that contribute to differences in bank behavior observed in the real world. In theory, there should not be much, if any, variation in the performance of affiliated and non-affiliated banks in a competitive market. If sufficient statistical control for variation in market conditions peculiar to individual banks were possible, the observed variation in behavior would vanish. Research on BHC affiliation, therefore, like most bank research, implicitly relies on the existence of disequilibrium, adjustment lags, or imperfections in the extent of competition to introduce durable variations in observed performance and the decision to affiliate.

In addition to this general criticism, specific criticisms of BHC studies concern the particular methods of control. Univariate studies, for example, have been criticized for the bias they introduce in limiting the comparisons to banks of a size that permits "pairing" of observations. Most independent banks tend to be small; using pairing as a control technique thus tends to bias sampling toward smaller institutions.[9] If scale economies or other size-related considerations are determinants of bank behavior, as seems likely, such a sampling bias may be important. The univariate studies also have tended to use pre- and post-affiliation comparisons of bank behavior. This technique has been criticized for failing to control for the time that elapses between independence and affiliation.[10]

Econometric studies have received less fundamental criticism. Most criticisms have been directed at alleged errors of omission or commission in selection of control variables and in the stress placed on simple cross-sectional comparison, rather than the pre- and post-affiliation comparison technique used in the univariate studies.

Self-Selection Bias

A more important criticism of traditional bank holding company research—both in its univariate and econometric manifestations—is that it has ignored the potential problem of self-selection bias.[11] Self-selection bias arises because the decision to affiliate with a BHC is not random; rather, it is an outcome of the same organizational forces that determine other aspects of bank behavior.

To see how self-selection processes may bias the estimation of the influence of BHC affiliation, consider the typical cross-section regression employed in econometric studies:

$$Y_1 = a + bX_1 + cH + e_1 \qquad (1)$$

where Y_1 is a performance measure, such as bank ROE, or a portfolio measure, H is a dummy variable indicating the bank's affiliation status (H = 1, if affiliated, and = 0 otherwise), X_1 is a vector of other bank or market characteristics suspected of influencing performance, and a, b, and c are coefficients.

The influence of BHC affiliation is measured by the coefficient, c, on the affiliation variable. For the estimate of c to be unbiased, however, it must be uncorrelated with the error term, e_1, in the performance equation. This will be the case if holding company affiliation is assigned independently of the X variables, but not otherwise.

For example, suppose that a bank chooses to become affiliated with a BHC on the basis of another (unobserved) factor (such as expectations of future profits) not included in X_1, which we might call Y_2. Specifically, if

$$Y_2 > 0, \text{ then } H = 1 \qquad (2a)$$

and if

$$Y_2 < 0, \text{ then } H = 0. \qquad (2b)$$

That is, if expected profits exceed a certain level, then the bank chooses affiliation; if they are equal to or below that level, then it does not choose affiliation. The value that Y_2 takes depends upon other conditions that prevail in the market or at the bank, X_2, and a random disturbance term, e_2. That is,

$$Y_2 = d + fX_2 + e_2. \qquad (3)$$

The relationships (1), (2), and (3) make up a simple, simultaneous equations system. Thus, if the covariance of e_1 and e_2 is not zero, ordinary regression analysis of equation (1) will not produce unbiased estimates of its coefficients. This is because any disturbance to e_2 will translate into a disturbance in H, which would then be correlated with the covarying e_1. Thus, H is a stochastic variable correlated with e_1.

As a practical matter, self-selection bias seems likely. That is, it seems likely that factors that disturb the bank's perception of its expected profits, for example, are likely also to disturb its performance. Therefore, it is likely that the disturbance terms of equations (1) and (3) do have non-zero covariance, and that simple regression analyses of BHC impact will produce biased estimates of the effects of BHC affiliation.

Treating Self-Selection Bias

The statistical solutions to the problems of self-selection bias belong to a class of econometric methods known as simultaneous equations techniques.[12] The general approach of these techniques is to "purge" the stochastic explanatory variable (H, in this case) of the influence of e_2. This is achieved by *estimating* H using only non-stochastic variables in a separate, "first-stage" regression. The predicted values of H are then mathematical combinations of non-stochastic variables and would be uncorrelated with e_1. If these predicted values are used instead of the actual values of H in regression (1) (the second stage), then the estimates of c would be unbiased.

Two problems arise in applying this technique to the model described by equations (1), (2), and (3). First, if all

of the non-stochastic variables in the first stage also logically belong in the second stage regression, then the predicted values of H are simply a linear combination of the X_1, and the second stage regression will not be estimable. (Of course, it need not be the case that all of the X variables in equation (1) belong in equation (3) and vice versa. In such a case, the exclusion of certain X variables will permit identification of the influence of BHC affiliation on the performance measure, Y_1.)

Second, H (the stochastic variable that introduces the simultaneous equations bias) is a dichotomous variable; it takes on values only of 0 or 1. Estimation of a linear regression equation with a dichotomous dependent variable (such as the first stage regression above) poses difficulties.

Both the identification problem and the dichotomous dependent variable can be addressed by estimating the first stage using a model known as a probit model. Specifically, a *probit* relationship can be used to estimate the first stage, producing predicted values of H. The probit model is nonlinear, and permits identification of the coefficient on H *even if exclusion is not possible*. The probit model also is intended specifically for use with dichotomous dependent variables.

Equation (1) would then be estimated in the form

$$Y_1 = a + bX_1 + c\hat{H} + e_1 \qquad (4)$$

where $\hat{H}$ is the *predicted probability* of being affiliated with a BHC derived from probit estimation of the affiliation decision characterized by (2) and (3). It can be shown that the coefficients of (4) will be unbiased, though the standard error estimates will not be precisely correct unless a joint maximum-likelihood estimation technique is used. As a practical matter, the standard error estimates tend to change little with maximum likelihood estimation.[13]

III. Application to a BHC Performance Study

In this section, the probit technique for controlling sample selection bias is applied to an econometric study of the performance of affiliated and non-affiliated banks. Results from conventional econometric techniques for identifying the effects of affiliation are compared to those obtained from a two-stage estimation procedure using a probit model to explain the BHC affiliation selection process.

The Sample

The study examines the performance of a cross-section of commercial banks in the Twelfth Federal Reserve District in 1985.[14] Because the circumstances of the banks in the sample in previous years were expected to be relevant to both the affiliation status of the banks and their performance, data were collected for these banks for the years

Table 1
Descriptive Statistics for a Sample of Banks in the Twelfth Federal Reserve District in 1985
(All Statistics are in percent, unless otherwise noted.)

Variable	Affiliated	Non-affiliated	Total
Total Assets ($000)	328,442	103,139	202,673
Total Deposits ($000)	279,250	88,557	172,802
Composition of Assets			
Cash	8.92	8.58	8.73
Treasury Securities	15.91	22.41	19.57
Loans and Other Assets	75.17	69.01	71.70
Loan Income/Loan Assets	4.11	3.78	3.92
Deposit Interest/Deposits	2.98	3.02	3.00
Return on Equity	1.56	4.25	2.89
Return on Assets	0.32	0.51	0.42
Capital/Risk Assets	12.97	12.51	12.71
Capital/Total Assets	8.29	8.76	8.56
Age of Banking Org. (years)	41.30	38.90	39.90
Sample Size (number)	163	161	324

1976, 1979, 1982, and 1985. The only selection criterion used in constructing the sample was that the banks have reported data on Reports of Condition and Income Statements continuously during the sample period. There were 324 such banks. Not all banks, however, report all variables. Thus, some of the analyses presented below using certain performance measures or other variables result in correspondingly smaller samples.

Bank holding company affiliation grew steadily in the sample period. In 1976, only 20 percent of the banks in the sample were affiliated with BHCs. By 1985, the proportion had risen to over 49 percent. Of the 324 banks in the sample, 110 changed their affiliation status at some point during the sample period. The growth in affiliation was primarily a small-bank phenomenon, however. The proportion of bank assets in affiliated institutions was already 87 percent in 1976, and grew to 95 percent in 1985.

The sample means are presented in Table 1. From a simple comparison of sample means, affiliated banks tend to be larger, with more loans and higher loan rates, fewer Treasury securities, and lower returns than non-affiliated banks. The degree of leverage appears to be approximately the same in both types of organizations.

Simple Econometric Tests

Using the simple econometric model described by equation (1), the effect of affiliation with a BHC can be estimated using ordinary least squares regression techniques. The effects of affiliation in the current period are estimated in this manner for current measures of leverage, profitability, portfolio composition, and pricing.

In addition to the dummy variable representing current BHC affiliation, several variables to control for cross-sectional variation in bank and market attributes were included in the regressions. Lagged bank size is used to control for the potential influence of size on the behavior of affiliated banks. The age of the institution is included, on the presumption that mature financial institutions may behave differently than start-up organizations. The length of time the bank has been affiliated with the BHC also is included—interacted with affiliation status—to capture potential vintage effects of affiliation on bank performance. State dummy variables are included to control for effects of variation in market conditions, variations in bank branching, state charter powers, or other regulations that might be expected to vary by state.[15]

The estimated impacts of affiliation using this simple regression technique are summarized in the first column of Table 2. Not all of the coefficients on the affiliation variable reported in this column are statistically significant. Their signs suggest, however, that affiliation with a

Table 2
Estimated Effects of BHC Affiliation on Various Performance Measures in 1985

(T-statistics are in parentheses.)

	Correction for Self-Selection?			
	No	Yes	No	Yes
Performance Measure	**Exclusion**		**Non-Exclusion**	
Return on Equity	−0.078	−0.147	−0.086	−0.573
	(−1.497)	(−0.961)	(−1.641)	(−0.964)
Return on Assets	−0.003*	−0.003	−0.004*	−0.003
	(−2.433)	(−1.356)	(−3.527)	(−1.253)
Capital/Risk Assets	0.049	0.584*	−0.025	0.166*
	(0.809)	(3.329)	(−0.581)	(2.340)
Capital/Total Assets	−0.010	0.150*	−0.030*	0.068*
	(−0.693)	(3.607)	(−2.633)	(3.521)
Deposit Rate	−0.000	0.000	0.000	−0.004*
	(−0.750)	(0.233)	(0.254)	(−2.275)
Loan Rate	0.002	0.012*	0.001	0.007*
	(1.278)	(4.017)	(1.633)	(2.635)
Muni Bonds/Assets	−0.034*	−0.089*	−0.034*	−0.105*
	(−2.146)	(−3.213)	(−2.043)	(−3.363)
Cash/Assets	0.000	0.019	0.002	0.052*
	(0.094)	(1.237)	(0.329)	(2.670)
Treas. Sec. /Assets	−0.050*	−0.211*	−0.052*	−0.134*
	(−2.362)	(−5.121)	(−2.642)	(−2.707)
Loans/Assets	0.125*	0.246*	0.124*	0.201*
	(3.241)	(3.654)	(3.302)	(2.765)

*Coefficient is different from zero at the 95 percent confidence level or better.

Note: The regressors in the exclusion model are: 1985 affiliation status; age of the banking organization; total assets in 1982; state dummies; the affiliation status interacted with BHC age; a constant term. The regressors in the non-exclusion model include these plus the capital/asset ratio in 1976; return on equity in 1976; the average loan rate in 1976; the average deposit rate in 1976; and total assets in 1976. In the regressions employing a self-selection correction, the actual affiliation status is replaced with an affiliation status predicted from the probit relationship presented in Table 3.

BHC appears to (1) increase the proportion of loans in total bank assets, (2) increase municipal bond holdings by the affiliated bank, (3) increase average loan income and deposit rates, (4) reduce holdings of cash and Treasury securities, and (5) reduce return on equity or assets. The effects on leverage are mixed; the use of equity is lower relative to total assets, but higher relative to risk-assets. These findings generally are consistent with the findings of other studies that have used other samples at other points in time.

Correcting for Self-Selection

As discussed above, the process for statistical correction of self-selection bias involves a two-stage estimation procedure. The first stage involves estimation of the "Affiliation Choice" relationship. The current affiliation status of the banks in the sample is modelled using probit representations. The selection of variables for inclusion in the probit regression is constrained somewhat by the availability of historical data on the study sample. The variables selected are intended to capture the influence of prior performance, prior affiliation status, and state location on the affiliation choice. The estimated parameters of the probit model of the affiliation choice relationship are presented in Table 3.

It appears from this regression that, in addition to prior affiliation status, prior performance of the banking organization bears importantly on whether it was affiliated in 1985. The probability of being affiliated with a BHC in 1985 appears to be positively related to the capital/asset ratio and the loan rate, and negatively related to the return on equity, the deposit rate, and total assets. The latter effect is consistent with the availability of more favorable double-leverage opportunities to smaller (less than $150 million in assets) banks. The state dummies are consistently insignificant, an observation in keeping with the notion that variations in state branching or charter powers are not important in determining BHC affiliation status—at least in the states that comprise the Twelfth District.

The probit estimates of affiliation choice are then included in two alternative representations of the performance relationships. The first, called the "Exclusion Model," excludes from the performance relationship some of the explanatory variables that were included in the affiliation choice relationship. The excluded variables are various bank performance measures from the year 1976.[16] This may help to identify the effects of the affiliation decision by excluding these variables from the performance relationships. In the second representation, called the "Non-Exclusion Model," these variables for 1976 are included in the performance regressions as well. Identification of the influence of BHC affiliation on performance is achieved exclusively by virtue of the nonlinearity of the probit relationship.

Identification by exclusion may or may not be justified. One must be willing to assume that some variables that influenced holding company affiliation can be excluded as influences on current bank performance. There is no *a priori* way of telling, however, whether that assumption is more reasonable than the alternative approach, which relies exclusively on the nonlinearity of the affiliation choice relationship.[17]

In the first column of Table 2, the estimated effects of affiliation are reported for an exclusion model with no correction for self-selection bias. In the second and fourth columns, the impact of affiliation is presented for the two-stage model that corrects for self-selection bias. The second column presents the results from the exclusion model and the fourth column presents the non-exclusion results. For comparison, the third column reports the results of a simple regression which does not exclude any performance variables and does not correct for self-selection bias.

Table 3
Estimated Coefficients of Probit Model of Affiliation Choice

Dependent Variable: Affiliation Status, 1985

Variable	Coefficient	T-Statistic
Constant	−1.632	−2.712
Age of Bank	0.005E-06	0.290
BHC Affilliation, 1982	2.880	6.990
Capital/Assets, 1976	2.663	2.116
ROE, 1976	−1.345	0.703
Loan Rate, 1976	41.801	3.260
Deposit Rate, 1976	−10.935	−2.577
Total Assets, 1976	−2.344E-06	−1.946
Alaska Dummy	−0.191	−0.453
Arizona Dummy	0.024	0.054
Hawaii Dummy	−0.179	−0.713
Idaho Dummy	0.045	0.132
Nevada Dummy	−0.302	−0.556
Oregon Dummy	0.067	0.252
Utah Dummy	0.171	0.670
Washington Dummy	0.012	0.060
Log Likelihood	−191.58258	
Cases with BHC in 1985 = 1	160	
Cases with BHC in 1985 = 0	164	

Findings

The results of these simple tests of the effects of bank holding company affiliation differ qualitatively between the models that treat self-selection bias and those that do not. The low levels of significance of some of the estimated coefficients permit few strong statistical statements. However, qualitatively at least, treatment of self-selection bias appears to reverse the estimated direction of the effect of BHC affiliation or change the point estimate of its magnitude in virtually all cases.

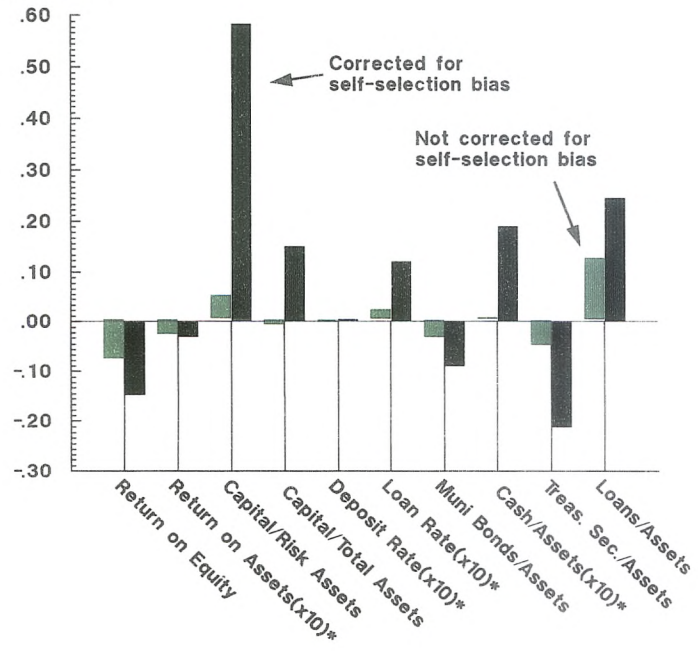
The effects are most easily seen in the two panels of Chart 1. The conventional finding that equity is lower in an affiliated bank is reversed in both of the models that treat self-selection bias. After correcting for self-selection, equity relative to risk assets and equity to total assets both appear to be higher at affiliated banks, although the finding for the equity/risk asset ratio is statistically significant only for the self-selection model that employs the exclusion assumption to identify BHC impact.

The models with self-selection corrections find that measured capital ratios are *higher* at BHC-affiliated banks. This is consistent with the view that affiliation is attractive because it allows BHCs to downstream debt as equity to the subsidiary bank. Indeed, the failure of earlier studies to find this impact consistently has been puzzling.

As the table and charts indicate, the self-selection correction models also change the findings regarding the impact of BHC affiliation on portfolio composition. The measured impact of affiliation on the share of loans in total assets is two times larger after correction for self-selection than before. Failing to correct for self-selection bias may underestimate the impact because banks choosing to affiliate with BHCs may tend to be those that, for other reasons, may wish to take on more risky assets and see the BHC vehicle as a convenient means of financing such a portfolio.

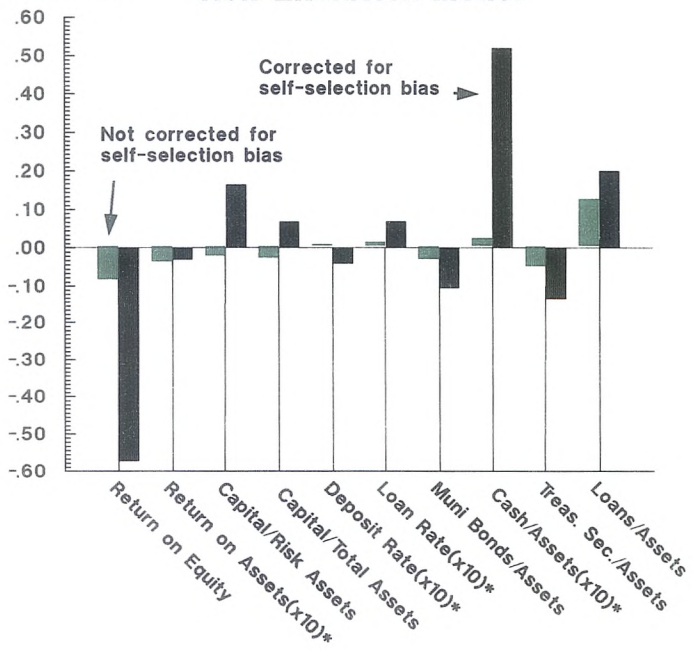
A larger impact on the average loan rate (measured as the ratio of average loan income and fees to total loans) also is found with the models using a self-selection correction. The impact of BHC affiliation on the measured loan rate is six to seven times higher in the corrected versus the uncorrected models. This finding is consistent with the argument that those seeking BHC affiliation may be seeking more risk if the higher rates reflect a risk-compensated return.

Chart 1A
Effects of BHC Affiliation on Bank Performance
Exclusion Model



*These effects have been multiplied by a factor of ten to make them more visible on the chart.

Chart 1B
Effects of BHC Affiliation on Bank Performance
Non-Exclusion Model



*These effects have been multiplied by a factor of ten to make them more visible on the chart.

Other impacts of BHC affiliation are less notably altered by employing the two-stage model for self-selection bias. The holding of Treasury securities appears to be reduced by BHC affiliation in the corrected models. The influence of affiliation on the holding of cash assets remains of low statistical significance. The measured negative impact on the return on equity is larger, but of low statistical significance. The measured impact on the return on assets is about the same, but of lower statistical significance. Difficulty in measuring impacts on earnings and returns is typical in banking research, due to the problems in using accounting measures of the components of these statistics.

Traditional models have tended to find a significant, positive effect of BHC affiliation on a bank's willingness to hold municipal bonds. This is not the case in my sample, perhaps because of the relatively recent data used. The tax treatment of municipal bonds held by banks changed with tax legislation in the 1980s and may have changed the direction of the effect of BHC affiliation. Both corrected models appear to amplify this effect.

Most of the other coefficients of the regression are not of policy interest and, for brevity, are not discussed here. However, it is interesting to note that the variable designed to capture the effect of the time elapsed since BHC formation—the age of the BHC interacted with affiliation status—is insignificant in all performance regressions. Hence, all of the effects of BHC affiliation appear to be captured by the affiliation status variable alone. (This is the measure reported in Table 2.) This suggests that, whatever the influence of BHC affiliation, the effects do not grow or fade with time. It also is interesting to note that the dummy variables for the various states in the region generally are not significant in the affiliation choice probit regression. There is considerable variation in the powers afforded banking organizations in the various states of the region. If state chartering was a viable alternative to obtaining some of the flexibility of BHC affiliation, presumably the affiliation choice regression would have been influenced accordingly by the state dummies.

IV. Conclusion and Policy Implications

The measured effects of BHC affiliation on subsidiary banks are sensitive to attempts to correct for self-selection bias. This suggests that the behavior of a bank and its decision to affiliate with a BHC are statistically related. This, in turn, implies that the findings of the large number of earlier BHC impact studies should be reconsidered in light of their failure to recognize and address this statistical problem directly.

In the specific population of banks examined here, several important measured effects of BHC affiliation are changed when self-selection correction procedures are employed. As important as the direction and magnitude of the changes, however, is the fact that BHC affiliation continues to be associated with significant differences in bank behavior *even when* self-selection bias is treated. This suggests that the behavior of a bank is not independent of the nonbank and holding company affiliations it forms, and contradicts the notion that banks can be "corporately" separated from the activities of their sister or parent organizations. Such separation often forms the basis of proposals that would give banking organizations additional nonbanking powers. My findings suggest that corporate separation cannot fully insulate the bank from the expanded risk-taking opportunities that such an expansion might imply.

ENDNOTES

1. In some states, a banking organization is allowed to engage in a wide variety of activities under a state charter. Thus, obtaining a state charter is one way to obtain broad banking powers. The fact that the BHC movement has dominated state chartering may suggest that other aspects of the BHC form of organization may be more important than the powers issue.

2. Specifically, if a BHC has substantial nonbank subsidiaries, its consolidated capital/asset ratio may appear high (and compatible with the subsidiary bank standard), but be lower than it would be if the bank truly had to be financed with equity. In addition, Regulation Y permits banks smaller than $150 million in assets to form a BHC and use as much as three times the debt in the parent as would be permitted in the bank affiliate.

3. In the population employed in this study, for example, fully 92 percent of bank assets are represented by BHC affiliated banks.

4. A good example of a market valuation approach that suffers from sample size problems is Varvel (1975). Frieder and Apilado use share price evidence in the 1982 study, and a synthetic valuation scheme in their 1983 study.

5. The paper by Frieder and Apilado (1982) provides a useful summary and synthesis of bank holding company research.

6. Frequently cited "matched pair" studies include Smith (1971), Talley (1972), and Hobson, Masten and Severiens (1978). The econometric studies cited are those by Johnson and Meinster (1975), Rose (1975), Mingo (1976), Mayne (1977), and Rhoades and Rutz (1982).

7. See Frieder and Apilado (1982).

8. See the study by Fraas (1974) summarizing the ambiguous findings of earlier studies.

9. This criticism is mentioned by Jessup (1974) and Frieder and Apilado (1982).

10. The Hobson, Masten and Severiens (1978) study was one of the first to emphasize the effects of the time elapsed since acquisition.

11. The author is not aware of any direct reference to the problems of self-selection bias in previous bank holding company research.

12. The literature on self-selection bias in economics arose out of studies of government program impact. See, for example, Barnow (1975) and Barnow and Cain (1977). The statistical properties of estimators of program impact in an environment of self-selection bias were studied by a number of authors, including Heckman (1976 and 1979) and Olsen (1979).

13. See Hausman and Wise (1977).

14. The cross-sectional design has been employed in most earlier studies of the effects of BHC affiliation. Other designs, such as a pooled time-series cross section, pose a number of difficulties for the analyst. Banking regulation and law changed significantly in the early 1980s, first with deposit deregulation in 1980, and then with changes in capital regulation in 1982. Also, the format of the Reports of Condition and Income changed several times during this period, making comparisons of certain financial measures suspect over time.

15. The use of variables lagged prior to change to BHC status also was examined. This modification turns out not to have significant effects on the regression analyses. More importantly, however, since any bank could conceivably change its status at any time—albeit with some implementation lag—a fixed lag in the explanatory variables across all observations is more appropriate.

16. For the regressions reported in the paper, the excluded independent variable set includes leverage, ROE, loan rate, deposit rate, affiliation status, and total asset size measures from the year 1976.

17. In the results presented here, the probit formulation of the choice regression is used to correct for self-selection bias in both the exclusion and non-exclusion models. This is not strictly necessary to achieve identification with an exclusion assumption. The choice regression used to produce predictions of affiliation status can be linear and identification still achieved. A linear formulation of the choice regression, however, has a number of undesirable properties, including the propensity to predict choice probabilities outside the range of zero to one.

REFERENCES

Barnow, B.S. "The Effect of Head Start and Socioeconomic Status on Cognitive Development of Disadvantaged Children," Ph.D. Dissertation, University of Wisconsin, 1975.

Barnow, B.S. and G.G. Cain. "A Reanalysis of the Effect of Head Start on Cognitive Development Methodology and Empirical Findings," *Journal of Human Resources*, 1977.

Bedingfield, J.P., P.M. Reckers, and A.J. Stagliano. "Distributions of Financial Ratios in the Commercial Banking Industry," *Journal of Financial Research*, Volume 8, Spring 1985.

Boyd, J.H., G.A. Hanweck and P. Pithyachariyakul. "Bank Holding Company Diversification," *Proceedings of a Conference on Bank Structure and Competition*, Federal Reserve Bank of Chicago, 1984.

Brewer, Virgil and William Dukes. "Empirical Evidence on the Risk Return Relationships Between Banks and Related Bank Holding Companies," *Review of Business and Economic Research*, Volume II, Spring 1986.

Curry, Timothy J. and John T. Rose. "Bank Holding Company Presence and Banking Market Performance," *Journal of Bank Research*, Winter 1984.

Fraas, Arthur G. "The Performance of Individual Bank Holding Companies," *Staff Economic Study*, #84, Board of Governors of the Federal Reserve System, 1984.

Frieder, Larry A. and Vincent P. Apilado. "Bank Holding Company Expansion: A Refocus on its Financial Rationale," *The Journal of Financial Research*, Volume VI, Spring 1983.

Heckman, James J. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 1976.

_____. "Sample Selection Bias as a Specification Error," *Econometrica*, Volume 47, January 1979.

Hobson, Hugh A., John T. Masten, and Jacobus T. Severiens. "Holding Company Acquisitions and Bank Performance: A Comparative Study," *Journal of Bank Research*, Summer 1978.

Jessup, Paul and Roger Upson. "Return from Bank Holding Companies," *The Bankers Magazine*, Volume 155, Spring 1972.

Johnson, Rodney D. and David R. Meinster. "An Analysis of Bank Holding Company Acquisition: Some Methodological Issues," *Journal of Bank Research*, Volume 4, Spring 1973.

Lawrence, Robert J. "The Performance of Bank Holding Companies," *Staff Economic Study*, #55, Board of Governors of the Federal Reserve System, June 1967.

Lawrence, Robert J. and Samuel H. Talley. "An Assessment of Bank Holding Companies," *Federal Reserve Bulletin*, Board of Governors of the Federal Reserve System, January 1976.

Lee, Warren F. and Alan K. Reichert. "Effects of Multibank Holding Company Acquisitions on Rural Community Banks," *Proceedings of a Conference on Bank Structure and Competition*, Federal Reserve Bank of Chicago, May 1975.

Light, Jack S. "Effects of Holding Company Affiliation on De Novo Banks," *Proceedings of a Conference on Bank Structure and Competition*, Federal Reserve Bank of Chicago, 1976.

Maddala, G.S. and Lung-Fei Lee. "Recursive Models with Qualitative Endogenous Variables," *Annals of Economic and Social Measurement*, 1976.

Mayne, Lucille S. "A Comparative Study of Bank Holding Company Affiliates and Independent Banks, 1969–1972," *Journal of Finance*, Volume 32, March 1977.

Meinster, D.R. and R.D. Johnson. "Bank Holding Company Diversification and the Risk of Capital Impairment," *The Bell Journal of Economics*, Volume 10, Autumn 1979.

Mingo, John J. "Capital Management and Profitability of Prospective Holding Company Banks," *Journal of Financial and Quantitative Analysis*, Volume 10, June 1975.

_____. "Capital Management by Holding Company Banks," *Journal of Business*, Volume 48, October 1975.

_____. "Managerial Motives, Market Structures and the Performance of Holding Company Banks," *Economic Inquiry*, Volume 14, September 1976.

_____. "More on the Performance Characteristics of Holding Company Acquisitions," *Proceedings of a Conference on Bank Structure and Competition*, Federal Reserve Bank of Chicago, June 1973.

Olsen, R.J. "A Least Squares Correction for Selectivity Bias," *Econometrica*, Forthcoming.

_____."Tests for the Presence of Selectivity Bias and their Relation to Specification of Functional Form and Error Distribution," Working Paper 812, Yale University, Institution for Social and Policy Studies, 1979.

Piper, Thomas R. "The Economics of Bank Acquisitions by Registered Bank Holding Companies," Research Report No. 48, Federal Reserve Bank of Boston, March 1971.

Piper, Thomas R. and Steven J. Weiss. "The Profitability of Bank Acquisitions by Multibank Holding Companies," *New England Economic Review*, Federal Reserve Bank of Boston, September-October 1971.

Rhoades, Stephen A. and Roger D. Rutz. "The Impact of Bank Holding Companies on Local Market Rivalry and Performance," *Journal of Economics and Business*, 1982.

Rose, John T. and Donald T. Savage. "Bank Holding Company *De Novo* Entry, Bank Performance, and Holding Company Size," *Quarterly Review of Economics and Business*, Volume 23, Winter 1983.

Rose, Peter S. and Donald R. Fraser. "The Impact of Holding Company Acquisitions on Individual Banks," *The Bankers Magazine*, Volume 156, Spring 1973.

Smith, David, "The Performance of Merging Banks," *Journal of Business*, Volume 44, April 1971.

Talley, Samuel H. "The Effect of Holding Company Acquisitions on Bank Performance," *Staff Economic Study*, #69, Board of Governors of the Federal Reserve System, 1972.

Varvel, Walter A. "A Valuation Approach to Bank Holding Company Acquisitions," *Economic Review*, Federal Reserve Bank of Richmond, July-August 1975.

Wall, Larry D. "Has Bank Holding Companies' Diversification Affected Their Risk of Failure?" *Journal of Economics and Business*, Volume 39, 1987.

Hotelling's Rule Repealed? An Examination of Exhaustible Resource Pricing

Ronald H. Schmidt

Contrary to the predictions of economic theory, prices of important exhaustible resources have not appreciated in real terms during the past century. Possible explanations for the lack of a trend in prices, such as changes in demand, discoveries of new reserves, and technological change are explored in this article. Based on the evidence, it appears that theoretical models consistently have underestimated the price elasticity of supply of and demand for exhaustible resources. Despite increasing consumption, resource availability has increased as well, suggesting that pressure for rising real resource prices will continue to be suppressed.

Economist, Federal Reserve Bank of San Francisco. The author would like to thank Steve Dean for his excellent assistance. Editorial committee members were Michael Keeley, Barbara Bennett, and Randall Pozdena.

An important contribution of economics to public policy is in the area of intertemporal resource allocation—particularly in the case of the optimal depletion of exhaustible resources. Models ranging from very simple to highly-sophisticated treatments of the topic have provided important insights into how market forces often can address problems of growing scarcity without intervention by centralized authorities.

One illustration of this role was provided in the early 1970s with the publication of *Limits to Growth* (LTG).[1] Using computer simulations of trends in consumption, output, resources, and population growth, LTG projected growing shortages of key raw materials and a declining standard of living in the world economy. Those projections, however, ignored the endogeneity of prices. Rebuttals to LTG based on dynamic optimization models were able to evaluate the likelihood of the LTG outcomes and suggest a far more adaptive environment. Prices would rise as commodities become scarce, they argued, causing automatic shifts in consumption patterns.

The predictions generated by the dynamic optimization models have become increasingly important in economic policy formation. Growing familiarity with dynamic optimization techniques led to the widespread adoption in economic and forecasting models of many of the "arbitrage equations" that are generated in the intertemporal optimization literature. For example, following the oil price spikes in 1973-74 and 1979-80, predictions of future oil prices routinely have been based on the intuitively appealing arbitrage relationship often referred to as "Hotelling's rule," which states that prices of exhaustible resources should rise at the same rate as other financial assets. That is, the rate of price increase should equal the interest rate. Otherwise, commodity holders would not be indifferent between current and future sales, and hence, would withhold or accelerate current sales until the present value of the future price equaled the current price.

This assumption that oil prices would be determined by Hotelling's rule continues to be embedded in most dynamic economic forecasting models. Especially in long-term forecasting models, oil prices are assumed to rise faster than the general level of inflation because the real interest rate is positive.

This assumption, however, fails to correspond to experience. As discussed in this article, oil prices, as well as most other major mineral prices, have not followed the predicted path. Since the 1870s, these real resource prices have not had a noticeable trend, in contrast to the rising trend predicted by Hotelling's rule. Moreover, prices have been highly volatile, rather than stable as arbitrage relationships would suggest.

In both the LTG and Hotelling scenarios, the explicit exhaustibility of the resource is a central assumption. New discoveries and innovations can alter the period over which extraction and consumption occur, but the resource eventually is fully consumed. Consequently, both theories predict declining per capita wealth unless the economy can substitute other factors of production for the resource.

Historical data contradict this view, however. As discussed in this article, the issue of scarcity and exhaustibility of natural resources is questioned by the evidence: prices have not appreciated in real terms, consumption has risen, and known reserves have risen sharply for nearly all resources examined here.

The thrust of this article is to suggest that neither the LTG nor the Hotelling scenarios, as commonly expressed, are likely. Because of a consistent tendency to underestimate the response of technological progress and innovations to perceptions of scarcity, the models using Hotelling arbitrage equations will tend to underpredict resource availability and overpredict price increases. Moreover, the LTG predictions are unlikely because technological progress appears to occur at a sufficiently rapid pace to prevent growing scarcity.

In Section I, the simple Hotelling model and the resulting arbitrage equations are derived. Empirical evidence testing the arbitrage condition for copper, lead, iron, zinc, and petroleum is then presented in Section II. The evidence generally provides poor support for a Hotelling price path, indicating the absence of a trend and the presence of large unexplained errors. Several explanations for this failure are presented in Section III. Some of the most important causes—uncertainty about reserves and the rate of technological change, the properties of the extraction cost function, shifts in tastes, changes in market structure, and problems caused by imperfect information—are discussed.

Concluding remarks are presented in Section IV. Based on the information problems, uncertainty, and the empirical evidence presented in this article, Hotelling's rule appears to be a poor guide for projecting prices of exhaustible resources, and the LTG model provides a poor prediction of resource scarcity. Rather, the lack of a trend in real resource prices suggests that economic forces are working to encourage expanding resource availability.

I. Hotelling's Rule

Exhaustible resources have received special attention in the economics literature. A resource is said to be exhaustible if its current use in some way reduces a finite stock of future uses:

> If we ignore the act of extraction as a production activity, such a resource is among the class of non-produced goods (i.e., it is a primary commodity). But then, so is agricultural land, and we do not usually regard land as being exhaustible in the same way as fossil fuels are. The distinguishing feature of an exhaustible resource is that it is used up as an input in production and at the same time its undisturbed rate of growth is nil. In short, the intertemporal sum of the *services* provided by a given stock of an exhaustible resource is finite. Land, if carefully tilled, can in principle provide an unbounded sum of services over time. This is the difference.[2]

Exhaustible resources, therefore, can command a scarcity premium that grows over time, and unlike land, this growth does not depend on the growth of demand for the service, but rather, on the diminishing availability of the stock of services because of previous consumption. Any consumption of the resource should increase the scarcity of the resource and, hence, affect the value of future scarcity rents. Optimal depletion of an exhaustible resource, therefore, is a problem of intertemporal allocation.

The most influential approach to modeling intertemporal depletion of exhaustible resources is generally attributed to Hotelling (1931). Hotelling derived the path of optimal prices and consumption in a model that assumes that the objective of society is to maximize the present discounted value of consumption of a resource that has a fixed stock. In its simplest form, the Hotelling problem can be stated as follows:

$$\underset{c}{\text{maximize}} \quad \int_0^T e^{-\delta t} U(c(t))\,dt \tag{1}$$

subject to:

$$\dot{R}(t) = -c(t) \tag{2}$$

$$R(0) \geq \int_0^T c(t)dt \tag{3}$$

$$R(t),\ c(t) \geq 0.^3 \tag{4}$$

where R is the level of remaining reserves, c is the consumption of the resource at time t, U(c) is the utility associated with consumption of the resource at time t (which is assumed equal to production, for simplicity), δ is the discount rate, and T is the (finite) date at which the resource is depleted. The problem is one of choosing an optimal consumption path, c(t), to yield the highest value to the agent subject to the constraints that production must always be positive and cumulative production cannot exceed the resource stock.

The fixed supply of the resource is the critical difference between exhaustible resources and other commodities produced at constant cost. Because the initial stock of resources is in fixed supply, a scarcity premium can be captured by the resource owner. Hence, as long as the scarcity is sufficiently apparent, prices can exceed production costs throughout the period of its consumption.

The mathematical solution of (1) - (4) involves straightforward application of the calculus of variations, and is available in a variety of sources [Hotelling (1931), Dasgupta and Heal (1974), Stiglitz (1974), and Schmidt (1984)]. It can be demonstrated that the arbitrage equation determining intertemporal allocations is:

$$\dot{U}'(t)/U'(t) = \delta. \qquad (5)$$

The solution affirms that resource use is optimal when the marginal utility of consumption rises at the agent's discount rate. When this occurs, the present value of the marginal utility of the last unit is the same in each time period, and because the marginal utility of consumption is assumed to be inversely related to consumption, no opportunity for arbitrage would remain.

In a competitive system, the marginal utility of consumption is proportional with the observed price for the commodity. Substituting the resource price for U'(t), the marginal utility of consumption, yields what has come to be known as "Hotelling's rule":

$$\dot{P}(t)/P(t) = \delta \qquad (6)$$

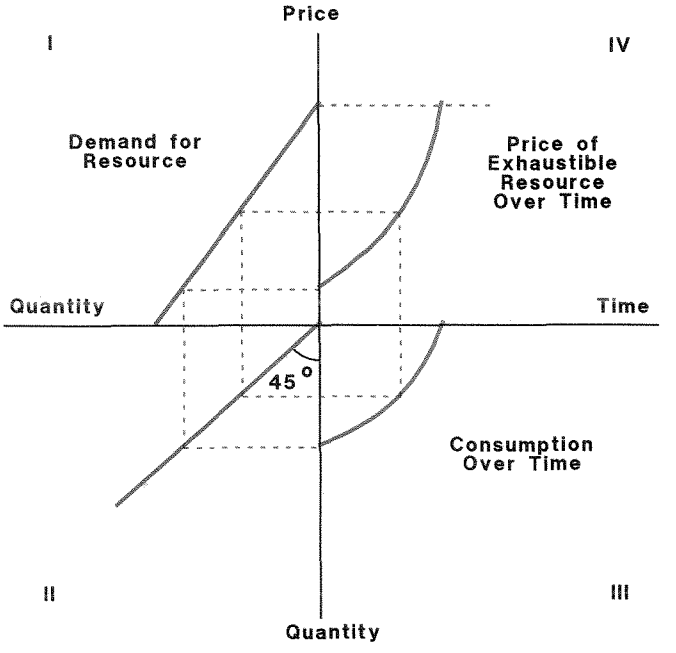
Equation (6) predicts that real prices will rise at the rate of time preference, which is often proxied by the observed rate of interest.[4]

The logic behind this rule is difficult to contest. Producers with perfect foresight and no holding or production costs should be indifferent between current and future production as long as the resource appreciates at the same rate that the proceeds from current production would earn if invested in other assets. If prices grow at a faster rate, arbitrage opportunities exist that would encourage reduced current production, because the returns to sales in the future would have a higher present value. This response would reduce the rate of price appreciation.

The solution to the problem can be seen graphically in Figure 1, a four-quadrant depiction of the optimal depletion problem [Herfindahl (1967)]. As demonstrated in Figure 1, an optimal price path can be determined by using the arbitrage equation to define the rate of change between periods, in conjunction with the resource constraint, which makes it possible to determine the starting price level.

The first quadrant depicts the demand curve for the resource at a point in time. For simplicity, it is assumed that demand is stationary; that is, the demand curve does not shift over time. The second quadrant simply maps the consumption at a particular point in time from the demand curve in quadrant I to its cumulative consumption in quadrant III. Quadrant III keeps track of resource use over time. The area bounded by the consumption path and the axes determines whether the chosen price and consumption paths violate the resource constraint. This area equals total consumption over time, and cannot exceed the available reserves of the resource. The fourth quadrant maps the price path described by Hotelling's rule.

Figure 1
The Optimal Depletion of an Exhaustible Resource



To determine the optimal path, an initial starting price is chosen in quadrant IV. Then, given the resulting price path in IV and the demand curve in quadrant I, it is possible to trace out the implied consumption path in quadrant III. The cumulative consumption resulting from the price path can be compared to the resource stock available. If the implied consumption exceeds the available stock, the starting price is raised in quadrant I and the exercise is repeated. When the starting price and resulting price path exactly exhaust the available resource at the time when the price reaches a level at which demand is choked off, the path is optimal.

The particular resource model developed above uses highly restrictive assumptions. In particular, it assumes constant demand, no extraction costs, known reserves, and no technological change. As discussed later in this article, more complicated depletion models have relaxed some of these assumptions. These enhancements modify the optimal price path and make the relationship expressed in equation (6) more complex, but the results continue to predict a positive relationship between price appreciation and interest rates. In other words, the model described in (1)-(4) is an abstraction, but the central prediction—that real prices should rise over time—is independent of many of these assumptions.

II. Empirical Evidence

In contrast to the theoretical predictions, however, Charts 1a-1e show that the real prices of copper, lead, iron, zinc, and petroleum have been highly volatile, but have not exhibited a significant trend over the period from 1870 to 1986. Current real prices for many of these minerals are at the levels of 100 years ago. None of the minerals has exhibited the real appreciation that would be predicted by a simple model.

Interestingly, the only mineral that visually demonstrates a rising real price is iron, which has little scarcity rent attributed to the resource. Also, the commodities that demonstrated some significant trend in the early 1980s have seen a sharp reversal. Copper is shown with a declining price since 1970, but the recent surge in copper prices (not shown) has raised the price close to the historical average price. Similarly, the explosion in oil prices in 1979-80 now has been reversed, although the current level remains above the historical average of $12.81 (in 1985 dollars).

One test of the Hotelling relationship between prices and the rate of interest follows directly from equation (6). As shown by Feige and Geweke (1979), a simple test of the relationship is to estimate the following equation:

$$\ln(P_{t+1}/P_t) = \alpha + \beta r_t + \epsilon_t, \qquad (7)$$

where r is the rate of return on alternative investments and P is the price of the resource. The Hotelling model would imply that $\alpha = 0$ and $\beta = 1$.

A joint test of this hypothesis for copper, iron, lead, zinc, and petroleum is presented in Table 1.[5] The annual data cover the period 1870 to 1986.[6] As shown in the table, there is little support for the Hotelling model. Interest rate coefficients are negative and in all cases not significantly different from zero. Furthermore, as shown in Table 1, the Durbin-Watson statistic suggests that there is little autocor-

Table 1
Testing Hotelling's Rule

Commodity	Intercept	Rate	adj. R^2	F-test*	DW	Eqn. F**
Iron	0.0838	−1.4866	−.004	1.54	2.27	0.94
	(1.10)	(−0.84)				
Copper	0.0932	−1.8386	−.002	1.78	1.73	0.72
	(1.10)	(−0.95)				
Lead	0.1020	−2.0221	.002	1.95	1.91	0.89
	(1.23)	(−1.06)				
Zinc	0.0775	−1.3853	−.013	0.27	3.21	0.13
	(0.41)	(−0.32)				
Petroleum	0.0686	−1.2341	−.009	1.21	2.06	0.45
	(0.77)	(−0.60)				

* Testing the joint hypothesis that the intercept equals zero and the coefficient on the interest rate equals one.

** Testing the joint hypothesis that both the intercept and interest rate coefficients are zero.

Chart 1A
Real Copper Prices
(1985 Dollars)

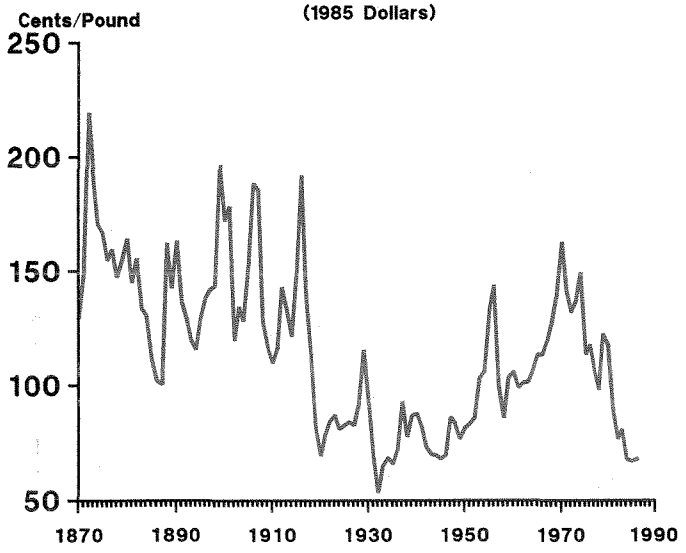


Chart 1B
Real Iron Prices
(1985 Dollars)

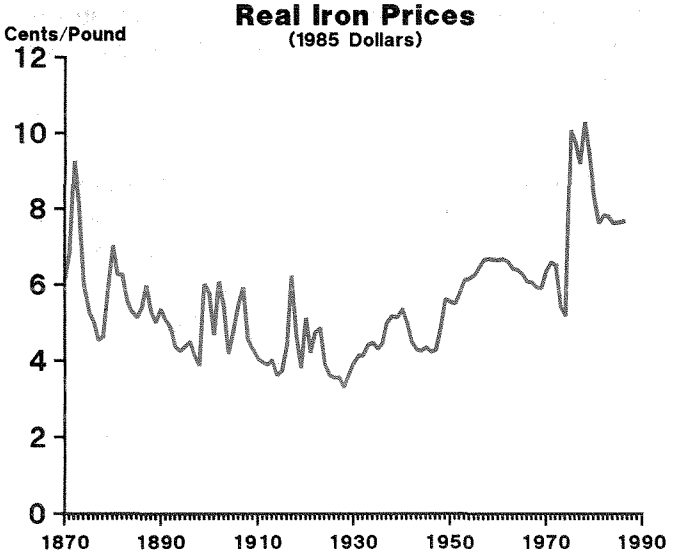


Chart 1C
Real Lead Prices
(1985 Dollars)

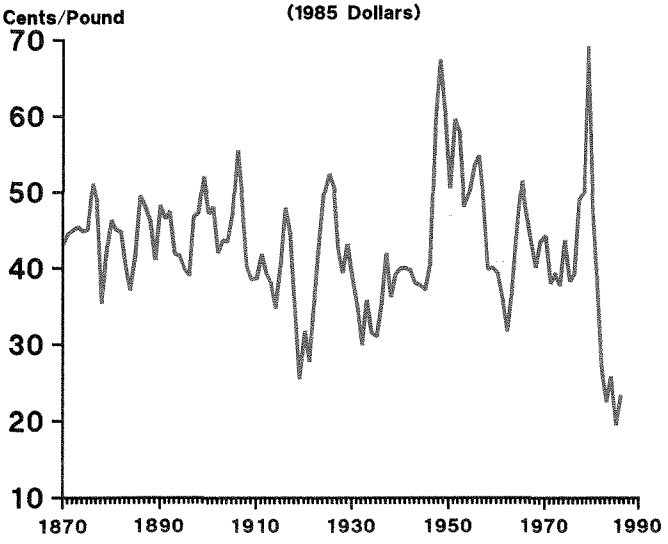


Chart 1D
Real Oil Prices
(1985 Dollars)

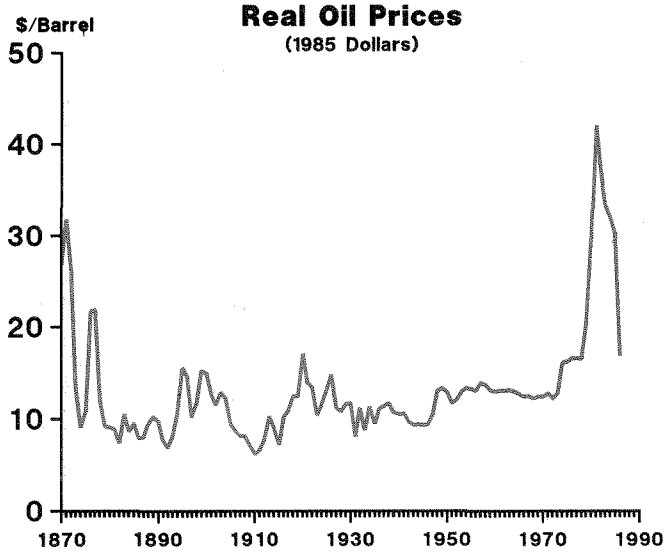
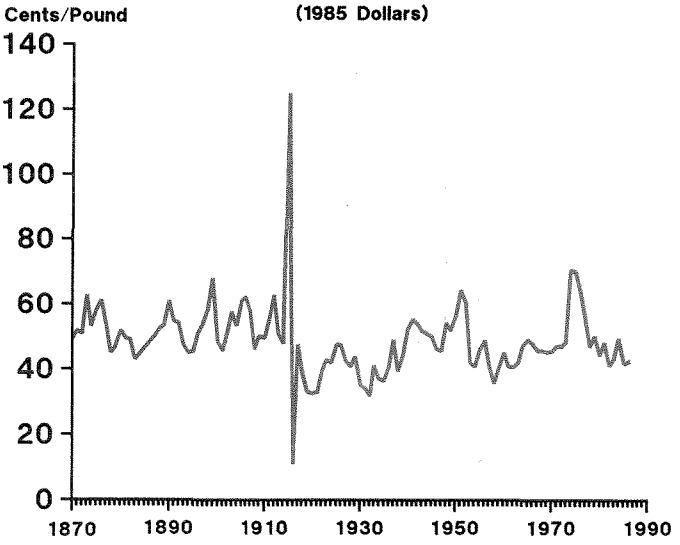


Chart 1E
Real Zinc Prices
(1985 Dollars)



relation in the errors for commodities other than zinc, indicating a lack of even short-term price trends. The lack of a significant constant term in the regressions also is consistent with a trendless process.

Similar to Feige and Geweke's findings, however, the joint hypothesis $\alpha = 0$ and $\beta = 1$ cannot be rejected. Given the theoretically incorrect signs on the coefficients, this evidence provides extremely weak support for the hypothesis. A more likely interpretation of the results would point to the low signal-to-noise ratio. The hypothesis cannot be rejected simply because the unexplained error swamps the explained variation. In fact, as shown by the second F statistic in the table, which tests the hypothesis that both coefficients are not significantly different from zero, those restrictions also cannot be rejected. Consequently, little confidence can be placed in the model's reliability.

Because of the naive specification of the model, it is not surprising that the data fail to confirm the Hotelling model. Clearly, other factors are important in shaping and explaining short-term movements. Smith (1981) and Heal and Barrow (1980), for example, have presented evidence demonstrating that arbitrage-based models that include several lagged price and interest rate terms have lower forecast errors than a simple univariate time series representation for some minerals (copper and lead in the Smith study).[7] However, the "best" specification was not consistent among models, and the best specification often was rejected as unacceptable because the coefficient on the interest rate term was negative.[8]

With more sophisticated versions of the Hotelling framework, optimal price paths need not be exponential. For example, as discussed in greater detail in the next section, optimal price paths can be shown to have U-shaped structures, given certain extraction cost schedules. In virtually all of these models, however, an upward trend would have been predicted in recent data.

Some have argued that the trendless nature of real resource prices does not violate the Hotelling model since *ex post* risk-free real interest rates have been close to zero over much of the period under study. However, even with zero real interest rates, the coefficient on the interest rate variable should not be negative in these models. Moreover, estimates of a model with *ex post* real interest rates and the *ex post* inflation rate as the explanatory variables (along with a constant) yielded similar results. Only in the case of lead was the inflation or interest rate variable significant, and in that case, the coefficients were both negative.

III. Factors Preventing Price Appreciation

The failure of Hotelling's rule to predict price behavior has been attributed to the restrictiveness of many of its underlying assumptions and may not reflect any inconsistency with intertemporal optimization.[9] In this section, several of these assumptions—no extraction costs, known reserves, no technological change, and static demand—are examined. This analysis suggests that the reason prices have failed to follow Hotelling's path is that technological innovations affecting both supply and demand consistently have made resource constraints less binding. At the same time, changes in market structure, along with these unexpected and abrupt changes in supply and demand, have contributed to the volatility of resource prices.

Extraction costs

A number of researchers have attempted to provide deterministic explanations for deviations from the Hotelling price path based on the properties of the extraction cost function [Solow and Wan (1976), Hanson (1980), and Roumasset, Isaak, and Fesharaki (1983)]. They argue that, holding technology and knowledge of the stock of the resource constant, the most easily accessible sources of the resource will be exploited first. This suggests that extraction costs should rise over time, and this will affect the resource price path [Dasgupta and Heal (1974, 1979)]. However, as demonstrated in this section, extraction costs alone—unless changed unexpectedly—do not explain why prices have not risen.

Inclusion of extraction costs in the optimal depletion problem results in a modified rule that requires prices net of marginal extraction costs to rise at the rate of interest:

$$\dot{P}(t) = r\,[P(t) - b(t)] \tag{8}$$

where b is the marginal extraction cost at time t, and r is the discount rate [Hanson (1980)]. Rearranging terms, (8) can be expressed as follows:

$$\dot{P}(t)/P(t) = r\,[1 - b(t)/P(t)]. \tag{9}$$

As can be seen by inspecting (9), if b(t) is zero, the arbitrage equation reverts to that shown in equation (6), assuming that $r = \delta$. If b is a positive constant, on the other hand, prices will grow at a slower rate than if extraction costs were zero, but the rate will rise over time, and eventually will approach the growth rate observed in (6).[10] Furthermore, if marginal extraction costs rise over time, the growth rate of prices remains below that in the

zero extraction case, and the rate of growth in prices can slow to nearly zero if costs rise faster than prices. Finally, if b(t) is discontinuous, involving discrete changes in extraction costs as the extractor moves to a lower grade of the resource, it is possible for the price path to exhibit periods of accelerating increase and periods of slowing growth.[11]

Most importantly in these models, however, prices should rise monotonically if the stock of reserves is known and fixed. As shown in (9), the only time prices fall (that is, grow at a negative rate) is when resource prices fall below marginal extraction costs, at which time production should not occur.

The price path can be U-shaped, however, if the model is further expanded to treat exploration and production costs separately, and if the initial reserve stock is small [Pindyck (1978)]. If production costs of the resource depend on both the exploration and the development of a resource, marginal costs could fall in initial stages as the resource is discovered and stocks of proven reserves grow. In this case, the decline in production costs exceeds the rise in exploration costs. In later stages, if costs of exploration continue to increase, costs would rise, forcing the price to rise as well.

Empirically, the impact of changes in extraction costs is especially difficult to isolate because of the lack of cost data. Evidence taken from various mineral census years is presented in Table 2. As can be seen in the table, real extraction costs of all minerals experienced step changes following World War II and in the 1970s.[12] Furthermore, the breakdown of costs between labor, on the one hand, and supplies and machinery, on the other, indicate a rapidly growing capital component to the cost function, suggesting exploitation of grades that are more difficult to extract.

However, these data reflect average unit costs, and do not indicate the path of marginal extraction costs. Furthermore, in the case of oil, the post-1973 observations include the effect of the rapid increase in oil prices and resulting development of high-cost energy supplies outside of OPEC. This high-cost development could not be construed as optimal development from a global standpoint, however, given that marginal extraction costs in the Middle East remained far below the marginal cost of the high-cost sources, and the Middle East had surplus capacity.

Similarly, rising commodity prices toward the end of the 1970s led to a sharp increase in exploration for other minerals, such as copper. Because of the high prices, marginal extraction costs rose significantly. When prices fell, the industry retrenched and closed down or modernized the higher-cost facilities. In recent years, prices again have risen, but unit costs are considerably lower because of the efficiency gains achieved when prices fell. Consequently, when examining extraction costs, it is important to distinguish temporarily high costs during periods of rapid price appreciation—when cost control is less apparent—and equilibrium situations where costs are closer to long-run equilibrium levels.

Other evidence by researchers finds limited support at best to indicate that rising extraction costs explain the

Table 2
Mineral Production and Extraction Costs

Year	Production Index	Costs			Unit Cost
		Payroll	Supplies & Machinery	Total	
	(1967 = 100)	(Millions of 1967 Dollars)			
1919	45	2044.8	1082.5	3127.3	69.5
1939	68	2959.7	1896.7	4856.4	71.4
1954	72	3961.1	7026.3	10987.4	152.6
1958	78	3963.0	7950.3	11913.3	152.7
1963	89	3960.9	9496.3	13457.1	151.2
1967	100	4187.0	10576.0	14763.0	147.6
1972	108	5261.1	12497.1	17758.2	164.0
1977	118	6780.1	23727.6	30507.7	259.0
1982	126	9568.0	36651.2	46219.2	366.5

Sources: Production data: Board of Governors, Federal Reserve System.
Cost data: Bureau of the Census, U.S. Department of Commerce.

trends in resource prices. Roumasset, *et. al* (1983) provide some evidence relating the oil price increases of the 1970s to rising extraction costs. Unfortunately, the marginal extraction costs they use are for U.S. producers, while the appropriate marginal extraction cost may be OPEC producers. Moreover, their results also fail to explain the pattern of prices over a longer period of time.

Rather, the absence of a rising trend in resource prices suggests two factors may be at work. First, technological change has offset rising extraction costs by developing more efficient extraction methods. Consequently, costs have been held down by productivity gains. Second, unexpected discoveries of reserves or technological progress in exploration have provided lower marginal cost extraction opportunities. Examples of discoveries of oil in Alaska, Mexico, and Columbia in the past 20 years suggest that this phenomenon is important.

Uncertain reserves

Changes in extraction and exploration technology all affect the size of the stock of proven, or extractible, reserves. This uncertainty about the reserve base contrasts with another underlying assumption in the Hotelling model. Constant real appreciation in exhaustible resource prices is derived in this model because the reserve stock is known with certainty (as are the demand function and extraction costs). In practice, however, reserves are not known with certainty and have increased dramatically over time, often in large, discrete leaps.

Table 3 presents estimates of reserves for several minerals for 1950 and 1974. Despite continued extraction and production of the minerals, reserves in 1974 were several times larger. In the case of asbestos and bauxite, for example, additions to reserves (new discoveries and extension of previously discovered reserves) were 11 to 17 times the known reserve bases in 1950. A similar pattern is found in petroleum and natural gas reserves, where additions to world reserves have tended to outstrip production.

The effect of uncertain reserves on the optimal depletion path has been examined in a number of studies [Arrow and Chang (1982), Pindyck (1980), Dasgupta and Heal (1979)]. An unanticipated shock to reserves can cause a shift among optimal paths. A sudden, unanticipated increase in proven reserves causes the price trajectory to fall to assure full resource exhaustion. Observed prices in these models fall sharply when the discovery is made.

In addition to unanticipated shocks to the reserve base, a number of these models address the impact of endogenous exploration behavior on the resource price path. As shown by Arrow and Chang (1982), exploration tends to accelerate as the stock of known reserves declines and the price of the resource rises. With major new discoveries, exploration tends to slow until scarcity again becomes important. The implied price path, therefore, is one that rises and falls, with little apparent trend.

As pointed out by Pindyck (1980), uncertainty about the stock of reserves is consistent with observed price behavior, although such uncertainty does not fully explain that behavior. Clearly, reserve shocks have played an important role in preventing the LTG scenario from occurring by consistently raising the size of the resource stock. The timing of reserve discoveries and shifts in price trajectories, however, do not coincide precisely as the theory would predict. Announcements of large new deposits have sometimes caused prices to move, but often there is little immediate response. For example, the major oil discoveries by Mexico in the mid-1970s may have contributed to pressure on OPEC in the mid-1980s, but those discoveries seemingly had little effect on prices in the mid-1970s.

In any case, the frequency with which shocks to the reserve base have occurred—either because of luck or because of the endogenous response of enhanced exploration activity—raises an important issue regarding the

Table 3
Changes in Selected Mineral Reserves, 1950–74
(Metric Tons)

Mineral	1950 Reserves	1974 Reserves	Reserve Additions as a Percent of 1950 Reserves*
Asbestos	3.9×10^7	8.7×10^7	281
Bauxite	1.4×10^9	1.6×10^{10}	1103
Chromium	1.0×10^8	1.7×10^9	1696
Copper	1.0×10^8	3.9×10^8	403
Iron	1.9×10^{10}	8.8×10^{10}	401
Lead	4.0×10^7	1.5×10^8	433
Manganese	5.0×10^8	1.9×10^9	313
Nickel	1.4×10^7	4.4×10^7	281
Tin	6.0×10^6	1.0×10^7	144
Zinc	7.0×10^7	1.2×10^8	210

Source: John E. Tilton, *The Future of Nonfuel Minerals* (Washington, D.C.: Brookings Institute, 1977), selected minerals from Table 2-2, page 10.

* Additions to reserves are calculated by adding cumulative production of the mineral between 1950 and 1974 to the difference between 1974 and 1950 reserve estimates. The reported figures are additions to reserves divided by 1950 reserves, in percentage terms.

degree to which these resources really are exhaustible. The steady rise in reserves, despite growing demand (see Charts 2a-2e, which depict a steady upward trend in consumption), may argue for decreasing scarcity value of the resource over time. If these resources are not exhaustible in practice, the failure of Hotelling's rule to predict trends in resource prices would not be surprising.

Uncertainty in demand

Technical change affecting the demand for a resource also may be an important factor in the observed failure of the Hotelling model. A key assumption of the model is that demand for the resource is known and predictable. In reality, however, dramatic changes in use patterns, the availability of alternatives, and variations in resource use intensity have caused frequent shifts in the demand for the resources. For example, the discovery of semiconductors and silicon chips significantly reduced the demand for copper wiring. Increased energy efficiency in automobiles, including substitution of aluminum and plastic for steel, had a direct impact on iron and petroleum demand.

These technological shocks result, in part, from a direct response to perceived shortages—reflected in rising prices—and from spin-off discoveries in other applications. In the short-run, most resource demand is highly inelastic. Over the longer-term, however, substitutes tend to develop that allow much greater substitutability. Often, the emergence of the substitutes leads to relatively sudden shifts in demand when the product appears, typically exceeding expectations of resource producers.[13] When these shifts occur, the expected consumption path is altered, and the optimal depletion path changes.

Such changes in demand can lead to consistent errors in the estimation of demand. Adjustments by producers to those errors then can affect the observed price path for resources. (See the accompanying Box.)

Relatively simple models of resource depletion have been developed for the case where alternative technologies exist. In the simplest form [Dasgupta and Heal (1979)], a "backstop" technology is assumed to exist in perfectly elastic supply at some price. The only effect of this modification is to affect the starting value of the arbitrage equation.

A more complicated version of the process [Kamien and Schwartz (1978)] considers the optimal depletion problem when the alternative technology is endogenously determined. The extractor must then choose a price and production schedule that maximizes profits taking into account the effect that the price level will have on encouraging alternatives. This approach, however, continues to predict monotonically rising resource prices.

A model of endogenous alternative production can generate observed price declines, however, if the assumption of complete information is relaxed. As is the case with unanticipated additions to reserves and sudden changes in technology or final demand, information limitations can lead to unstable price paths. If, for example, development of an alternative is characterized by high initial investment and low marginal costs, a sudden increase in resource prices can cause a large increase in the availability of the alternative. This increase, in turn, can force prices to fall.

Models in which prices can fall depend on the existence of uncertainty. Prices fall because supply or demand conditions change in a way the resource producer cannot anticipate. Presumably, if the producer could anticipate all responses to a given price path, the producer would follow an extraction path that would avoid these price declines. Otherwise, the arbitrage condition would be violated.

The fact that prices do fall suggests that these information problems are significant. Furthermore, the information problems are not merely the result of luck, but also because information is not often fully disseminated to the affected parties. If the supply of and demand for the resource depends on the actions of many agents, and the involved agents do not have all the information on how the other agents will react, this imperfect information can lead to unstable prices.

Consider, for example, the case of alternative production [Schmidt (1988)]. If the extractor and the alternative producer are different agents, and information is proprietary, so that: a) the extractor does not know the nature of the relationship between resource price levels and changes in the price level on the future supply of alternatives; and b) the alternative producer does not know with certainty the desired price path of the extractor, unstable pricing can be generated.

Consider the simplest case: no extraction costs, known reserves, and constant total demand for the resource and the alternative, which is a perfect substitute. The resource extractor will seek to find the price path that maximizes the present value of extraction rents, taking into account the *expected* effect of the selected price path on the supply of the alternative.

The alternative producer is assumed to choose current investment in research and development to bring the substitute on line at some future period. The optimal investment level is determined, among other factors, by the substitute producer's expectations of resource price appreciation.

In both cases, expectations are based on imperfect information. Furthermore, it is typically the case that the gestation period of an alternative product is considerable.

Chart 2A
Copper Consumption

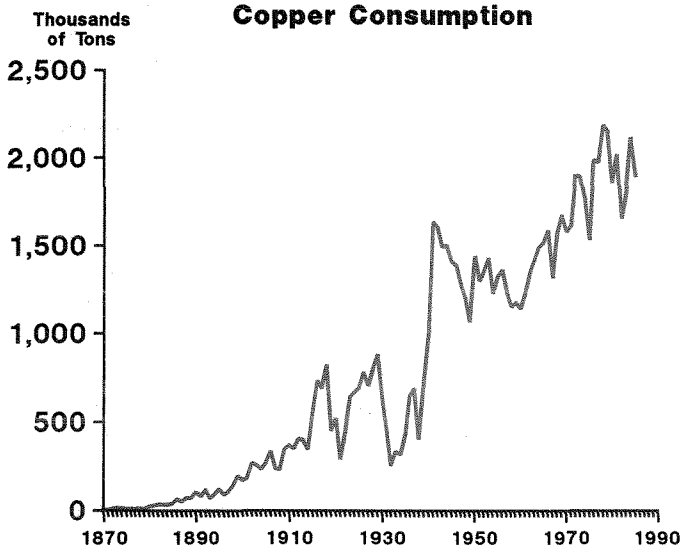


Chart 2B
Iron Consumption

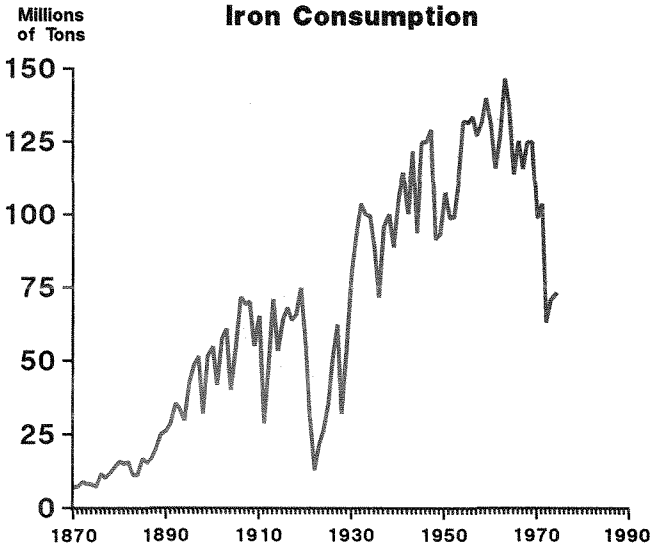


Chart 2C
Lead Consumption

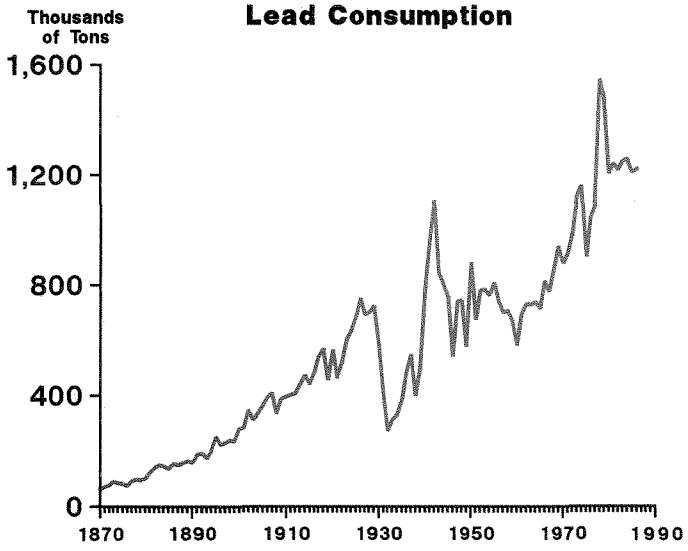


Chart 2D
Oil Consumption

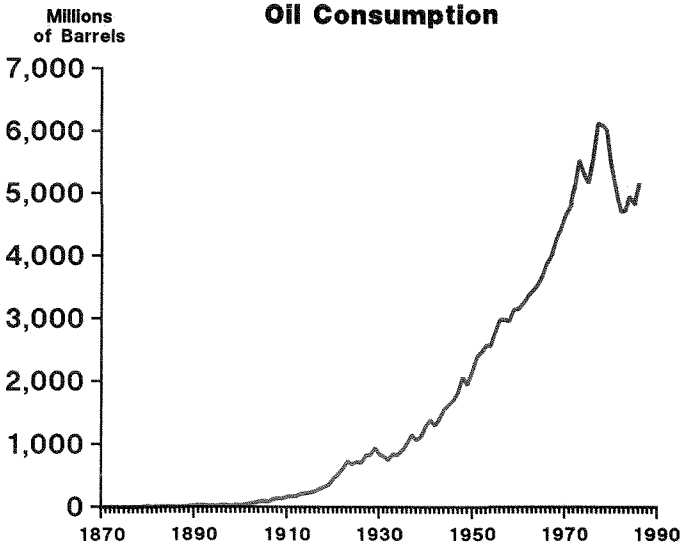
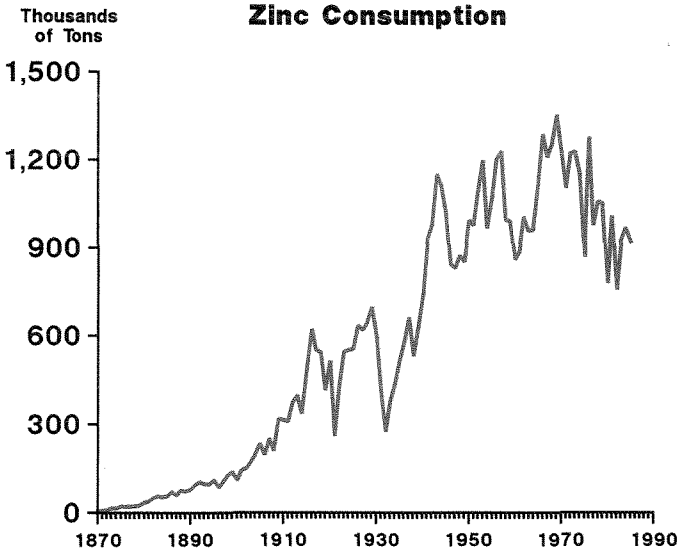


Chart 2E
Zinc Consumption



Consequently, it is possible for the extractor to underestimate the response of the alternative producer, and choose a higher initial price.[14] In this case, the response of alternative production will exceed that expected by the extractor in future periods, forcing the extractor to shift to a lower optimal price path. This shift to a lower price path would appear as a sudden drop in observed prices.

Similarly, sudden price declines might convince the alternative producer that prices will remain low, and lead to sharp cutbacks in development activities. Such a cutback would then result in lower alternative production in the future than would be expected by the extractor, possibly leading the extractor to shift to a higher price path in the future.

Unless the agents learn the true nature of each other's optimization solution—which may entail acquiring proprietary information—the process of observed price shocks can continue indefinitely. This phenomenon explains, in part, why extractive industries also frequently are major investors in the development of substitute products in order to internalize these informational externalities. For example, synfuels are developed by oil companies. Nevertheless, some substitutes inevitably emerge from non-extractive sources, surprising the market (for example, the replacement of copper by fiber optics).

Moreover, miscalculations of demand elasticities may cause extractors to raise prices so rapidly as to encourage the development of alternatives that have high fixed costs, but competitive marginal costs. The sharp increase in oil prices led to major new investments in production capability outside OPEC, where marginal production costs after drilling were low enough to continue production even after prices fell. Furthermore, the price spike was sufficiently dramatic to encourage enormous investments

Forecast Errors Caused by Misspecified Models

Incorrect forecasts of demand may be partially the result of misreading short term trends and extrapolations based on the assumption of exponential growth. Estimates of demand often are based on logarithmic specifications that explicitly force exponential growth forecasts. Population growth, which is exponential, is often used as a justification for this approach.

In practice, however, growth of consumption for many commodities has had a linear growth trend. As shown in Charts 2a–2e, the growth of consumption in the five commodities appears to have a slight exponential growth rate, but that rate is relatively small. To demonstrate the effect of this functional form misspecification, evidence on ten-year-ahead forecast errors is presented in the table for linear and exponential forecast models using rolling forecasting horizons. The models are each assumed to be reestimated each year using the last 30 years of data and then projected forward 10 years.

Results in the table suggest a strong bias towards overestimation in the exponential specifications compared to the projections of the linear model. Forecast errors made using an exponential model consistently were larger than those made by linear projections. On average, mean absolute forecast errors were 3.5 times larger in the exponential model than in the linear version. In all cases except the linear model for petroleum, the models over-estimated demand.

Demand miscalculations could occur systematically if the model is not adjusted to reflect this pattern. In fact, in the case of electricity demand forecasts, this pattern of projecting what should be a linear process with an exponential function apparently has been repeated frequently.*

*A discussion of this phenomenon is presented in Schmidt (1987), pp. 22-23.

Estimates of Consumption Trends: Comparison of Linear and Exponential Models

(Percent)

	Linear Model		Exponential Model	
Commodity	Mean Error	St. Dev.	Mean Error	St. Dev.
Iron	−23.4	87.1	−77.3	133.0
Copper	−1.5	49.5	−47.3	84.3
Lead	−2.3	38.5	−16.0	50.1
Petroleum	29.1	25.5	−33.7	55.1
Zinc	−3.0	37.3	−37.1	58.7

Note: Mean absolute forecast errors and standard deviations are calculated from a time series of ten-year-ahead forecast errors from linear and exponential models, each of which contain time and a constant as the explanatory variables. The models were reestimated each period using 30 prior years of observations, and the 10-year-ahead forecast was obtained and compared to the observed value.

in energy-efficient equipment and structure—investments that were not reversed when oil prices fell.

Consequently, the lack of perfect information among agents and the slow nature of adjustment can lead to a series of price fluctuations. Moreover, elimination of these shocks through learning often is not possible because the technology changes over time, resulting in new substitute producers with different production functions.

Market Power

Finally, a number of researchers have suggested that changes in the institutional structure of resource markets help to explain short-term movements in resource prices. Many important exhaustible resources are not sold in competitive markets. In particular, tin and petroleum are produced in cartelized market environments, and other minerals largely are owned and produced by state-owned enterprises that may have different objectives than those embedded in the Hotelling model.

Beginning with Hotelling (1931), economists have compared the implications for extraction and prices in competitive and monopolistic markets [Dasgupta and Heal (1974, 1979), Stiglitz (1974), Hnyilicza and Pindyck (1976), and Pindyck (1978)]. Researchers have found the production and price paths under the two market structures to be quite different. If marginal extraction costs are constant, monopolists will choose a price path that allows marginal revenues, rather than prices, to rise at the discount rate. Such a price path will tend to have higher initial prices and slower appreciation, leading to depletion over a longer period of time compared to the price path of a competitive producer.

This difference in optimal pricing patterns may explain part of the observed price behavior for some of the resources. For example, oil extraction has been characterized by major institutional changes. In the pre-World War II period, oil production was highly competitive in the United States—so much so, that the Governor of Texas called out the national guard to halt "cutthroat" competition in 1933, which had driven prices as low as ten cents per barrel. Beginning in 1933, prices were stabilized (and held virtually constant over long periods of time) by the prorationing policies of the Texas Railroad Commission. Production levels of Texas producers were set so as to meet refiner demand at prices then prevalent. This power waned in 1973 as imports became the marginal supply, and pricing since 1973 has reflected frequent shifts in the cartel unity of OPEC.

Similarly, a buyer's cartel has dominated the tin market for decades. Prices have risen and fallen over time with the cohesiveness of the cartel. Other industries also have had important structural changes as the market shares of government-controlled production have changed. In the case of copper, market shares have fluctuated sharply between U.S. producers (which produce in accordance with profit maximization goals) and Latin American producers (which produce to maximize foreign exchange). As the market shares change, the different objectives of the producing groups force changes in the optimal price path. Furthermore, competition for market share has at times forced production capacity to be idled as excessive supplies are dumped on the market, reducing world prices.

In cases such as these, where market institutions shift frequently, the price path can be expected to be discontinuous. Shifts in institutions reflect changes in underlying goals, changes in discount rates as different players become market-makers, and differing degrees of monopoly power. At each transition point, optimal price paths (optimal from the perspective of the dominant market participants) shift, and prices shift abruptly from the old path to the new path. In the cases of oil and copper, shifts in cartel cohesion have had immediate short-term effects on the direction of prices.

IV. Conclusions

Examination of exhaustible resource price data over the past century leads to a simple conclusion. Even with the enormous sociological, political, technological, and economic changes of the past 115 years, real prices of important exhaustible resources have not increased significantly. Certainly, those prices have at times risen or fallen sharply. But if one were attempting to forecast prices in the future, this historical behavior would nudge the forecaster toward a prediction of little future appreciation in real prices.

Does this mean that the dire consequences of resource exhaustion spelled out in LTG will occur? After all, one argument against the LTG model was the economic rationale that prices would rise and allow a gradual shift away from the resource, avoiding major disruptions. If prices do not rise, what forces are available to shift production and consumption patterns prior to the emergence of shortages?

Results of this study suggest that the corrective forces attributed to the pricing mechanism remain viable and have allowed consumption patterns to change in a nondisruptive fashion. Rather than projecting a gloomy decline in standards of living as we run out of resources, the interpretation in this article argues for the best of both worlds. Not

only is the LTG scenario unlikely, but so are the price increases associated with the Hotelling scenario.

Rather, the trendless nature of real resource prices over the past 115 years suggests that the Hotelling and LTG approaches seriously underestimated the ability of agents to substitute other resources and develop alternatives.

Shifts in consumption, changes in reserves caused by new discoveries or gains in extractive technology, technological change that affects the output mix and production function of the economy, and shifting market power of cartels all have the effect of changing the optimal depletion trajectory. Moreover, growing reserves even with rising consumption, also brings into question the degree of scarcity that truly should be attributed to these resources.

The Hotelling model predicts a rising price path when reserves do not grow, alternative technologies do not exist, and demand does not change. But history would suggest that these conditions always will change. Moreover, rising prices in the short run seem to have a larger effect on the supply of alternatives than we generally expect, given our current state of knowledge and technology.

ENDNOTES

1. Meadows, Meadows, Randers, and Behrens (1972).
2. Dasgupta and Heal (1979), p. 153
3. A "dot" above a variable indicates the rate of change.
4. Prices and interest rates are usually expressed in real terms in the theory. A similar relationship (including that tested in Section III) exists between nominal prices and nominal interest rates.
5. Similar results can be demonstrated for a wide variety of other minerals.
6. Data for the estimation for 1870-1973 were taken from *Natural Resource Commodities—A Century of Statistics*, Robert S. Manthy, Johns Hopkins University Press (1978). Data for 1973-1986 were constructed using the methodology described by Manthy from more recent publications of the original sources. The interest rate is based on railroad bonds for the early part of the series, and based on Moody's Aaa corporate bonds in more recent years.
7. Smith, in particular, experimented with interest rates of differing maturities, but found that the term of the interest rate had little impact on the general relationship to price appreciation.
8. Smith (1981), p. 110.
9. Other studies that have attempted less direct tests of the theory have found mixed results for some implications of the Hotelling model in the data. Farrow (1985), using data from mining firms, found that the *in situ* value of the resource did not follow a time path consistent with the theory. On the other hand, Miller and Upton (1985) found some support for the theory by examining cross-sectional evidence of the stock market value of U.S. domestic oil and gas firms. Schmidt (1984) also found evidence that exploration activity was consistent with a Hotelling model, with drilling responding to expected real price appreciation.
10. Note that the limit of $b(t)/P(t)$ is zero as $t \rightarrow \infty$, if $r > 0$.
11. See Hanson (1980) for a proof of this property.
12. The more recent data contrast with the findings of Barnett and Morse (1963), who found that extraction costs fell for almost all extractive products between 1870 and 1957.
13. This sequence of events characterized the pattern of oil prices in the 1979-86 period. Prices rose sharply in 1979-80, causing major investments in energy-saving technologies and the development of other sources of energy. As these sources emerged, the demand for OPEC oil diminished rapidly, leading to a sharp price decline over the 1982-86 period.
14. The producer also might choose this path if the gestation period is long and the producer's discount rate is high. In that case, the producer might attempt to capture short-term higher profits by exploiting the inelastic short-run demand, even though that action reduces future profits.

REFERENCES

Arrow, Kenneth J., and Sheldon Chang. "Optimal Pricing, Use, and Exploration of Uncertain Natural Resource Stocks," *Journal of Environmental Economics and Management* 9 (March 1982): 1–10.

Barnett, Harold J., and Chandler Morse. *Scarcity and Growth: The Economics of Natural Resource Activity.* Baltimore: Johns Hopkins Press, 1963.

Dasgupta, Partha S., and Geoffrey M. Heal. "The Optimal Depletion of Exhaustible Resources," *Review of Economic Studies, Symposium on the Economics of Exhaustible Resources* (1974): 3–28.

_____. *Economic Theory and Exhaustible Resources.* Cambridge: Cambridge University Press, 1979.

Farrow, Scott. "Testing the Efficiency of Extraction from a Stock Resource," *Journal of Political Economy* 93 (1985): 452–87.

Feige, Edgar, and John Geweke. "Testing the Empirical Implications of Hotelling's Principle," SSRI Workshop Series, no. 7924 (University of Wisconsin-Madison, Social Systems Research Institute, September 1979).

Frank, Jeff, and Mark Babunovic. "An Investment Model of Natural Resource Markets," *Economica* 51 (February 1984): 83–95.

Hanson, Donald A. "Increasing Extraction Costs and Resource Prices: Some Further Results," *Bell Journal of Economics* 11 (Spring 1980): 335–42.

Heal, Geoffrey, and M. Barrow. "The Relationship Between Interest Rates and Metal Price Movements," *Review of Economic Studies* 47 (January 1980): 161–82.

Herfindahl, Orris C. "Depletion and Economic Theory," in *Extractive Resources and Taxation*, M. Gaffney, ed. Madison, Wisconsin: University of Wisconsin Press, 1967.

Hnyilicza, Esteban, and Robert S. Pindyck. "Pricing Policies for a Two-Part Exhaustible Resource Cartel: The Case of OPEC," *European Economic Review* 8 (1976): 139–54.

Hotelling, Harold. "The Economics of Exhaustible Resources," *Journal of Political Economy* 39 (April 1931): 137–75.

Kamien, Morton I., and Nancy L. Schwartz. "Optimal Exhaustible Resource Depletion with Endogenous Technical Change," *Review of Economic Studies* 45 (1978): 179–96.

Meadows, Donnella H., Dennis L. Meadows, Jorgen Randers, and William W. Behrens III. *Limits to Growth.* New York: The New American Library, Inc., 1972.

Miller, Merton H., and Charles W. Upton. "A Test of the Hotelling Valuation Principle," *Journal of Political Economy* 93 (1985): 1–25.

Ott, Deborah A., and Donald A. Norman. "An Empirical Analysis of the Determinants of Petroleum Drilling," American Petroleum Institute Research Study no. 032 (Washington, D.C., December 1983).

Peterson, Frederick M., and Anthony C. Fisher. "The Exploitation of Extractive Resources: A Survey," *The Economic Journal* 87 (December 1977): 681–721.

Pindyck, Robert S. "The Optimal Exploration and Production of Nonrenewable Resources," *Journal of Political Economy* 86 (October 1978): 841–61.

_____. "Uncertainty and Exhaustible Resource Markets," *Journal of Political Economy* 88 (December 1980): 1203–1225.

Roumasset, James, David Isaak, and Fereidun Fesharaki. "Oil Prices Without OPEC: A Walk on the Supply-Side," *Energy Economics* 5 (July 1983): 164–70.

Schmidt, Ronald H. "Deregulating Electric Utilities: Issues and Implications," *Economic Review*, Federal Reserve Bank of Dallas (September 1987): 13–26.

_____. "Exhaustible Resource Pricing with Private Information," unpublished manuscript (September 1988).

_____. "The Effect of Price Expectations on Drilling Activity," *Economic Review*, Federal Reserve Bank of Dallas (November 1984): 1–12.

Smith, V. Kerry. "The Empirical Relevance of Hotelling's Model for Natural Resources," *Resources and Energy* 3 (1981): 105–17.

Solow, Robert M. "Intergenerational Equity and Exhaustible Resources," *Review of Economic Studies, Symposium on the Economics of Exhaustible Resources* (1974): 29–45.

Solow, Robert M. and F.Y. Wan. "Extraction Costs in the Theory of Exhaustible Resources," *Bell Journal of Economics* 7 (Autumn 1976): 359–70.

Stiglitz, Joseph E. "Growth with Exhaustible Natural Resources: Efficient and Optimal Growth Paths, *Review of Economic Studies, Symposium on the Economics of Exhaustible Resources* (1974): 123–37.