Interest Rate Expectations and the Slope of the Money Market Yield Curve

Timothy Cook and Thomas Hahn *

What determines the relationship between yield and maturity (the yield curve) in the money market? A resurgence of interest in this question in recent years has resulted in a substantial body of new research. The focus of much of the research has been on tests of the "expectations theory." According to the theory, changes in the slope of the yield curve should depend on interest rate expectations: the more market participants expect rates to rise, the more positive should be the slope of the current yield curve. The expectations theory suggests that variation in the slope of the yield curve should be systematically related to the subsequent movement in interest rates. Much of the recent research has focused on whether this prediction of the theory is supported by the data. A surprising finding is that parts of the yield curve have been useful in forecasting interest rates while other parts have not.

A novel and interesting aspect of some of the recent literature is its emphasis on the possible role of monetary policy in explaining the behavior of the yield curve. This literature views the Federal Reserve's policy instrument as the federal funds rate, and it posits that money markets rates at different maturities are strongly influenced by current and expected levels of the funds rate. In this view, explaining the behavior of the yield curve requires understanding how the Federal Reserve moves the funds rate over time. A key paper in this area (Mankiw and Miron [1986]), for example, argues that the *persistence* of changes in the federal funds rate engineered by the Federal Reserve helps explain why the yield curve from three to six months has had negligible forecasting power.

This paper surveys the recent literature on the determinants of the yield curve. It begins by reviewing the expectations theory and recent empirical tests of the theory. It discusses two general explanations for the lack of support for the theory from these tests. Finally, the paper discusses in more detail the behavior of market participants that might influence the yield curve, and the role that monetary policy might play in explaining this behavior.

I. THE EXPECTATIONS THEORY

Concepts

Two concepts central to the tests of the expectations theory reviewed below are the "forward rate premium" and the "term premium." Suppose an investor can purchase a six-month Treasury bill now or purchase a three-month bill now and reinvest his funds three months from now in another three-month bill. The forward rate is the hypothetical rate on the three-month bill three months in the future that equalizes the rate of return from the two options, given the current three- and six-month rates.[1] The forward rate calculated from the current six-month rate (R6) and the current three-month rate (R3), which we denote F(6,3), is defined as:

$$(1 + R6) = (1 + R3)(1 + F(6,3)), \text{ or} \qquad (1)$$
$$F(6,3) = [(1+R6)/(1+R3)] - 1$$

where the yields are simple unannualized yields.

Virtually all of the studies surveyed in this paper use continuously compounded yields, which enable the forward rate to be expressed as an additive (rather than a multiplicative) function of the current six- and three-month rates. Using continuously

* Timothy Cook is an economist at the Federal Reserve Bank of Richmond and Thomas Hahn is a consultant with TKH Associates, San Francisco, California. The authors benefitted from comments by Marvin Goodfriend, Ward McCarthy, Yash Mehra, and Roy Webb. They are also grateful to J. Huston McCulloch for providing interest rate data used in the paper and to the publishers of the *Washington Bond & Money Market Report* for permitting use of their survey data on interest rate expectations.

[1] The intuition behind the term "forward rate" is that a market participant who can borrow and lend at currently quoted three- and six-month rates can fix the rate at which he borrows or lends funds three months forward by an appropriate set of current transactions. See Shiller [1987, pp. 6-7].

compounded annualized yields (denoted here by lower case letters) the forward rate becomes[2]:

$$f(6,3) = 2r6 - r3 \qquad (2)$$

The "forward rate premium" is defined as the difference between the forward rate and the current short-term spot rate:

$$f(6,3) - r3 = (2r6-r3) - r3 = 2(r6-r3) \qquad (3)$$

When the maturity of the long-term rate is twice the maturity of the short-term rate, as in this case, the forward rate premium is simply twice the spread between the long- and short-term rates.

The "term premium (θ)" is generally defined as the difference between the forward rate and the corresponding expected spot rate:

$$\theta = f(6,3) - Er(3{:}t+3), \qquad (4)$$

where r(3:t+3) denotes the three-month rate three months in the future and E denotes the current expectation of that rate. Spot and forward rates not followed by a colon are measured as of time "t". Equation (4) can be rewritten in terms of the forward rate premium by rearranging terms and subtracting r3 from both sides:

$$f(6,3) - r3 = [Er(3{:}t+3) - r3] + \theta \qquad (5)$$

This expression now decomposes the forward rate premium into the expected change in interest rates and a term premium.

To illustrate these concepts, suppose the current three-month rate is 6 percent, the current six-month rate is 7 percent, and the expected three-month rate three months in the future is 7½ percent. Then the implied forward rate on a three-month security three months in the future is 8 percent and the forward rate premium is 2 percentage points. The forward rate premium can be decomposed into an expected change in the three-month rate of 1½ percentage points and an expected term premium of ½ percentage point.

An equivalent decomposition of the forward rate premium used in some papers employs the concept of "holding period yield," which is the return earned on a security sold prior to maturity. The forward rate premium can be divided into (1) the expected change in the three-month rate and (2) the difference between the expected holding period yield earned by investing in a six-month bill and selling it when it is a three-month bill three months in the future, Eh(6,3:t+3), and the return from investing in a three-month bill[3]:

$$f(6,3) - r3 = [Er(3{:}t+3) - r3] + [Eh(6,3{:}t+3) - r3] \qquad (6)$$

In the above example, the forward rate premium of 2 percentage points can be decomposed into an expected change in the short-term rate of 1½ percentage points and an expected excess return of ½ percentage point for holding six-month bills for three months rather than investing in three-month bills.

Assumptions

The "expectations theory" is based on two assumptions about the behavior of participants in the money market. The first is that the term premium that market participants demand for investing in one maturity rather than another (and issuers are willing to pay to issue that maturity) is constant over time.[4] Under this assumption equation (5) becomes:

$$Er(3{:}t+3) - r3 = -c + [f(6,3) - r3] \qquad (7)$$

where c is now a constant term premium. Note that equation (7) can be rewritten using equation (3) as:

$$r6 = \tfrac{1}{2}c + \tfrac{1}{2}[r3 + Er(3{:}t+3)], \qquad (8)$$

which says that under the expectations hypothesis the long-term rate is equal to an average of the current and expected short-term rates plus a constant which reflects the term premium.

[2] The relationship between a simple yield (R) and the corresponding continuously compounded yield (r) is:

$$(1+R) = \exp(r).$$

Hence, using continuously compounded yields, equation (1) in the text can be rewritten:

$$\exp(r6) = \exp(r3)\exp(f(6,3)),$$

which taking logarithms of both sides becomes:

$$f(6,3) = r6 - r3.$$

If we now let the lower case letters stand for *annualized* continuously compounded yields, the expression for the forward rate becomes:

$$\tfrac{1}{4}f(6,3) = \tfrac{1}{2}r6 - \tfrac{1}{4}r3, \text{ or}$$

$$f(6,3) = 2r6 - r3.$$

[3] Fama [1986, pp. 180-182] and Fama and Bliss [1987, pp. 681-682] derive this decomposition.

[4] Some papers equate the term "expectations theory" with the assumption of a constant term premium, while others include the hypothesis of rational expectations (discussed below) as part of the theory. In this paper we follow the latter procedure.

Equation (7) is the focus of most of the recent empirical work testing the expectations hypothesis. Researchers using equation (7) to test the expectations hypothesis do not know the values of Er(3:t+3). The procedure generally used to get these values is to assume that interest rate expectations are formed "rationally," so that:

$$r(3{:}t+3) = Er(3{:}t+3) + e{:}t+3, \qquad (9)$$

where e:t+3 is a forecast error that has an expected value of zero and is assumed to be uncorrelated with any information available at time t. The ideas behind the rational expectations assumption are that (1) there is a stable economic environment, (2) market participants understand this environment, (3) therefore, they should not systematically over- or under-forecast future interest rates, and (4) they should not ignore any readily available information that could improve their forecasts. This assumption specifically requires that forecast errors are not correlated with the forward rate premium at time t or its two components, the expected change in interest rates and the expected term premium. Substituting (9) into (7) yields the following regression equation:

$$r(3{:}t+3) - r3 = a + b[f(6,3) - r3] + u{:}t+3 \qquad (10)$$

Under the rational expectations assumption the error term in equation (10) is uncorrelated with the right-hand side variable so that the coefficient b can be estimated consistently. The theory predicts that b should not differ significantly from one. A significantly different value would contradict either the assumption of a fixed term premium or the rational expectations assumption.[5] An estimated coefficient of zero would be evidence that the forward rate premium has no forecasting power for the subsequent behavior of the three-month rate.

While equation (10) is the most common regression estimated in this literature, a number of other specifications have also been used.[6] An alternative used by Fama [1984a, 1986] replaces the change in the three-month rate in equation (10) with the holding period premium:

$$h(6,3{:}t+3) - r3 = a1 + b1[f(6,3) - r3] + u{:}t+3 \qquad (11)$$

The estimates of the coefficients of equation (11) provide the same information as the estimates of equation (10) because the dependent variables in the two equations sum to the common independent variable (as indicated by equation 6). Hence, b plus b1 equal one, and the sum of the constants in the two equations equals zero.[7] A value of b1 greater than zero is evidence that the current yield curve has forecasting power for the excess return earned by investing in six-month bills for three months over the return from investing in three-month bills. Given the rational expectations assumption, a value of b1 equal to one would indicate that all variation in the yield curve is due to variation in expected excess returns (i.e. the term premium) and none due to variation in the expected change in rates.

II.
Regression Estimates

Three major sets of postwar monthly interest rate data have been used by the studies surveyed in this article to estimate equations (10) and (11): (1) Treasury yields from the Center for Research in Security Prices (CRSP) at the University of Chicago, (2) Yield series constructed from Treasury rate data by a cubic spline curve-fitting technique by McCulloch [1987] and (3) Yields for Treasury and private sector securities from Salomon Brothers' *An Analytical Record of Yields and Yield Spreads*. In addition, Hardouvelis [1989] uses weekly data on Treasury bills obtained from the quotation sheets of the Federal Reserve Bank of New York, and Mankiw and Miron [1986] construct a quarterly series on three- and six-month loan rates at New York banks from 1890 through 1958. The regression results we report in this paper use the McCulloch data for Treasury rates and the Salomon Brothers data for private sector rates. We also used Treasury rates from the CRSP data and the Salomon Brothers data and found little difference in the results. All interest rates used in the paper are converted to continuously compounded annual rates as described in the Appendix I.

[5] We discuss in detail in Sections III and IV and in Appendix II the expected effect on the estimate of "b" if either of these assumptions is not valid.

[6] Campbell and Shiller [1989] derive and estimate two other specifications to test the expectations theory using a short m-period rate and a longer n-period rate. In the first the difference between the yield on an n – m period bond m periods ahead and the current yield on an n-period is regressed on the spread between the current n-period and m-period rates, where the spread is weighted by m/(n – m). In the second a weighted-average change of the m-period rate over (n – m)/m periods is regressed on the current spread between the n-period and m-period rates.

[7] This statement is correct if the long maturity (n) is equal to twice the short maturity (m). If n is not equal to 2m, then the statement is still true if the dependent variable is multiplied by an appropriate constant.

Estimates of the Standard Regression

The standard test of the expectations theory uses a long-term rate with a maturity equal to twice that of the short rate. Numerous studies have used the three- and six-month rates to calculate a three-month forward rate three months in the future and estimate the coefficients of equation (10) or a comparable equation using data over the postwar period. These include Hamburger and Platt [1975], Mankiw and Miron [1986], Mankiw and Summers [1984], and Shiller, Campbell and Schoenholtz [1983]. All these studies report coefficients for the forward rate premium that are not significantly different from zero, indicating that the yield curve from three to six months has had negligible power to forecast the changes in the three-month rate. Fama [1986] finds that the Treasury bill yield curve from six to twelve months has had no forecasting power for the subsequent six-month rate, although he does find some forecasting power for the CD yield curve from six to twelve months.[8]

The lack of support for the expectations theory using postwar Treasury bill rates at the three-, six-, and twelve-month maturities is shown in the top of Table I, which reports regression results using the McCulloch data.[9] Table I also shows little support for the theory using private security rates. The coefficients in these regressions all are positive, but only one is significant at the five percent level, and the explanatory power of the regressions is negligible. The results for the private rates are similar to those reported by Fama [1986], except that his dependent variable is the holding period premium so that his coefficients are roughly 1 minus the coefficients reported in Table I.

Mankiw and Miron [1986] estimate equation (10) from 1890 to 1914, prior to the founding of the Federal Reserve, and over four subperiods from 1914 through 1979. They find that the spread between the six- and three-month rates had substantial forecasting power for the three-month rate only in the period prior to 1914. In fact, the estimated slope coefficient in this period is only slightly below the value predicted by the expectations theory. We discuss this interesting result in more detail below.

Estimates of Non-Standard Regressions

A number of recent studies also report regression results for sections of the yield curve over which the maturity of the long-term rate is not equal to twice that of the short rate. One type of regression measures the "cumulative" predictive power of the slope of the yield curve between a one-period rate and longer-term rates at various maturities. For example, we can estimate the predictive power of the yield curve from one to six months with the regression:

$$r(1{:}t+5) - r1 = a + b[(f6,5) - r1] + u{:}t+5 \qquad (12)$$

The dependent variable in this regression is the change in the one-month rate over the following five months. The independent variable is the difference between the forward rate for a one-month bill five months in the future and the current one-month spot rate. The forward rate on a one-month bill five months in the future can be calculated from the current five- and six-month yields — hence, the notation f(6,5).[10] A coefficient of 1 for b in this regression supports the expectations hypothesis, and a coefficient less than 1 but significantly greater than zero provides evidence that the yield curve over this range has forecasting power for the subsequent movement in rates.

A second type of non-standard regression estimates the "marginal" ability of small sections of the yield curve to forecast the subsequent movements in rates over a corresponding future period. For example, the predictive power of the yield curve for the change in rates from four to five months in the future can be estimated with the regression:

$$r(1{:}t+5) - r(1{:}t+4) = a + b[f(6,5) - f(5,4)] + u{:}t+5 \qquad (13)$$

where the dependent variable is the change in the one-month rate from four to five months in the future

[8] Also, Hendershott [1984] finds forecasting power for the bill yield curve from six to twelve months after adding unexpected changes in inflation and unexpected changes in other variables to his estimated equation.

[9] The forecast horizon in these regressions is generally longer than the monthly period between observations. As a result there will likely be serial correlation in the error term of the regressions. For example, a regression of the three-month change in the three-month rate on the forward rate premium using monthly three- and six-month rates will likely generate a moving average error term of order 2 because the forecasts in months two and three are made before the error from month one's forecast is known. The standard errors provided in the tables are calculated using the consistent variance-covariance estimate from Hansen [1982] with the modification by Newey and West [1987]. For discussion of this procedure see Mishkin [1988, pp. 307-309]

[10] The formula used to calculate the forward rate on an n−m month bill m months in the future is:

$$f(n,m) = [1/(n-m)][nr(n) - mr(m)].$$

Table I

ESTIMATES OF THE STANDARD REGRESSION (n = 2m)*

$$r(m:t+m) - rm = a + b[f(n,m) - rm] + u:t+m$$

Dependent Variable	a	b	R^2	Estimation Period
Treasury Bills				
r(3:t+3) − r3	0.10	−0.15	0.00	52:1-86:8
	(0.09)	(0.19)		
r(6:t+6) − r6	0.04	0.04	0.00	52:1-86:8
	(0.17)	(0.30)		
r(3:t+3) − r3	0.13	−0.20	0.01	66:12-86:8
	(0.15)	(0.22)		
r(6:t+6) − r6	0.04	−0.01	0.00	66.12-86:8
	(0.25)	(0.32)		
Certificates of Deposit				
r(3:t+3) − r3	−0.05	0.36	0.02	66:12-86:8
	(0.17)	(0.19)		
r(6:t+6) − r6	0.07	0.52	0.06	71:10-86:8
	(0.32)	(0.26)		
Eurodollars				
r(3:t+3) − r3	−0.06	0.38	0.02	66:12-86:8
	(0.20)	(0.25)		
Commercial Paper				
r(3:t+3) − r3	−0.02	0.40	0.03	66:12-86:8
	(0.16)	(0.22)		

* Standard errors are in parentheses and are calculated as described in footnote 9. Interest rates are continuously compounded annual rates in percentage points. "m" and "n" refer to maturity in months.

and the independent variable is the difference between the one-month forward rate five months in the future (calculated from the current five- and six-month rates) and the one-month forward rate four months in the future (calculated from the current four- and five-month rates). A coefficient not significantly different from one supports the expectations hypothesis, and a coefficient less than one but significantly greater than zero provides evidence that the yield curve from four to six months has predictive power for the movement in the one-month rate four to five months in the future.

Estimates for the non-standard regressions using the McCulloch data are shown in Table II. The estimates of the cumulative regressions in the top of the table show positive and steadily declining coefficients over the money market yield curve out to six months, although only the coefficient in the first regression is significant at the 5 percent level. The results of the marginal predictive power regressions show that virtually all of the forecasting power of the bill yield curve is in the spread between the one-month ahead one-month forward rate and the current one-month spot rate.

Fama [1984a] estimates cumulative and marginal predictive power regressions using Treasury bill rates with maturities up to six months from the CRSP data from 1959 through 1982, and Mishkin [1988] repeats Fama's regressions using the same data set extended through 1986. Both studies report full sample results for the cumulative predictive power regressions roughly similar to those reported in the top of Table II. One difference is that Fama finds coefficients significant at the five percent level in his regressions covering the cumulative change in rates one, two, and three months in the future, and Mishkin finds significant coefficients in regressions covering the cumulative change in rates one and two months

Table II

ESTIMATES OF NON-STANDARD REGRESSIONS*

A. Cumulative Regressions: $r(1{:}t+n-1)-r1 = a + b[f(n,n-1)-r1] + u{:}t+n-1$

Dependent Variable	a	b	R^2
$r(1{:}t+1)-r1$	−0.18	0.50	0.09
	(0.04)	(0.12)	
$r(1{:}t+2)-r1$	−0.19	0.36	0.03
	(0.11)	(0.20)	
$r(1{:}t+3)-r1$	−0.21	0.33	0.03
	(0.14)	(0.21)	
$r(1{:}t+4)-r1$	−0.04	0.09	0.00
	(0.10)	(0.15)	
$r(1{:}t+5)-r1$	0.03	0.02	0.00
	(0.12)	(0.14)	

B. Marginal Regressions: $r(1{:}t+n-1)-r(1{:}t+n-2) = a + b[f(n,n-1)-f(n-1,n-2)] + u{:}t+n-1$

Dependent Variable	a	b	R^2
$r(1{:}t+1)-r1$	−0.18	0.50	0.09
	(0.04)	(0.12)	
$r(1{:}t+2)-r(1{:}t+1)$	−0.01	0.12	0.00
	(0.06)	(0.21)	
$r(1{:}t+3)-r(1{:}t+2)$	0.02	−0.07	0.00
	(0.04)	(0.14)	
$r(1{:}t+4)-r(1{:}t+3)$	−0.00	0.09	0.00
	(0.04)	(0.20)	
$r(1{:}t+5)-r(1{:}t+4)$	−0.03	0.62	0.02
	(0.04)	(0.34)	

*Standard errors are in parentheses. Interest rates are continuously compounded annual rates in percentage points. "n" refers to maturity in months. Estimation period is 1952:1 to 1986:8.

in the future. As with the McCulloch data, however, the full sample marginal predictive power regressions reported by Fama and Mishkin have significant coefficients only in the regression for the change in rates one month ahead, $r(1{:}t+1) - r1$, confirming that virtually all of the forecasting power of the bill yield curve is in the shortest maturities.

Fama estimates subperiod regressions for 1959 to 1964, 1964 to 1969, and 1969 to 1982, and Mishkin reports regressions for these subperiods and also for 1982 to 1986. They find that in each subperiod the difference between the one-month ahead one-month forward rate and the current spot rate had forecasting power for the movement in the one-month rate over the following month. They also find that in some subperiods—notably those in the 1960s—the difference between the two-month ahead forward rate and the one-month ahead forward rate had significant forecasting power for the change in rates one to two months in the future.

Hardouvelis [1988] uses weekly data on Treasury bill rates from 1972 through 1985 to calculate two-week forward rates at one week intervals from one to twenty-four weeks in the future. Hardouvelis estimates coefficients for cumulative and marginal forecasting regression equations over three periods corresponding to three Federal Reserve policy regimes from 1972 through October 1979, October 1979 through October 1982, and October 1982 through November 1985. In the first period the yield curve has forecasting power for only one week, while in the latter two periods the marginal forecasting power of the yield curve lasts eight or nine weeks. These results are roughly consistent with those of Fama and Mishkin, who also find that the forecasting power of the money market yield curve was weakest in the 1970s.[11] A striking feature of Hardouvelis' results is that the coefficient in the regression for the one week ahead change in rates is close to 1 in each of the three periods, which suggests that in these periods the shortest end of the yield curve behaved closely in accordance with the expectations theory.

The Forecasting Power of the Yield Curve from One to Five Years

A final set of regression results that we briefly review relate to the forecasting power of the yield curve from one to five years. Fama and Bliss [1987] find that the yield curve from one to five years has had substantial forecasting power for the change in rates over the following three or four years. For example, they find that the difference between the forward rate on a one-year Treasury security four years in the future (calculated from the current four- and five-year rates) and the current one-year rate explains 48 percent of the variance of the 4-year change in the one-year rate. Table III reports these regressions using the McCulloch data. The results are generally similar to those reported by Fama, although the explanatory power of the four-year rate change regression is smaller.

Campbell and Shiller [1989] use the McCulloch data to test a different specification of the expectations theory in which the current spread between an n-period maturity rate (such as a five-year rate) and a shorter m-period maturity (one-year) rate forecasts a weighted average change of the m-period rate over the next $n-1$ periods (4 years). They regress the weighted average change of the m-period rate on the current spread and get results similar to those of Fama and Bliss. Specifically, they find that the spread between the 4-year and 1-year rates and the spread between the 5-year and 1-year rates have significant forecasting power for the weighted average change in the one-year rate over the next 3 or 4 years.

Table III

FORECASTING POWER OF YIELD CURVE FROM ONE TO FIVE YEARS*

$$r(1{:}t+n-1) - r1 = a + b[f(n,n-1) - r1] + u{:}t+n-1$$

Dependent Variable	a	b	R^2
$r(1{:}t+1) - r1$	0.15	0.38	0.02
	(0.25)	(0.27)	
$r(1{:}t+2) - r1$	0.25	0.73	0.08
	(0.55)	(0.52)	
$r(1{:}t+3) - r1$	0.17	1.28	0.23
	(0.55)	(0.31)	
$r(1{:}t+4) - r1$	0.10	1.53	0.29
	(0.51)	(0.33)	

* Standard errors are in parentheses. Interest rates are continuously compounded annual rates in percentage points. "n" refers to maturity in years. Estimation period is 1952:1 to 1983:2.

III. EVIDENCE OF A VARIABLE TERM PREMIUM

The studies surveyed in the previous section strongly reject the expectations theory, especially when the theory is tested with the standard regression using three- and six-month or six- and twelve-month rates. The rejection of the theory implies that either (1) the term premium is not constant, (2) the rational expectations assumption is not valid, or (3) both. We discuss evidence regarding the variable term premium in this section and evidence regarding the rational expectations assumption in the following section.

Most explanations of the lack of empirical support for the expectations theory have focused on the possibility that the expected term premium is not constant, as assumed by the theory, but varies substantially over time. If the term premium is variable, the estimate of b in equation (10) will differ from the value of one predicted by the expectations theory. A number of papers have discussed the determinants of the estimated coefficient and derived expressions for the probability limit of the coefficient when the variance of the term premium is positive. (See Hardouvelis [1988, pp. 342-343] and Mankiw and Miron [1986, pp. 218-220].) The derivation of one of these expressions is shown in Appendix II. One conclusion of these papers is that, generally, the greater the fraction of the variance in the spread between the forward and spot rates due to the variance in the expected term premium—and the smaller the fraction due to the variance of the

[11] In a related paper Simon [1990] tests the forecasting power of the spread between the three-month Treasury bill rate and the overnight federal funds rate for the average funds rate over the following three months. His full sample covers the period from 1972 to 1987, and his three subperiods correspond to those in Hardouvelis's paper. Simon [p. 574, Table II] finds that the spread has forecasting power in the latter two subperiods but not in the 1970s.

expected change in rates—the lower will be the coefficient below the value of one predicted by the expectations theory.[12] If the variance of the expected change in rates is equal to the variance of the expected term premium, then the estimate of the coefficient converges to one-half.

From this perspective the relevant questions are (1) does the expected term premium vary and (2) how much does it vary relative to the expected change in rates. Evidence from a variety of sources suggests that the expected term premium does vary substantially over time and, moreover, that the magnitude of the variance is comparable to the variance in the expected change in rates.

Evidence from Holding Period Premium Regressions

As discussed in Section I, an alternative and complementary way to estimate the standard regression is to make the dependent variable the holding period premium rather than the expected change in rates:

$$h(6,3{:}t+3) - r3 = a1 + b1[f(6,3) - r3] + u{:}t+3 \qquad (14)$$

A value of b1 greater than zero is evidence that the forward rate premium has had forecasting power for the excess return from holding six-month versus three-month securities over a three-month period. A value of b1 equal to one would be evidence that virtually all variation in the yield curve is due to variation in expected returns. (This conclusion, of course, depends on the rational expectations assumption.)

Fama [1986] estimates equation (14) using one- and three-month rates, three- and six-month rates, and six- and twelve-month rates. He reports values of b1 that are close to one for bills and average a little over one-half for CDs and commercial paper. These results indicate that variation in the slope of the yield curve provides systematic information about expected excess returns. As Fama [1984a, p. 512] emphasizes, this is evidence that the current slope of the money market yield curve is influenced by expected term premiums that change over time.

Evidence from Lower Bound Estimates

A few papers have tried to measure the variance of the term premium by estimating interest rate forecasting equations using data that was available to market participants at the time of their forecasts. Startz [1982] regresses the current interest rate, r, on lagged values of spot and forward rates. He then uses the standard error of this equation as a maximum estimate or "upper bound" of the standard deviation of the market's forecast error, assuming that the set of variables used in the regression represents a minimum set of information available to market participants to forecast rates.

Startz then decomposes the spread between the forward rate and the subsequent matching spot rate (which he labels the "forward deviation") into the expected term premium (P) and the forecast error (e):

$$\begin{aligned} f - r{:}t+3 &= (f - Er{:}t+3) + (Er{:}t+3 - r{:}t+3) \\ &= \qquad P \qquad + \qquad e \end{aligned} \qquad (15)$$

The variance of (f − r:t + 3) is:

$$\mathrm{var}(f - r{:}t+3) = \mathrm{var}(P) + \mathrm{var}(e) + 2\mathrm{cov}(P,e) \quad (16)$$

The covariance of P and e is zero under the rational expectations assumption, however, because P is known at the time of the forecast and should not be correlated with forecast errors. Hence,

$$\mathrm{var}(P) = \mathrm{var}(f - r{:}t+3) - \mathrm{var}(e) \qquad (17)$$

From equation (17) we can see that if $\hat{\mathrm{var}}(e)$—the standard error of the regression squared—is an upper bound estimate for the true variance of the market's forecast error, then var(f − r:t + 3) − $\hat{\mathrm{var}}(e)$ is a *lower* bound estimate of the true variance of the term premium.

Startz calculates lower bound estimates over the period from 1953 through 1971 of the proportion of the variance of the spread between the forward rate for a one-month bill and the subsequent matching spot rate that was due to variation in the term premium. These estimates range from one-third to two-thirds over horizons from one to twelve months. This conclusion implies that lower bound estimates of the ratio of the variance of the premium to the variance of the forecast error ranged from one-half to two. Of course, this is a lower bound estimate of the ratio of the variance of the premium to the variance of the forecast error, not to the variance of the expected change in rates. Nevertheless, these results suggest that the variation in the premium is substantial.[13]

[12] Specifically, these papers find that if the correlation coefficient between the expected change in rates and the expected term premium is zero or greater than zero, the probability limit of the estimated coefficient in equation (10) is a strictly increasing function of the ratio of the variance of the expected change in rates to the variance of the expected term premium.

[13] Moreover, in the interest rate survey data discussed in the following section, the variance of the expected change in rates is less than the variance of forecast errors, in which case one-half to two would also be a lower bound estimate for the ratio of the variance of the premium to the variance of the expected change in rates.

DeGennaro and Moser [1989] employ essentially the same procedure as Startz to calculate lower bound estimates over the period from 1970 through 1982 of the proportion of the variance of the spreads between the forward rates for four- and eight-week bills and the subsequent matching spot rates that was due to variation in the term premium. Their estimates range from one-fifth to three-fifths for horizons from one to 49 weeks.

IV. Evidence on the Rational Expectations Assumption

The previous section presented evidence that a variable term premium contributes to the rejection of the expectations hypothesis in the tests reported in Section II. The remaining question is whether violation of the rational expectations assumption also contributes to the regression results. A way to evaluate this question is to use survey data to get an estimate of the market's interest rate expectations. For instance, suppose ESr(3:t+3) is the mean response from survey participants of the expected level of the three-month rate three months in the future. Then the coefficients of the standard equation can be estimated with the regression:

$$ESr(3{:}t+3)-r3 = a + b[f(6,3)-r3] + u \quad (18)$$

The variables in equation (18) are all measured at time t, the time of the survey. Consequently, the expected coefficient estimates do not depend on the rational expectations assumption. That is, if the term premium is constant, then the estimated coefficient of the forward rate premium in equation (18) should be close to 1 regardless of how expectations are formed.

A small number of studies, including Friedman [1979] and Froot [1989], have used the "Goldsmith-Nagan" survey data to estimate versions of equation (18). The survey data are based on a quarterly survey of 25 to 45 market participants on the interest rates they expect three and six months in the future. The survey was originally conducted by the *Goldsmith-Nagan Bond and Money Market Letter* and is now published in the newsletter *Washington Bond & Money Market Report*. The survey collects forecasts of the three-month bill rate, the twelve-month bill rate, and a private sector three-month rate, along with forecasts of a number of long-term interest rates. Through the March 1978 survey the private rate was the three-month Eurodollar rate. Since then, the private rate has been the three-month commercial paper rate.

There is typically about a two-week period between the time the survey forms are mailed to the respondents and the latest market close prior to publication of the responses. The average timing of the latest close prior to publication is the end of the quarter, and in estimating equation (18) we matched the survey data with the end-of-quarter data on Treasury bill rates from McCulloch and the end-of-quarter data on Eurodollar and commercial paper rates from Salomon Brothers.[14] We also used the six- and nine-month Treasury bill rates from the McCulloch data to calculate the six-month ahead forward rate for a three-month bill, f(9,6), and estimated the coefficients of an equation with the survey expected change in the three-month bill rate six months in the future as the dependent variable:

$$ESr(3{:}t+6)-r3 = a + b[f(9,6)-r3] + u. \quad (19)$$

Equation (19) can not be estimated for private sector rates because Salomon Brothers does not have the nine-month rates needed to calculate f(9,6).

The top part of Table IV shows the regression results for equations (18) and (19) using the Goldsmith-Nagan survey data. The coefficients of the forward rate premium in these regressions are all positive and significant. The low Durbin-Watson statistics, however, suggest that serial correlation is a serious problem, and inspection of the regression residuals indicated that they fall sharply in recessions. Consequently, we reestimated the regressions with a dummy variable set equal to one for all the survey dates that occurred in recessions.[15] The coefficients of the forward rate premium in these regressions range from 0.45 to 0.59 and are significant at the one percent level in the Treasury bill rate and

[14] On average over the period covered by the survey regressions the latest market close prior to publication of the survey results falls on the last day of the quarter. The average absolute difference between the latest close and the last day of the quarter is four days. We know of no reason to expect that the differences between the timing of the survey and the timing of the calculation of the forward rate premium would bias the estimate of b in equations (18) and (19). Froot [1989, p. 285, footnote 9] experiments with data sets one week and two weeks before the end of the quarter and finds that the regression results are the same as with end-of-quarter data.

[15] The dummy variable equals 1 from the fourth quarter of 1969 through the third quarter of 1970, the fourth quarter of 1973 through the fourth quarter of 1974, and the first quarter of 1980 through the third quarter of 1982. The latter period covers two recessions that are separated by three quarters.

Table IV

TEST OF THE EXPECTATIONS THEORY USING SURVEY DATA*

A. Dependent Variable: Survey Expected Change in Rates

$$ESr(m{:}t+n-m) - rm = a + b[f(n,n-m) - rm] + cD + u$$

	a	b	c	R^2	DW	Estimation Period (quarter)
Treasury Bills						
ESr(3:t+3) − r3	−0.33	0.44		0.19	0.43	69:3-86:2
	(0.11)	(0.14)				
ESr(3:t+3) − r3	−0.11	0.54	−0.96	0.70	1.24	69:3-86:2
	(0.07)	(0.11)	(0.10)			
ESr(3:t+6) − r3	−0.43	0.53		0.31	0.59	69:3-86:2
	(0.13)	(0.11)				
ESr(3:t+6) − r3	−0.08	0.50	−1.09	0.68	1.27	69:3-86:2
	(0.10)	(0.08)	(0.15)			
Eurodollars						
ESr(3:t+3) − r3	−0.67	0.75		0.42	0.67	69:3-78:1
	(0.12)	(0.19)				
ESr(3:t+3) − r3	−0.37	0.45	−0.70	0.56	1.17	69:3-78:1
	(0.10)	(0.18)	(0.23)			
Commercial Paper						
ESr(3:t+3) − r3	−0.12	0.90		0.35	1.35	78:2-86:2
	(0.10)	(0.19)				
ESr(3:t+3) − r3	0.14	0.59	−0.86	0.58	2.11	78:2-86:2
	(0.08)	(0.15)	(0.23)			

* Standard errors are in parentheses. Interest rates are continuously compounded annual rates in percentage points. "n" and "m" refer to maturity in months. D is a dummy variable set equal to 1 from the fourth quarter of 1969 through the third quarter of 1970, the fourth quarter of 1973 through the fourth quarter of 1974, and the first quarter of 1980 through the third quarter of 1982.

B. Dependent Variable: Actual Change in Rates

$$r(m{:}t+n-m) - rm = a + b[f(n,n-m) - rm] + u{:}t+m$$

	a	b	R^2	DW	Estimation Period (quarter)
Treasury Bills					
r(3:t+3) − r3	0.20	−0.35	0.02	2.56	69:3-86:2
	(0.26)	(0.42)			
r(3:t+6) − r3	−0.16	0.14	0.00	1.61	69:3-86:2
	(0.35)	(0.24)			
Eurodollars					
r(3:t+3) − r3	−0.31	0.62	0.07	1.57	69:3-78:1
	(0.28)	(0.36)			
Commercial Paper					
r(3:t+3) − r3	−0.04	0.23	0.00	2.65	78:2-86:2
	(0.42)	(0.82)			

commercial paper rate regressions.[16] The coefficients of the dummy variable are all negative and significantly different from zero. Moreover, the Durbin-Watson statistics for these regressions rise sharply, although they still indicate some serial correlation in three of the four regressions. A plausible explanation for the significance of the dummy variable coefficient is that the term premium rises in recessions. We discuss this possibility in Section VI.

For comparison with the survey regression results, the bottom part of Table IV shows estimates of the regressions over the same sample period and the same quarterly observations but with the actual change in rates as the dependent variable. The negligible explanatory power of these regressions is in sharp contrast to the survey regressions.

We can derive an estimate of the term premium implied by the survey results by subtracting the expected change in rates from the forward rate premium at the time of the survey. This estimate can be used to calculate an estimate of the relative magnitude of the variances of the premium and the expected change in rates. These variances are summarized in Table V for both Treasury bills and private securities at the three-month forecast horizon. The ratio of the variance of the premium to the variance of the expected change in rates is 1.11 for bills and about 0.65 for private securities. These numbers appear roughly consistent with the evidence from the studies reviewed in Section III that used forecasting equations to generate lower bound estimates of the variance of the premium.

The survey regression results suggest that the rational expectations assumption used in the studies surveyed in Section II is not valid for the time period covered by the survey data. To see this, note that the actual change in the three-month rate used as the dependent variable in these studies can be decomposed into the expected change in the rate plus a forecast error. If the actual change in interest rates is uncorrelated with the forward rate premium—as indicated by the regression results reported in Section II—but the expected change is positively correlated with the forward rate premium, then the survey forecast error must be negatively correlated with the forward rate premium. This is a violation of the rational expectations assumption specified by equation (9).

Table V

VARIANCE OF SURVEY EXPECTED CHANGE IN RATES AND SURVEY PREMIUM

Treasury Bills (69:3-86:2)	
Variance of Premium	0.42
Variance of Expected Change in Rates	0.38
Ratio = 1.11	
Eurodollars (69:3-78:1)	
Variance of Premium	0.29
Variance of Expected Change in Rates	0.46
Ratio = 0.63	
Commercial Paper (78:2-86:2)	
Variance of Premium	0.41
Variance of Expected Change in Rates	0.62
Ratio = 0.66	

As shown in Appendix II, a negative correlation between forecast errors and the forward rate premium reduces the coefficient of the forward rate premium in tests of the expectations theory (estimated with actual changes in rates) below the value of 1 predicted by the theory. Following Froot [1989, pp. 290-92], we can use the survey data to get estimates of the effects of the variable term premium and forecast errors on the coefficient of the forward rate premium. The probability limit of the coefficient of the forward rate premium can be written as one minus a deviation due to the variable term premium plus a deviation due to systematic expectational errors:

$$b = 1.0 - \frac{\mathrm{cov}(\theta,\mathrm{FP})}{\mathrm{var}(\mathrm{FP})} + \frac{\mathrm{cov}(\mathrm{e},\mathrm{FP})}{\mathrm{var}(\mathrm{FP})} \qquad (20)$$

where FP refers to the forward rate premium, θ refers to the term premium, and e refers to expectational errors. The survey data can be used to derive estimates of the terms on the right-hand side of equation (20). According to these estimates, shown in Table VI, roughly half the deviation from 1.0 of the

[16] Froot [1989, p.293, Table III] reports estimates of equations for the three-month ahead expected changes in the three-month bill rate, the twelve-month bill rate, and the three-month Eurodollar rate, and six-month ahead expected changes in the three-month bill rate and the twelve-month bill rate. He also finds a strong positive correlation between forward rate premiums and the survey expectations. The major difference between his results and those reported here is that he reports a negative coefficient of −0.05 in the regression for the six-month ahead forecast of the change in the three-month bill rate. Hamburger and Platt [1975, p. 191, footnote 5] find the correlation between forward rates and the survey's expected rates to be so strong that they cite it as evidence that forward rates are the market's expectations of future spot rates.

Table VI

DECOMPOSITION OF THE COEFFICIENT OF THE FORWARD RATE PREMIUM IN TESTS OF THE EXPECTATIONS THEORY*

		(1)	(2)	(3)
Instrument	Forecast Horizon	Regression Coefficient	Component Attributable to Term Premium	Component Attributable to Forecast Errors
3-Month Treasury Bill	3 Months	−0.35	0.56	−0.79
3-Month Treasury Bill	6 Months	0.14	0.47	−0.39
3-Month Eurodollar	3 Months	0.62	0.25	−0.13
3-Month Commercial Paper	3 Months	0.23	0.10	−0.67

* The construction of this table follows Froot [1989, p. 291, Table II]. The regression coefficients are from Part B of Table IV. Columns (2) and (3) are calculated using the Goldsmith-Nagan survey data from the third quarter of 1969 through the second quarter of 1986. Column (1) equals 1.0 minus column (2) plus column (3).

coefficient of the forward rate premium is due to a variable term premium and half results from the correlation of forecast errors with the forward rate premium.

The survey regression results suggest that market participants build their expectations into the yield curve, but their forecasts have been so poor at the three- and six-month horizons that the yield curve has had negligible forecasting power for the subsequent three-month and six-month rates.[17] Of course, this interpretation depends critically on the assumption that the mean of the survey forecasts is an unbiased estimate of the market forecast, and that the survey expectations can be interpreted as *determining* the current slope of the yield curve. One can imagine circumstances under which this might not be true. For example, the forecasters used in the survey might be influenced by the current shape of the yield curve in determining their interest rate forecasts, in which case the regression results would be spurious. Or they might systematically differ in their forecasts from the market in general for other reasons, perhaps because their forecasts are made public or because they are not actively involved in buying and selling securities.[18] We know of no evidence that either of these possibilities is true.[19]

A final point to make here is that there is a distinction between the specialized form of the rational expectations hypothesis used in the literature surveyed in this article—indicated by equation (9)—and the general principle of rational expectations, which is that market participants use available information efficiently in forming their expectations. Webb [1987] discusses a number of reasons why rational market participants might not behave over a given time period according to the specialized form of the hypothesis. A general point is that it is difficult to say anything definite about whether market participants have formed expectations rationally without a clear understanding of the process determining interest rate movements.

[17] The poor forecasting of market participants at the three- and six-month horizons is documented by Hafer and Hein [1989, p. 37, Table 1], who evaluate the forecasting power of both the Goldsmith-Nagan survey data and the Treasury bill futures market. They find that naive forecasts of no change in the three-month bill rate over the following three and six months do about as well as the changes forecast by the Goldsmith-Nagan survey or by the futures market. Similarly, Belongia [1987, p. 13, Table 1] finds that a forecast of no change in rates over six months does as well as the consensus forecast of a group of economists surveyed regularly by the *Wall Street Journal.*

[18] Kane [1983, pp. 117-118] emphasizes that survey respondents should be decision-makers with the authority and willingness to commit funds in support of their forecasts. He finds [p. 119] that in his survey the response of "bosses" (i.e., decision-makers) differs from the response of non-bosses. Many of the respondents in the Goldsmith-Nagan survey are senior officials in their respective organizations and would seem to fit the label of "decision-maker." We are not aware, however, of any general classification of the Goldsmith-Nagan respondents along these lines.

[19] More detailed discussions and evaluations of the Goldsmith-Nagan survey data are found in Prell [1973], Throop [1981], Friedman [1980], and Hafer and Hein [1989].

To illustrate this point, consider the behavior of the Goldsmith-Nagan forecasts over the period from late 1979 through mid-1982. In reaction to rising inflation, the Federal Reserve at the beginning of the period unexpectedly raised short-term interest rates sharply and then kept them at an unusually high level for most of the following 2½ years. The Fed's policy over this period was generally not anticipated by market participants. As a result, following the initial increase in rates the survey participants forecasted large declines in rates for several quarters in a row. (See Chart 1 in the following section.) It seems reasonable that in this episode the expectations of market participants at the three- and six-month horizons would be influenced by their judgment that monetary policy actions had driven short-term rates to a level that could not be sustained. Moreover, the expectation that rates were going to fall sharply eventually proved correct. Yet ex post the expected declines in rates at the three- and six-month horizons over this period were accompanied by large positive forecast errors. This contributed to a negative correlation over the estimation period between the expected change in rates and forecast errors, but it does not seem accurate to say that market participants formed expectations irrationally over this period.

V.
FEDERAL FUNDS RATE EXPECTATIONS AND THE BEHAVIOR OF INTEREST RATES

The regression results reported in Section II indicate that the slope of the yield curve from three to twelve months has had no forecasting power for three- and six-month rates. A puzzling aspect of the strong rejection of the expectations theory from this type of test is that it seems at odds with the standard view among money market participants that money market rates are largely determined by the current and expected levels of the shortest-term rate, the federal funds rate. A second puzzling aspect of the regression results is that the yield curve from one week to three months and from one year to five years has had forecasting power, even though the yield curve from three to twelve months has had no forecasting power. This section discusses possible explanations for these two puzzles.

Federal Funds Rate Expectations and the Mankiw-Miron Hypothesis

Market participants view the federal funds rate as the instrument used by the Federal Reserve to carry out its policy decisions. In forming expectations of the future level of the funds rate they attempt to identify Federal Reserve actions signaling changes in the funds rate target, and they attempt to forecast values of macroeconomic variables they believe influence the Fed's decisions.[20] Many studies over the past decade have found that Treasury bill rates respond to monetary policy announcements or actions that influence funds rate expectations. Similarly, many studies have found that bill rates respond to incoming news on variables—such as the money supply—that market participants believe are likely to influence policy actions. If money market rates are so sensitive to funds rate expectations, as these studies suggest, why do tests of the expectations theory using rates from three to twelve months fail so badly?

A possible answer focuses on the way market participants form expectations of the future behavior of the federal funds rate. Mankiw and Miron [1986, p. 225] suggest that at each point in time the Federal Reserve sets the short rate (i.e., the federal funds rate) at a level that it expects to maintain given the information affecting its policy decisions. They hypothesize that market participants understand this behavior and therefore expect *changes* in the short rate at any point in time to be zero: "Under this characterization of policy, while the Fed might change the short rate in response to new information, it always (rationally) expected to maintain the short rate at its current level." If this view is correct, then the whole spectrum of money market rates would adjust up and down in response to changes in the funds rate targeted by the Fed, but the *slope* of the yield curve would be unchanged. Hence, expected changes in interest rates would be negligible, and the variance of expected changes in rates would be small. This expectations behavior coupled with a variable term premium could explain the regression results in Section II. The paradox according to this explanation is that tests of the expectations theory using three- and six-month rates provide little support for the theory, even though rates at these maturities are, in fact, responding strongly to funds rate expectations.

Mankiw and Miron provide support for this argument by testing the expectations theory using three- and six-month interest rates over the 25-year period prior to the founding of the Fed and over four periods

[20] See McCarthy [1987] for a description of "Fed-watching" behavior by market participants and Goodfriend [1990] for a description of the Federal Reserve's interest rate targeting procedures.

since. They find virtually no support for the expectations theory in any of the latter periods. In the period prior to the founding of the Fed, however, they find that the yield curve from three to six months had substantial forecasting power and that the slope coefficient for this time period is only modestly below the value predicted by the expectations theory. Mankiw and Miron also present evidence that after the founding of the Fed there was a sharp deterioration in the ability of time series forecasting equations to forecast changes in interest rates three months in the future. In light of this evidence they conclude that the ability of market participants to forecast changes in short-term rates fell sharply after the founding of the Fed, resulting in a sharp rise in the ratio of the variance of the premium to the variance of the expected change in interest rates.

Cook and Hahn [1989, p. 345] catalogue the reactions of the three-month, six-month, and twelve-month Treasury bill rates to events changing federal funds rate expectations and find these reactions to be broadly consistent with the Mankiw-Miron hypothesis. These reactions are summarized in Table VII, which shows the estimated coefficients of regression equations of the form:

$$\Delta RTBi = a + b\Delta Xj + u, \qquad (21)$$

Where ΔRTBi is the change in the Treasury bill rate at maturity i and ΔXj is the change in a variable j that influences the market's funds rate expectations. The top of the table shows the reaction of bill rates to changes in the Federal funds rate target over the period from September 1974 to September 1979.[21] The middle shows the reaction of bill rates to discount rate announcements with policy content in the 1973-1985 period (i.e., announcements indicating the discount rate is being changed for reasons other than to simply realign it with market rates).[22] The bottom of the table shows the reaction of bill rates to announcements of unexpected changes in the money supply. Under the "policy anticipations hypothesis"—which is the most widely accepted explanation for this phenomenon—this reaction occurs because the Fed is expected to raise or lower the funds rate in response to deviations of money from its target path.

A striking aspect of the regression coefficients in Table VII is the relative stability from the three-month through the twelve-month maturities. This suggests that new information influencing expectations of the future level of the funds rate—even though it has a strong effect on bill rates at all maturities—has little effect on the slope of the Treasury yield curve from three to twelve months. In light of this evidence it seems plausible that the variance of the yield curve over this range has been dominated by movement in the term premium, as suggested by Mankiw and Miron.

A More General Monetary Policy Explanation for the Regression Results

While the Mankiw-Miron hypothesis can help explain the absence of forecasting power of the yield curve from three to twelve months, it is inconsistent with the evidence that the yield curve up to three months and from one to five years has had forecasting power. One can pose a more general version of the monetary policy explanation that is consistent with this evidence, and, we believe, more in line with the way market participants actually view monetary policy.

The Mankiw-Miron hypothesis assumes that the Fed reacts continuously to new information affecting its policy decisions, whereas in practice Fed policy changes are of a more discontinuous nature. That is, changes in the Fed's target for the funds rate typically occur infrequently after they are triggered by the cumulative weight of new information on economic activity and inflation. Consequently, at times there is a gap between the release of new information influencing policy expectations and when policy actually changes. This information could take the form of a policy announcement—such as a discount rate announcement—which signals an upcoming change in the funds rate target. Or it could take the form of news on an economic variable—such as the money supply or employment—that is viewed by market participants as likely to influence the Fed's target for the funds rate.

If policy and news announcements affect expectations of changes in the funds rate over a relatively short term, then the slope of the bill yield curve out to three months will vary more in response to changing interest rate expectations than will the slope from three to twelve months. In this case the reaction of market participants to such announcements could generate a pattern of funds rate expectations that is consistent with the regression results. For example,

[21] This period is unique in that the Fed controlled the funds rate so closely that market participants could identify most changes in the funds rate target on the day they were first implemented by the Fed. See Cook and Hahn [1989, pp. 332-338].

[22] Cook and Hahn [1988] find that throughout this period the Fed systematically used discount rate announcements with policy content to signal persistent changes in the federal funds rate.

Table VII

THE REACTION OF TREASURY BILL RATES BY MATURITY TO EVENTS CHANGING FEDERAL FUNDS RATE EXPECTATIONS*

(Coefficients of Treasury Bill Rate Regressions)

	3-month	6-month	12-month
Federal funds rate target changes:			
Sept. 1974-Oct. 1979	0.554 (8.10)	0.541 (10.25)	0.500 (9.61)
Discount rate announcements:			
Jan. 1973-Oct. 1979	0.26 (2.66)	0.32 (3.54)	0.30 (3.15)
Oct. 1979-Dec. 1985	0.73 (7.38)	0.61 (7.61)	0.59 (7.54)
Money announcements:			
Sept. 1977-Oct. 1979	0.072 (3.11)		0.072 (4.73)
Oct. 1979-Oct. 1982	0.364 (6.58)		0.338 (7.59)
Oct. 1982-Sept. 1984	0.190 (5.77)		0.216 (5.62)

* The funds rate target regression coefficients and the discount rate announcement regression coefficients are from Cook and Hahn [1988, 1989]. The money announcement regression coefficients are from Gavin and Karamouzis [1984]. t-statistics are in parentheses.

suppose a discount rate announcement generates expectations of a 50 basis point change in the funds rate the following week, after which no further change in the rate is expected. Under the expectations theory the effect on the slope of the yield curve out to one or two months would be considerable, but the effect on the slope from three to six months and six to twelve months would be negligible.

Hegde and McDonald [1986] find that Treasury bill futures rates have substantially outperformed a no-change forecast from one to four weeks prior to delivery, even though they have not been superior to a no-change forecast from five to thirteen weeks prior to delivery. This evidence is consistent with the hypothesis that market participants are at times able to forecast rate changes over the near-term and build these expectations into the yield curve.

A second modification one could make to the Mankiw-Miron hypothesis notes that funds rate target changes are persistent (i.e., not quickly reversed) but not permanent.[23] If so—and if market participants expected this type of funds rate behavior—then increases in the funds rate target would be associated with decreases in the slope of the yield curve between short-term rates and rates on longer maturities of five to ten years, and changes in this slope would have some forecasting accuracy.

A number of recent papers have suggested that the forecasting power of the spread between long- and short-term rates is at least partially a reflection of monetary policy. (See Bernanke and Blinder [1989], Laurent [1989], and Stock and Watson [1990, pp. 25-26].) The basic reasoning is that monetary policy has a strong influence over short-term rates but that

[23] Fama and Bliss [1987] find that the one-year Treasury rate is highly autocorrelated but slowly mean-reverting, which is consistent with the view that changes in the funds rate are highly persistent but not permanent.

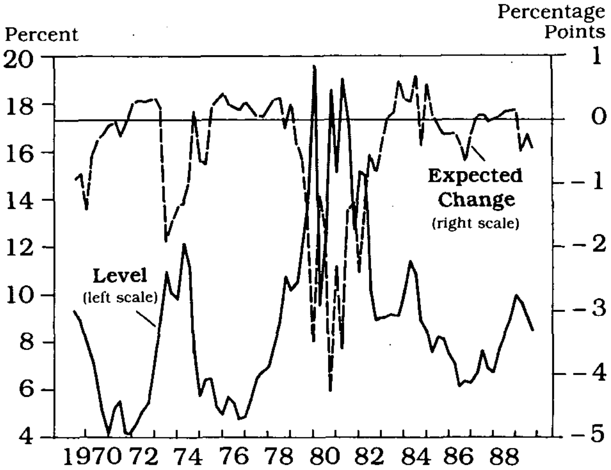
this influence diminishes at longer maturities. Hence, if short-term rates are high relative to long-term rates that is an indication that monetary policy is contractionary and a decline in inflation and interest rates is likely in the future.

This explanation for the forecasting power of the yield curve between long- and short-term rates seems especially relevant for the periods from late 1969 to late 1970, mid-1973 to mid-1974, and late 1979 to mid-1982. Near the beginning of each of these periods, the Fed raised the funds rate sharply because of its concern over accelerating inflation, and short-term rates rose well above long-term rates. (See Laurent [1989, Figure 2, p. 26].) In each period the rise in the funds rate and the downward-sloping yield curve were eventually followed by a recession and falling interest rates.[24] As illustrated in Chart 1, the Goldsmith-Nagan survey participants expected large declines in the funds rate throughout these periods.[25] These expectations had considerable accuracy at longer horizons of two to four years, even though they were not very accurate at the three- and six-month horizons.[26]

If the above adjustments to the Mankiw-Miron hypothesis are correct, then one would expect to see the slope of the yield curve out to three months and the slope from one to five years vary more than the slope from three to twelve months in response to policy actions or announcements signaling changes in the funds rate.[27] Numerous studies have provided evidence that the response of interest rates to policy actions and announcements influencing policy expectations gradually declines at maturities greater than a year. For example, Cook and Hahn [1989] find that the reaction of Treasury rates to funds rate target changes falls from 0.50 at the one-year maturity to 0.29 at the five-year maturity and 0.13 at the ten-year maturity. Likewise, several papers including Hardouvelis [1984] and Gavin and Karamouzis [1984] have reported that the reaction of Treasury rates to money announcements declines at longer maturities.

Chart 1

EXPECTED CHANGE IN AND LEVEL OF FEDERAL FUNDS RATE



The evidence at the short-end of the yield curve is more limited and somewhat more ambiguous. The reaction of interest rates to money announcements has been studied by many authors, but only a few have looked at the reaction of rates with shorter maturities than three months. These studies have found that the reaction of the one-month rate to money announcements is smaller than the reaction of longer-term money market rates, which is consistent with the notion that the yield curve out to three months varies more in response to changing policy anticipations than the curve from three months to a year. Husted and Kitchen [1985, p. 460] find that the reaction of Eurodollar rates to announcements of unexpected increases in the money supply—as determined by the coefficients of a regression similar to equation (21) above—rose from 0.28 at the one-month maturity to 0.46 at the three-month maturity and 0.44 at the six-month maturity. Hardouvelis [1984] finds that the reaction of the one-month bill rate (0.24) was smaller than the reaction of the one- to two-month forward rate (0.45), the

[24] For more discussion of these episodes see Romer and Romer [1989], who also suggest that the sharp rise in interest rates in these periods resulted from monetary policy actions intended to lower the rate of inflation.

[25] The funds rate used in Chart 1 is the average rate for the week at the end of the quarter, as determined by the weekly rate that had the greatest overlap with the last five trading days of the quarter. Special factors at the end of 1985 and 1986 caused the year-end weekly average rate to rise sharply above its level over the surrounding weeks. In these two cases Chart 1 uses the average rate for the previous week.

[26] Of course, the evidence from the survey data that market participants expected large declines in interest rates three and six months in the future in these episodes is inconsistent with the Mankiw-Miron hypothesis that market participants always forecast small changes in rates at the three- and six-month horizons. These episodes constitute a relatively small part of the period covered by the survey data, however, and they may be unique to this era. It may be that over the longer period studied by Mankiw and Miron the generalization that expected changes in interest rates at the three- and six-month horizons were generally small is an accurate one.

[27] In the case of *actual* changes in the funds rate target, however, one would expect very short maturity rates to vary as much as three- and six-month rates, since in this case the level of the funds rate rises immediately.

two- to three-month forward rate (0.40) and the three- to six-month forward rate (0.35). (The sample period for the studies cited in this and the following paragraph is from late 1979 or early 1980 to late 1982.)

Surprisingly, however, the money announcement literature indicates that the reaction of the one-day funds rate and the one-week bill rate to money announcements is not smaller than the reaction of longer-term money market rates. Hardouvelis finds a coefficient of 0.38 for the one-day funds rate, and Roley and Walsh [1985] find a coefficient of 0.43 for the one-day funds rate, 0.37 for the one-week bill rate, and 0.36 for the three-month bill rate. A possible explanation for the relatively large response of the one-day and one-week rates is that under the lagged reserve accounting system prevailing prior to February 1984 the weekly money announcement provided information on the current statement week's aggregate demand for reserves that influenced the expected average funds rate for the statement week—and, hence, the one-week bill rate—independent of any policy anticipations effect.[28]

VI.
Behavior of the Term Premium

The evidence presented in Section III suggests that a variable term premium plays an important role in explaining the negligible forecasting power of the yield curve from three to twelve months. This conclusion raises a final set of questions. First, how does the term premium behave on average and at different maturities? Second, what causes the term premium to change over time? The literature in this area—especially regarding the second question—is voluminous, yet largely inconclusive. Our purpose here is simply to provide a brief review of this literature and the difficulties researchers have faced in trying to measure the term premium.

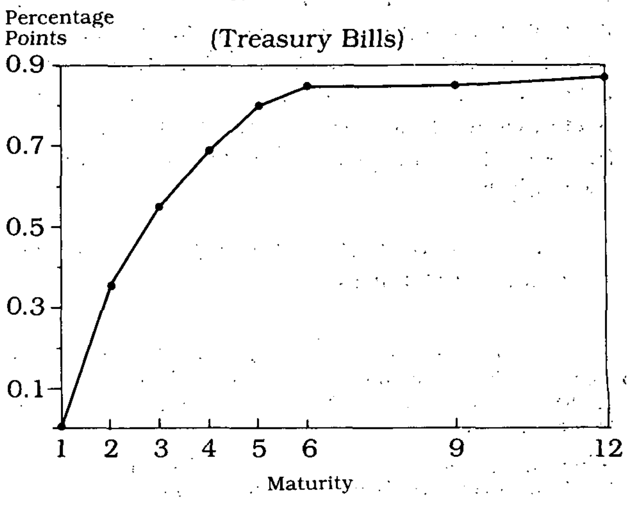
The Average Term Premium in the Money Market

Researchers have generally estimated the average term premium in the money market by calculating over long periods of time the average excess returns from holding n-month securities for m months versus the return from holding m-month securities. The most common practice is to use the one-month rate for the benchmark (m = 1). The literature in this area has found a positive average term premium in the Treasury bill market at all maturities. The average term premium rises sharply for the first couple of months, increases at a decreasing rate out to around five or six months, and then flattens out. This behavior is illustrated in Chart 2 which shows estimates of the average premium at different maturities using the McCulloch data.[29] McCulloch [1987] shows that the CRSP data provide a similar picture of the relationship between the average premium and maturity in the bill market.[30] Researchers have found no evidence of a significant term premium, on average, in the markets for private money market securities such as commercial paper, CDs, and Eurodollar CDs. Fama [1986, p. 178, Table 1], for example, finds that average term

Chart 2

ESTIMATES OF AVERAGE TERM PREMIUM

(Treasury Bills)



[28] Along these lines, Strongin and Tarhan [1990, pp. 151-152] conclude that "The response of the Fed funds rate [to money announcements] cannot be explained by either [policy anticipations or expected inflation], but instead by the peculiar way money shocks are transmitted to the reserve market under lagged reserves accounting."

[29] For maturities up to six months, the estimate of the average term premium in Chart 2 is the average of the annualized returns earned by holding a Treasury bill at a given maturity for one month less the returns on a one-month bill. For maturities of nine and twelve months, the estimate of the average premium is the average return from holding a nine- or twelve-month bill for three months less the return on a one-month bill. (In the latter two cases three months is the shortest holding period yield that can be calculated using the McCulloch data.)

[30] Fama [1984b] provides evidence using the CRSP data that the premium declines between nine and ten months. McCulloch [1987], however, shows that this evidence results from the small bid-ask spread on nine-month Treasury bills in the period from 1964 through 1972 when the Treasury was issuing new bills at that maturity. He concludes that the description that best fits the CRSP data is that the premium rises monotonically to about five months and has no further significant change.

premiums for privately issued securities over his whole sample period from January 1967 to January 1985 are close to zero. Fama, however, also divides his sample into months when the yield curve was monotonically upward sloping and months when the yield curve was "humped" (i.e initially rising and then falling). He finds that in months when the private yield curve was upward sloping, the average term premium was positive and rose with maturity, and in months when the private yield curve was humped, the average term premium initially was positive but then became negative at the longer maturities.

One type of explanation for the positive average premium in the bill market focuses on the preferences of individual investors. Hicks [1946, pp. 144-52] argues that investors have a preference for shorter-term securities because of the greater price volatility of long-term securities when interest rates change. In contrast, he reasons that many borrowers have a preference for long-term borrowing. Hence, there is a "constitutional weakness" on the long side of the market such that in equilibrium investors have to be offered a premium to invest in longer-term securities. In a similar vein, Kessel [1965, p. 45] argues that the market has a preference for shorter-term securities because of their greater liquidity: "The shorter the term to maturity of a security, the smaller is its vulnerability to capital loss, and hence the greater its liquidity and the smaller the yield differential between that security and money."[31]

More recent papers have analyzed the term premium in the context of individuals who maximize the expected utility of their lifetime consumption.[32] An idea that comes out of this literature is that the term premium is likely to be positive if unexpected capital losses (i.e. positive future interest rate surprises) are generally positively correlated with negative consumption surprises. In other words, investors are likely to demand a higher yield on long-term securities if they are likely to experience unexpected capital losses when times are unexpectedly bad and their marginal utility of consumption is relatively high.

A second explanation for the positive average premium in the bill market, suggested by Rowe, Lawler, and Cook [1986, pp. 9-10] and Toevs and Mond [1988], focuses on the unique characteristics of the market. Treasury bills can be used to satisfy numerous institutional and regulatory requirements, such as serving as collateral for tax and loan accounts at commercial banks and satisfying margin requirements on futures contracts. To the extent that the holding period for these purposes tends to be short, investors might prefer to minimize capital risk by holding short-term bills to satisfy them. Moreover, the Treasury is not sensitive to interest rates at different maturities in its supply of bills, so there is no pressure from the supply side to equalize the expected cost of issuing bills at different maturities. In contrast, banks might be expected to issue more three-month CDs if the expected cost of raising funds this way were systematically lower than the cost of raising funds by issuing six-month CDs, and this behavior would raise the three-month rate relative to the six-month rate and reduce the premium.

Measuring the Behavior of the Term Premium over Time

A number of approaches have been taken in the literature to measure the behavior of the term premium over time. The simplest approach is to assume that the forward rate premium is an accurate representation of the term premium. Suppose the expected change in rates at any point in time is negligible so that the forward rate premium is completely dominated by variation in the term premium. Then, as Fama [1986, p. 187] suggests, the forward rate premium can "provide a direct picture of the behavior of the expected term...premium." As discussed earlier, however, the Goldsmith-Nagan survey data suggest that at times market participants have expected large changes in rates. If so, then in these periods the forward rate premium provides an inaccurate picture of the term premium.

A second approach to measuring the term premium is to subtract the expected interest rate level from the Goldsmith-Nagan survey from the comparable forward rate at the time of the survey. Chart 3 shows (a) the difference between the forward rate on three-month bills three months ahead and the expected three-month bill rate three months ahead and (b) the difference between the forward rate on three-month bills six months ahead and the expected three-month bill rate six months ahead.[33] The chart shows a clear

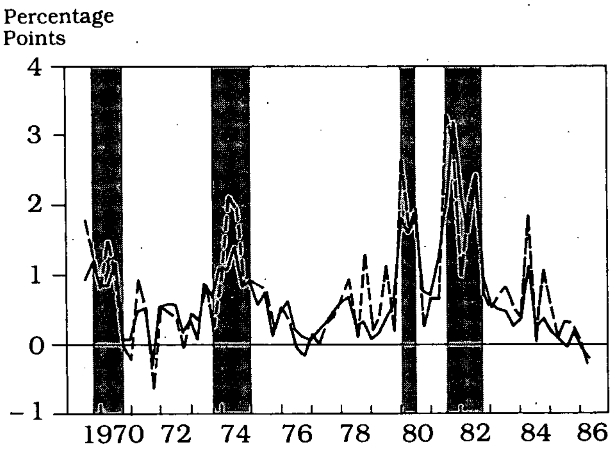
[31] There was a huge amount of literature on the expectations theory and the term premium in the 1960s and early 1970s. For a review of this literature see Van Horne [1984, Chapter 4] and Malkiel [1970].

[32] For an example of this approach see Sargent [1987, pp. 102-105]. Abken [1990] discusses this literature.

[33] The vertical lines in Charts 3, 4, and 5 show quarterly business cycle peaks and troughs. Peaks are the fourth quarter of 1969, fourth quarter of 1973, first quarter of 1980, and third quarter of 1981. Troughs are the fourth quarter of 1970, the first quarter of 1975, third quarter of 1980, and the fourth quarter of 1982.

Chart 3

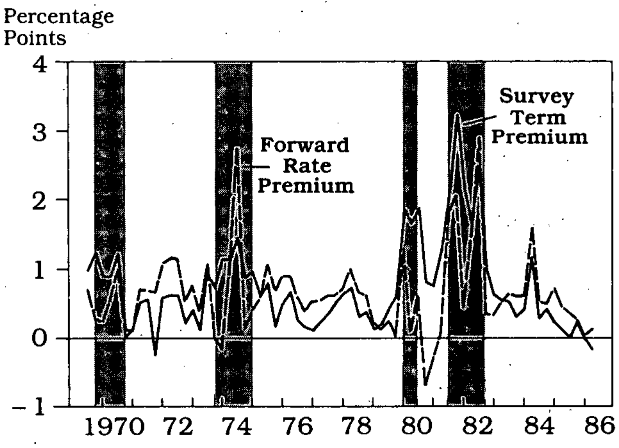
SURVEY TREASURY BILL TERM PREMIUMS



Note: The dashed line is the difference between the forward and expected three-month rates six months ahead. The solid line is the difference between the forward and expected three-month rates three months ahead. The shaded areas represent recessions (peak to trough).

Chart 4

FORWARD RATE PREMIUM AND SURVEY TERM PREMIUM



Note: The forward rate premium is the forward rate on three-month Treasury bills three months ahead minus the current three-month spot rate. The survey term premium is the forward rate minus the expected three-month spot rate three months ahead. Shaded areas represent recessions.

tendency for the survey term premiums to be relatively high in recessions and low in expansions. (This tendency was captured in the regression results reported in Section IV by the signficant negative coefficient of the recession dummy variable.)

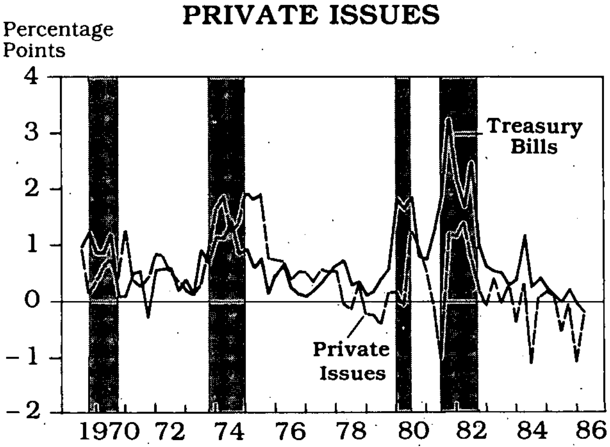
Chart 4 shows that the survey term premium and the forward premium generally move together, but large differences occasionally occur when the survey indicates large expected declines in rates. The most striking difference in the two estimates of the term premium is in the period from late 1979 though mid-1982 when interest rates were unusually high and were expected to fall by the survey participants. In this situation the survey term premium rose well above the forward rate premium. Chart 5 compares the survey premiums for bills and private money market securities. The private premium generally follows the same pattern as the bill premium—rising in recessions and falling in expansions—although occasionally there are significant differences in the two premiums.

A third estimate of the term premium is the forward deviation, i.e. the difference between the forward rate and the subsequently realized spot rate. As discussed in Section III, the forward deviation can be decomposed into an expected term premium and an interest rate forecast error. Both the Goldsmith-Nagan survey data and futures market data indicate that market participants have had little ability to forecast rates at the three- and six-month horizons. As a result, the forward deviation is an extremely volatile measure dominated by interest rate forecast errors.

A final approach used to estimate the term premium is to employ regression methods to generate "expected" interest rates with data available to market participants at the time of the forecast. These

Chart 5

SURVEY TERM PREMIUMS ON TREASURY BILLS AND PRIVATE ISSUES



Note: The term premiums are the difference between the forward and expected three-month spot rates three months ahead. Shaded areas represent recessions.

estimates can then be used along with contemporaneous forward rates to calculate estimates of the term premium. Since these forecasting equations have little power to predict changes in interest rates, one might expect this approach to provide estimates of the term premium that are similar to the forward rate premium. We are not aware, however, of any studies that have made this comparison.

Determinants of the Variable Term Premium

The estimates of the term premium shown in Charts 3-5 suggest that term premiums in the money market tend to be low in periods of economic expansions and high in periods of weakness. This is consistent with a recent conclusion by Fama and French [1989, p. 43] that term premiums "move opposite to business conditions," This is not a universally accepted description of the behavior of term premiums, however. Numerous variables are correlated with economic conditions, and the charts might be capturing a correlation of the premium with some other variable such as the level of interest rates.[34] Moreover, even if one accepts the description that term premiums move opposite to business conditions as accurate, there is still no generally accepted explanation for why term premiums rise around recessions and fall in expansions.

Numerous papers have attempted to make judgments about the determinants of the term premium by regressing one of the measures of the premium described above on various possible explanatory variables.[35] Two explanatory variables often included in these regressions are the volatility of interest rates and the level of interest rates. Hicks [1946] reasons that the term premium is compensation for the capital risk resulting from interest rate movements and, therefore, increases in interest rate volatility should increase the premium demanded by investors. The argument that the level of rates should be a determinant of the term premium is generally associated with Kessel [1965, pp. 25-26]. He argues that short-term bills are better money substitutes than long-term bills, and since an increase in interest rates increases the cost of holding money, the yield on short-term bills should fall relative to the yield on long-term bills when rates rise.

Papers that find the volatility of interest rates to be a significant determinant of the premium include Fama [1976], Heuson [1988], and Lauterbach [1989].[36] Papers that find the level of rates to be a determinant of the premium include Kessel [1965], Pesando [1975], and Friedman [1979]. Other explanatory variables that have been used in studies of the premium include the relative supplies of securities at different maturities, the unemployment rate, industrial production, and the spread between yields on high- and low-risk securities. As Shiller [1987, pp. 56-57] concludes in his survey article on the term structure of interest rates: "It is difficult to produce a useful summary of [the] conflicting results" from the empirical studies of the term premium. The main conclusion is that no consensus has emerged in the literature on what macroeconomic variable the term premium is most closely related to or on why the term premium varies so much.[37]

VII. CONCLUSION

The studies surveyed in this paper find that over long periods of time the yield curve from three to twelve months has had negligible power to forecast interest rates three and six months in the future. The

[34] For example, on the basis of the Goldsmith-Nagan survey data and a chart similar to Chart 3, Froot [1989, p. 299, Figure 1] concludes that "the surveys suggest that term premia rose substantially during periods of high interest rate volatility [p. 300]." He also concludes that the survey premia "are highly positively correlated with nominal interest rates and inflation [p. 303]." Friedman [1979, p. 972] on the basis of regressions using the Goldsmith-Nagan data from September 1969 through March 1977 concludes that "the results make clear that the basic relation is between the term premium and interest rate levels, not economic activity . . .".

[35] For example Friedman [1979] uses the premium calculated from the survey data as the dependent variable in his regressions, Kessel [1965] uses the forward deviation in his regressions, and Pesando [1975] estimates interest rate forecasting equations to calculate an estimate of the term premium to use as the dependent variable in his regressions.

[36] Fama [1976] assumes that the expected real return on a one-month Treasury bill is constant over his sample period and therefore concludes that his measure of the volatility of interest rates is capturing the positive effect of inflation uncertainty on expected term premiums on multimonth bills.

[37] One possibility that we do not discuss here is that the variable term premium results from factors related to specific Treasury bill issues and maturities. For example, Park and Reinganum [1986] find that Treasury bill yields maturing at the end of the month and especially at the end of the year have lower yields than surrounding maturities, and Nelson and Siegel [1987] find evidence of both maturity-specific and issue-specific effects on bill yields. Also, it is also widely believed in the financial markets that a shortage or abundance of a particular bill issue can cause that issue's yield to differ significantly from the yields on surrounding maturities. The McCulloch data used in this paper, however, are constructed from a curve-fitting technique and therefore should generally not be affected by such factors. Moreover, the evidence presented earlier in the paper suggests the variable term premium is pervasive throughout the money market and not just due to special factors operating in the bill market.

yield curve out to three months has had forecasting power for the one-month ahead rate, however, and the yield curve from one to five years has had forecasting power for the one-year rate over the following three or four years.

The research in this area has suggested two broad reasons for the poor forecasting power of the yield curve from three to twelve months. The first is that the variation in the term premium at the three- and six-month horizons has been substantial relative to the variation in the expected change in rates. The second is that even when market participants have forecasted significant changes in interest rates at the three- and six-month horizons, their forecasts have been poor at these horizons.

An understanding of how market participants form monetary policy expectations may provide insight into some of the results in this literature. A monetary policy explanation for the poor forecasting power of the yield curve from three to twelve months is that market participants expect changes in the Fed's federal funds rate target to be persistent. According to this explanation, three-, six-, and twelve-month rates tend to move the same amount in reaction to changes in the funds rate target and, therefore, changes in the slope of the yield curve over this range are dominated by movement in the term premium. The forecasting power of the yield curve out to three months may reflect the ability of market participants to forecast over short horizons the reaction of the Fed to new information influencing its policy decisions. And the forecasting power of the yield curve from one to five years may partially reflect the belief of market participants that over longer periods of time changes in the funds rate target are likely to be reversed, especially after the Fed has raised the funds rate sharply in reaction to rising inflation.

The evidence cited in this paper in favor of a monetary policy explanation for the regression results is limited, however, and the explanation has not been universally, or even widely, accepted. There is also no general agreement on why the term premium varies so much, although the Goldsmith-Nagan survey data strongly suggest that the premium rises when economic conditions deteriorate. A brief assessment of the literature surveyed in this paper is that it has done a good job of documenting the forecasting power of various parts of the yield curve, and it has suggested some plausible and interesting answers to some of the major questions in this area. A comprehensive explanation for these questions, however, awaits further research.

APPENDIX I
INTEREST RATE CONVERSIONS

All interest rates in the paper are continuously compounded annual rates. No conversion is necessary for the McCulloch Treasury bill rate data, which come in this form. Three-month Treasury bill rate forecasts from the Goldsmith-Nagan survey are on a 360-day discount basis, however, as are all commercial paper rates used in the paper. Eurodollar, CD, and federal funds rates are quoted on a simple interest 360-day basis. Prices per $1 of return are calculated from the quoted yields, Q, using the formulas:

$$P = 1 - [(Q/100)/(t/360)] \qquad (1)$$

for bills and commercial paper and

$$P = 1/[(Q/100)(t/360) + 1] \qquad (2)$$

for Eurodollars, CDs, and federal funds rates. "t" is the days from settlement to maturity: 30, 90, and 180 days for commercial paper, CDs, and Eurodollars; 91 days for Treasury bills; and 1 day for federal funds. Prices are converted to continuously compounded yields using the formula:

$$r = -(365/t)\ln P \qquad (3)$$

where lnP is the natural logarithm of P.

Appendix II

The Coefficient of the Forward Rate Premium in the Standard Regression

The standard regression equation is:

$$r(3{:}t+3) - r3 = a + b[f(6,3) - r3] + u{:}t+3 \qquad (1)$$

To simplify the notation rewrite this as:

$$\Delta r = a + b(FP) + u{:}t+3 \qquad (2)$$

where Δr is the rate change and FP is the forward rate premium. Recall also that the forward rate premium can be decomposed into the expected rate change and the expected term premium, θ, and the actual change in the interest rate can be decomposed into the expected change and a forecast error, e:

$$FP = E(\Delta r) + \theta \qquad (3)$$

$$\Delta r = E(\Delta r) + e \qquad (4)$$

The probability limit (abbreviated as plt) of the ordinary least squares estimate of b in equation (2) is:

$$\text{plt } b = \frac{\text{cov}(FP,\ \Delta r)}{\text{var}(FP)} \qquad (5)$$

Substituting (3) and (4) into (5) yields:

$$\begin{aligned}\text{plt } b &= \frac{\text{cov}[E(\Delta r)+\theta,\ E(\Delta r)+e]}{\text{var}[E(\Delta r)+\theta]} \\ &= \frac{\text{cov}[E(\Delta r)+\theta,\ E(\Delta r)]}{\text{var}[E(\Delta r)+\theta]} \\ &\quad + \frac{\text{cov}[E(\Delta r)+\theta,\ e]}{\text{var}[E(\Delta r)+\theta]} \qquad (6)\end{aligned}$$

Suppose the rational expectations assumption is valid. Then the forecast error, e, is not correlated with information available at the time of the forecast, which includes the expected change in rates and the expected premium. Then the second term on the right-hand side of equation (6) equals 0 and we get the expression:

$$\text{plt } b = \frac{\text{cov}[E(\Delta r)+\theta,\ E(\Delta r)]}{\text{var}[E(\Delta r)+\theta]} \qquad (7)$$

Denote the variance of x as $\sigma^2(x)$, the standard deviation as $\sigma(x)$, and the correlation coefficient between x and y as ρ. Recall that $\text{cov}(x,y) = \rho\sigma(x)\sigma(y)$. Then equation (7) can be written:

$$\begin{aligned}\text{plt } b &= \frac{\sigma^2[E(\Delta r)] + \text{cov}[\theta, E(\Delta r)]}{\sigma^2[E(\Delta r)] + \sigma^2(\theta) + 2\text{cov}[E(\Delta r),\theta]} \\ &= \frac{\sigma^2[E(\Delta r)] + \rho\sigma(\theta)\sigma[E(\Delta r)]}{\sigma^2[E(\Delta r)] + \sigma^2(\theta) + 2\rho\sigma[E(\Delta r)]\sigma(\theta)} \qquad (8)\end{aligned}$$

This is the expression in Hardouvelis [1988, p. 342]. It is also similar to the expression in Mankiw and Miron [1986, p. 219], except that the term premium in their framework is equal to one-half the premium above. Note that the probability limit of b is one if the premium is a constant and one-half if the standard deviation of the term premium equals the standard deviation of the expected change in rates.

Now substitute equation (4) into (5) to get:

$$\begin{aligned}\text{plt } b &= \frac{\text{cov}(FP,\ E(\Delta r) + e)}{\text{var}(FP)} \\ &= \frac{\text{cov}(FP,\ E(\Delta)) + \text{cov}(FP,e)}{\text{var}(FP)} \qquad (9)\end{aligned}$$

Substituting (3) into (9) yields:

$$\begin{aligned}\text{plt } b &= \frac{\text{cov}(FP,\ FP-\theta) + \text{cov}(FP,e)}{\text{var}(FP)} \\ &= \frac{\text{var}(FP) - \text{cov}(FP,\theta) + \text{cov}(FP,e)}{\text{var}(FP)} \\ &= 1 - \frac{\text{cov}(FP,\theta)}{\text{var}(FP)} + \frac{\text{cov}(FP,e)}{\text{var}(FP)} \qquad (10)\end{aligned}$$

Equation (10) says that a positive correlation of the term premium with the forward rate premium or a negative correlation of forecast errors with the forward rate premium will reduce the coefficient of the forward rate premium below the value of one predicted by the expectations theory.

References

Abken, Peter A. "Innovations in Modeling the Term Structure of Interest Rates." Federal Reserve Bank of Atlanta *Economic Review* LXXV (July/August 1990): 2-27.

Belongia, Michael T. "Predicting Interest Rates: A Comparison of Professional and Market-Based Forecasts." Federal Reserve Bank of St. Louis *Economic Review* 69 (March 1987): 9-15.

Bernanke, Ben S., and Alan S. Blinder. "The Federal Funds Rate and the Channels of Monetary Transmission." Working Paper 89-10, Federal Reserve Bank of Philadelphia, February 1989.

Campbell, John Y., and Robert J. Shiller. "Yield Spreads and Interest Rate Movements: A Bird's Eye View." Working Paper 3153. National Bureau of Economic Research Working Paper Series, October 1989.

Cook, Timothy, and Thomas Hahn. "The Information Content of Discount Rate Announcements and Their Effect on Market Interest Rates." *Journal of Money, Credit and Banking* 20 (May 1988): 167-180.

————. "The Effect of Changes in the Federal Funds Rate Target on Market Interest Rates in the 1970s." *Journal of Monetary Economics* 24 (November 1989): 331-351.

DeGennaro, Ramon P., and James T. Moser. "Variability and Stationarity of Term Premia." Working Paper 89-16. Federal Reserve Bank of Chicago, September 1989.

Fama, Eugene F. "Inflation Uncertainty and Expected Returns on Treasury Bills." *Journal of Political Economy* 84 (June 1976): 427-448.

————. "The Information in the Term Structure." *Journal of Financial Economics* 13 (December 1984a): 509-528.

————. "Term Premiums in Bond Returns." *Journal of Financial Economics* 13 (December 1984b): 529-546.

————. "Term Premiums and Default Premiums in Money Markets." *Journal of Financial Economics* 17 (September 1986) 175-196.

Fama, Eugene F., and Robert R. Bliss. "The Information in Long-Maturity Forward Rates." *The American Economic Review* 77 (September 1987): 680-692.

Fama, Eugene F., and Kenneth R. French. "Business Conditions and Expected Returns on Stocks and Bonds." *Journal of Financial Economics* 25 (November 1989): 23-49

Friedman, Benjamin M. "Interest Rate Expectations Versus Forward Rates: Evidence From an Expectations Survey." *The Journal of Finance* 34 (September 1979): 965-973.

————. "Survey Evidence on the 'Rationality' of Interest Rate Expectations." *Journal of Monetary Economics* 6 (October 1980): 453-465.

Froot, Kenneth A. "New Hope for the Expectations Hypothesis of the Term Structure of Interest Rates." *The Journal of Finance* 44 (June 1989): 283-305.

Gavin, William T., and Nicholas V. Karamouzis. "Monetary Evidence and Real Interest Rates: New Evidence from the Money Stock Announcements." Working Paper 1984-6, Federal Reserve Bank of Cleveland, December 1984.

Goodfriend, Marvin. "Interest Rates and the Conduct of Monetary Policy." Working Paper 90-6, Federal Reserve Bank of Richmond, August 1990.

Hafer, R. W., and Scott E. Hein. "Comparing Futures and Survey Forecasts of Near-Term Treasury Bill Rates." Federal Reserve Bank of St. Louis *Economic Review* 71 (May/June 1989): 33-42.

Hamburger, Michael J., and Elliott N. Platt. "The Expectation Hypothesis and the Efficiency of the Treasury Bill Market." *The Review of Economics and Statistics* LVII (May 1975): 190-199.

Hansen, Lars P. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 (July 1982): 1029-1054.

Hardouvelis, Gikas A. "Market Perceptions of Federal Reserve Policy and the Weekly Monetary Announcements." *Journal of Monetary Economics* 14 (September 1984): 225-240.

————. "The Predictive Power of the Term Structure during Recent Monetary Regimes." *The Journal of Finance* 43 (June 1988): 339-356.

Hegde, Shantaram P., and Bill McDonald. "On the Informational Role of Treasury Bill Futures." *The Journal of Futures Markets* 6 (Spring 1986): 629-643.

Hendershott, Patric H. "Expectations, Surprises and Treasury Bill Rates: 1960-82." *Journal of Finance* 39 (July 1984): 685-696.

Heuson, Andrea J. "The Term Premia Relationship Implicit in the Term Structure of Treasury Bills." *The Journal of Financial Research* XI (Spring 1988): 13-20.

Hicks, J. R. *Value and Capital*, second edition. London: Oxford University Press, 1946.

Husted, Steven, and John Kitchen. "Some Evidence on the International Transmission of the U.S. Money Supply Announcement Effects." *Journal of Money, Credit and Banking* 17 (November 1985, Part I): 456-466.

Kane, Edward J. "Nested Tests of Alternative Term-Structure Theories." *Review of Economics and Statistics* 65 (February 1983): 115-123.

Kessel, Reuben A. *The Cyclical Behavior of the Term Structure of Interest Rates*. National Bureau of Economic Research Occasional Paper 91, 1965.

Laurent, Robert D. "Testing the Spread." Federal Reserve Bank of Chicago *Economic Perspectives* (July/August 1989): 22-33.

Lauterbach, Beni. "Consumption Volatility, Production Volatility, Spot-Rate Volatility, and the Returns on Treasury Bills and Bonds." *Journal of Financial Economics* 24 (September 1989): 155-179.

Malkiel, Burton G. "The Term Structure of Interest Rates: Theory, Empirical Evidence, and Applications." Morristown, New Jersey: General Learning Press, 1970.

Mankiw, N. Gregory, and Jeffrey A. Miron. "The Changing Behavior of the Term Structure of Interest Rates." *The Quarterly Journal of Economics* CI (May 1986): 211-228.

Mankiw, N. Gregory, and Lawrence H. Summers. "Do Long-Term Interest Rates Overreact to Short-Term Interest Rates?" *Brookings Papers on Economic Activity* (1:1984): 223-242.

McCarthy, F. Ward. "Basics of Fed Watching" in Frank J. Fabozzi (ed.) *The Handbook of Treasury Securities*. Chicago: Probus Publishing Company, 1987.

McCulloch, J. Huston. "The Monotonicity of the Term Premium: A Closer Look." *Journal of Financial Economics* 18 (March 1987): 185-192.

————. "U.S. Term Structure Data." Appendix II in Robert J. Shiller "The Term Structure of Interest Rates." Working Paper No. 2341. National Bureau of Economic Research, August 1987.

Mishkin, Frederic S. "The Information in the Term Structure: Some Further Results." *Journal of Applied Econometrics* 3 (October-December 1988): 307-314.

Nelson, Charles R., and Andrew F. Siegel. "Parsimonious Modeling of Yield Curves." *Journal of Business* 60 (October 1987): 473-489.

Newey, Whitney K., and Kenneth D. West. "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (May 1987): 703-708.

Park, Sang Yong, and Marc R. Reinganum. "The Puzzling Price Behavior of Treasury Bills that Mature at the Turn of Calendar Months." *Journal of Financial Economics* 16 (June 1986): 267-283.

Pesando, James E. "Determinants of Term Premiums in the Market for United States Treasury Bills." *The Journal of Finance* 30 (December 1975): 1317-1327.

Prell, Michael J. "How Well do the Experts Forecast Interest Rates." Federal Reserve Bank of Kansas City *Monthly Review* (September-October 1973): 3-13.

Roley, V. Vance, and Carl E. Walsh. "Monetary Policy Regimes, Expected Inflation, and the Response of Interest Rates to Money Announcements." *The Quarterly Journal of Economics* 100 (Supplement 1985): 1011-1039.

Romer, Christina, and David Romer. "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." Working Paper No. 2966. National Bureau of Economic Research, May 1989.

Rowe, Timothy D., Thomas A. Lawler, and Timothy Cook. "Treasury Bill Versus Private Money Market Yield Curves." Federal Reserve Bank of Richmond *Economic Review* 72 (July/August 1986): 3-12.

Sargent, Thomas J. *Dynamic Macroeconomic Theory*. Cambridge, Massachusetts: Harvard University Press, 1987.

Shiller, Robert J. "The Term Structure of Interest Rates." Working Paper No. 2341. National Bureau of Economic Research, August 1987.

Shiller, Robert J., John Y. Campbell, and Kermit L. Schoenholtz. "Forward Rates and Future Policy: Interpreting the Term Structure of Interest Rates." *Brookings Papers on Economic Activity* (1:1983): 173-217.

Simon, David P. "Expectations and the Treasury Bill-Federal Funds Rate Spread over Recent Monetary Policy Regimes." *The Journal of Finance* 45 (June 1990): 567-577.

Startz, Richard. "Do Forecast Errors or Term Premia Really Make the Difference Between Long and Short Rates?" *Journal of Financial Economics* 10 (November 1982): 323-329.

Stock, James H., and Mark W. Watson. "Business Cycle Properties of Selected U.S. Time Series, 1959—1988." Working Paper No. 3376. National Bureau of Economic Research, June 1990.

Strongin, Steven, and Vefa Tarhan. "Money Supply Announcements and the Market's Perception of Federal Reserve Policy." *Journal of Money, Credit and Banking* 22 (May 1990): 135-153.

Toevs, Alden L., and David J. Mond. "Capturing Liquidity Premiums." Morgan Stanley Fixed Income Research, January 1988.

Throop, Adrian W., "Interest Rate Forecasts and Market Efficiency." Federal Reserve Bank of Kansas City *Economic Review* (Spring 1981): 29-43.

Van Horne, James C. *Financial Market Rates and Flows*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1984.

Webb, Roy H. "The Irrelevance of Tests for Bias in Series of Macroeconomic Forecasts." Federal Reserve Bank of Richmond *Economic Review* 73 (November/December 1987): 3-9.

The Macroeconomic Effects of Government Spending

Ching-Sheng Mao *

I. INTRODUCTION

Peacetime government spending has risen steadily from less than 10 percent of GNP in the 1920s to about 30 percent of GNP today.[1] The larger role of government has generated increasing interest in the macroeconomic effects of government spending. This paper examines the effects of government spending in a simple macromodel. A small-scale neoclassical model is used for analyzing a classical problem in the literature, namely, the effects of temporary and persistent changes in government spending under a balanced budget. It is found that under a simple lump-sum tax financing scheme, persistent changes in government spending have much larger effects on economic aggregates (such as consumption, output, labor, and investment) than do temporary changes. This result replicates the findings of recent studies by King (1989) and Aiyagari, Christiano, and Eichenbaum (1990).

The second purpose of this paper is to analyze the effects of government spending under different tax financing regimes. For simplicity or technical reasons, the above studies assume that government purchases are financed by lump-sum taxes.[2] This assumption severely limits the applicability of the theory because most taxes are distortionary. The current paper extends the existing literature to the important case of income tax financing. The results stemming from this extension are fundamentally different from those of lump-sum tax financing. For example, an increase in government spending that is financed by a lump-sum tax under a balanced budget will increase labor effort and real output because of the dominating income effect. Under income tax financing, however, both labor effort and output decline instead of rise in response to an increase in government spending.

This paper is organized as follows: Section II describes a model economy that will be used for analyzing the effects of government spending. Section III analyzes the consumer's problem. Section IV then calibrates the model economy and considers a specific example. The effects of temporary and persistent changes in government spending, under both the lump-sum tax and the income tax regime, are discussed in Section V. Section VI concludes the paper and points out possible extensions for future studies.

II. THE ECONOMY

The hypothetical economy is assumed to be populated by a large number of identical and infinitely lived consumers. Since consumers are all alike by assumption, their behavior can be represented in terms of a single representative agent. At each date t, the representative agent values services from consumption of a single commodity c_t and leisure l_t. It is assumed that both leisure and the consumption goods are normal in the sense that more is always desired to less and that the utility function $u(c_t,l_t)$ satisfies the usual restrictions, namely, it is strictly increasing, concave, and twice differentiable.

The consumer derives his income from three different sources. First, at time t the consumer provides labor services n_t (hours of work) to the market and earns wage income w_tn_t, where w_t is the market-determined hourly real wage rate expressed in consumption units. Labor hours are constrained by the total time endowment, which is normalized to one. Thus, $l_t + n_t = 1$. The second source of income derives from the holding of a single asset called capital. At the beginning of each period, the consumer rents to firms the amount of capital k_t carried from the previous period and collects capital income r_tk_t, where r_t is the market-determined rental rate expressed in consumption units. In each period, the government imposes a uniform tax rate τ_t on wage income and capital income so that the net-of-tax earned income for the consumer is $(1-\tau_t)(w_tn_t + r_tk_t)$.[3] The final source of income is the lump-sum

* I would like to thank Tim Cook, Michael Dotsey, Marvin Goodfriend, Robert Hetzel, and Tom Humphrey for helpful comments. All errors are my own.

[1] For a statistical review of government spending, see Barro (1984).

[2] A notable exception is Baxter and King (1990) who considered the case of a proportional tax. Barro (1984) also discussed the implications of income tax financing.

[3] For simplicity, wage income and capital income are assumed to be taxed at the same rate. This assumption may not represent the actual tax scheme in the U.S. where capital income (e.g., capital gains) is usually taxed at a lower rate than is wage income.

transfer v_t from the government. Depending upon the budget constraint of the government, the lump-sum payment may be negative, in which case there is a lump-sum tax imposed on the consumer. The total disposable income for the consumer at time t is therefore $(1-\tau_t)w_tn_t + (1-\tau_t)r_tk_t + v_t$, which will be allocated between consumption and investment. In short, the budget constraint for the consumer at time t is:

$$c_t + i_t = (1-\tau_t)w_tn_t + (1-\tau_t)r_tk_t + v_t, \qquad (1)$$

where $i_t = k_{t+1} - (1-\delta)k_t$ is gross investment[4] and δ is the depreciation rate of capital $(0 < \delta < 1)$. While the capital stock will always be positive, gross investment is allowed to become negative. That is, investment is reversible in the sense that the consumer may actually eat some existing capital stock if he decides to do so.[5]

The consumer's problem is to choose a sequence of contingent plans for consumption and labor supply, taking prices as given, so as to maximize his discounted expected lifetime utility subject to the budget constraint. Formally, the consumer solves the following maximization problem:

$$\max E_0 \left[\sum_{t=0}^{\infty} \beta^t u(c_t, l_t) \right], \ 0 < \beta < 1,$$

$$\text{subject to} \quad c_t + i_t = (1-\tau_t)(w_tn_t + r_tk_t) + v_t,$$

$$\text{and}$$

$$l_t + n_t = 1, \quad \text{for all t,}$$

where β is the time preference discounting factor and E_0 is the conditional expectation operator. Expectations are taken conditional on the future course of government spending, which will be discussed shortly. The optimal solution of the consumer's problem will be characterized in the next section.

As in the case of consumers, there are a large number of identical firms in the economy; each firm accesses a constant returns to scale technology represented by the production function $F(k_t,n_t)$. During each period, the firm chooses inputs in order to maximize the current profit (or output) at the market-determined wage rate and rental rate. Let y_t denote output at time t; then the firm solves the following problem:

$$\max \ [y_t - w_tn_t - r_tk_t]$$

$$\text{subject to} \quad y_t = F(k_t,n_t).$$

Note that the firm's problem is much simpler than that of consumers; it does not involve any intertemporal trade-off as in the consumer's problem. Since the market is assumed to be competitive, the zero profit condition dictates that capital and labor will be employed up to the point where the rental rate r_t and the wage rate w_t equal the marginal product of capital and labor, respectively. That is:

$$w_t = F_n(k_t,n_t) \text{ and } r_t = F_k(k_t,n_t) \qquad (2)$$

where F_n and F_k are the marginal product of labor and capital, respectively.[6] To focus on government fiscal shocks, it has been assumed that there is no uncertainty in the firm's production process. Incorporating such uncertainty into the model is easy, but unnecessary. Also, for simplicity, it is assumed that the firm's income or profit is not taxed.[7]

The role of the government in this hypothetical economy is a simple one. It collects taxes and consumes portions of real output each period. It is assumed that government spending is not utility- or production-enhancing; the resources claimed by the government are simply "thrown into the ocean" and vanish. This assumption may not be the most interesting way to model the function of the government, but it serves as a useful point of departure. Thus, let g_t be the percentage of output that the government claims each period. Then government purchases at time t are g_ty_t. In order to finance its purchases, the government collects taxes $\tau_t(w_tn_t + r_tk_t)$, which are equal to τ_ty_t in view of the constant returns to scale technology. As noted before, the variable τ_t is the income tax rate. The budget constraint of the government at time t, expressed in per capita terms, is:

$$g_ty_t + v_t = \tau_ty_t. \qquad (3)$$

In short, equation (3) states that the sum of government purchases g_ty_t and lump-sum transfers v_t must equal tax revenues τ_ty_t. I rule out the possibility of debt financing and money creation as alternative means to finance government purchases. That is, the

[4] The gross investment i_t is the sum of the net investment $(k_{t+1} - k_t)$ and the replacement investment δk_t.

[5] Later on, the shock I choose turns out to generate negative gross investment at the time of impact, but not later.

[6] Throughout the paper, the notation $F_n(.)$ will be used to denote the partial derivative of the function F with respect to the argument n_t, which is the marginal product of labor. Similar quantities are defined accordingly.

[7] It should be mentioned, however, that the personal income tax in the hypothetical economy is equivalent to a production tax.

government fiscal policy will be conducted under a balanced budget constraint.

The variable g_t is an exogenous policy instrument that is assumed to be a random variable. Ideally, the government would make g_t contingent on certain variables in the economy such as output and labor hours. However, a simplistic view will be taken regarding the policy process $\{g_t\}$. Specifically, g_t is assumed to follow a first-order Markov process with a given transition probability that is known to all agents in the economy. For the bulk of the analysis, the transition probability will be further structured so that it gives rise to the following autoregressive representation:

$$E[g_{t+1}|g_t] = (1-\rho)g^* + \rho g_t, \qquad 0 \leq \rho < 1. \qquad (4)$$

In this representation, the conditional mean of g_{t+1} depends only on its immediate past plus a constant term $(1-\rho)g^*$. The quantity g^* is the steady state or long-run level of the government share of GNP. The autoregressive parameter ρ, assumed to be nonnegative and less than one, will determine the persistence of government spending. The larger ρ is, the more lasting will be the displacement of g_t. If $\rho = 0$, then changes in government spending will be completely temporary.

Although the government is not allowed to print money or issue debts to finance its purchases, it still has some latitude in choosing different tax schemes. Two idealized tax systems will be considered in this paper: (1) $\tau_t = 0$ and (2) $\tau_t = g_t$. In the first case, the government finances all its purchases by a lump-sum tax. That is, the transfer v_t is negative and equals $g_t y_t$ in absolute value. In the second case, all government purchases are financed by an income tax and the lump-sum transfer will be zero (i.e., $v_t = 0$). This policy exerts the greatest distortion on the behavior of consumers.

It is not difficult to conceive that the effects of government spending will depend upon the way it is financed. For instance, if the spending is financed by an income tax, there will be substitution effects that will distort market outcomes. Even in the case of a lump-sum tax, market prices will still have to adjust in response to changes in quantities that are induced by income effects. It is impossible to assess the impact of government spending without explicitly considering the market equilibrium.

III. THE EQUILIBRIUM

The equilibrium of the model economy requires that the commodity market clear at each date and that consumers and firms solve their maximization problems at the given market prices. A formal definition of the equilibrium is discussed in the appendix. Here we focus on characterizing the firm's and consumer's equilibrium.

As noted before, the firm's problem is straightforward. It requires, as stated in equation (2), that the rental rate and the real wage rate equal the marginal product of capital and labor, respectively. The consumer's problem requires that the following two first-order necessary conditions be satisfied in equilibrium:

$$u_l(c_t,l_t)/u_c(c_t,l_t) = (1-\tau_t)w_t. \qquad (5)$$

$$u_c(c_t,l_t) = \beta\, E_t\big[u_c(c_{t+1},l_{t+1})\,[1 + (1-\tau_{t+1})r_{t+1} - \delta]\big]. \qquad (6)$$

Equation (5) states that the rate of substitution of consumption for leisure (i.e., the ratio of their marginal utilities) should equal the opportunity cost of leisure, which is the after-tax wage rate. Equation (6) states that the utility-denominated price of current consumption (i.e., marginal utility of consumption) should equal the discounted expected return on saving, which is the expected value of the product of the after-tax return to investment $[1 + (1-\tau_{t+1})r_{t+1}-\delta]$ and the next period's marginal utility of consumption discounted at the rate β.[8] This condition implies that in equilibrium the consumer is indifferent between consuming one extra unit of output today and investing it in the form of capital and consuming tomorrow. Equations (5) and (6) together with the budget constraint (1) and the time constraint $l_t + n_t = 1$ completely characterize the consumer's equilibrium.

Figure 1 presents a two-quadrant diagram to illustrate the determination of the consumer's equilibrium. For this purpose, we assume that there is no uncertainty in the economy and that the utility function is homothetic.[9] The right-hand quadrant depicts the

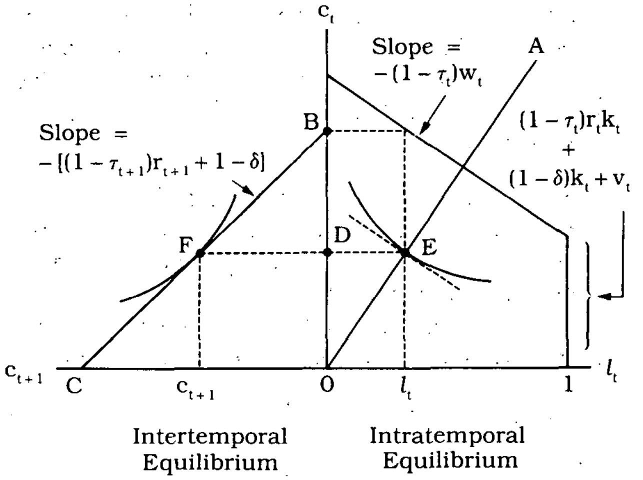
[8] Note that in a deterministic context, the gross return to investment will be equal to one plus the real interest rate, which is the ratio of marginal utilities of consumption between time t and time t+1.

[9] A utility function is called homothetic if the marginal rate of substitution depends only on the consumption-leisure ratio. A homothetic utility function has the property that the slope of the indifference curve is constant along a given ray from the origin.

Figure 1

CONSUMER'S EQUILIBRIUM

Taking the real interest rate, the after-tax wage rate, and the after-tax rental rate as given


c_t
Slope = $-(1-\tau_t)w_t$
A
Slope = $-[(1-\tau_{t+1})r_{t+1}+1-\delta]$
B
$(1-\tau_t)r_tk_t$ + $(1-\delta)k_t+v_t$
D
E
F
c_{t+1}
C
c_{t+1}
0
l_t
1
l_t
Intertemporal Equilibrium
Intratemporal Equilibrium

trade-off between consumption (measured along the vertical axis) and leisure (measured along the horizontal axis) for a given wage rate and tax rate at time t. The budget line in the right-hand quadrant for the consumer at time t has two components: the vertical segment corresponds to the nonwage income which is fixed at the beginning of the period and equals $[1+(1-\tau_t)r_t-\delta]k_t + v_t$; the sloping segment corresponds to labor income $(1-\tau_t)w_tn_t$ and has the slope $-(1-\tau_t)w_t$. From equation (5), the slope of the indifference curve must equal the after-tax wage rate in equilibrium. Since the utility function is homothetic, this condition determines an equilibrium consumption-leisure ratio that is represented by the ray OA extended from the origin.

Equation (5) alone cannot pin down the equilibrium point, however. To locate the equilibrium, one must determine saving from equation (6). Consider the point E along the ray OA. There is an indifference curve tangent at E with slope equal to $-(1-\tau_t)w_t$. The total income associated with this point, OB, is divided between consumption and investment. If invested, the income available at time t+1 is OC, which is measured from right to left along the horizontal axis in quadrant 2. The absolute value of the slope of the budget line BC is the after-tax rate of return to capital [i.e., $1 + (1-\tau_{t+1})r_{t+1}-\delta$]. According to equation (6), the intertemporal equilibrium will be achieved at point F, where the indifference curve is tangent to the budget line BC. The point F determines the optimal saving (i.e., k_{t+1}) BD and time t consumption OD which coincide with those implied by the intratemporal equilibrium point E.[10] Points E and F jointly characterize the consumer's equilibrium. Other quantities such as leisure (labor hours) and time t+1 consumption can be easily derived once the equilibrium point is determined.

The appendix sketches a numerical procedure which permits computation of the equilibrium and quantitative assessment of the effects of government spending. This approach requires one to take an explicit stand on the parameter structure of the economy. The rest of the paper therefore focuses on a specific example and works out the equilibrium implications of changes in government spending.

IV. CALIBRATING THE MODEL

The example considered here involves a logarithmic utility function:

$$u(c_t, l_t) = \theta \ln c_t + (1-\theta) \ln l_t, \quad 0 < \theta < 1,$$

and a Cobb-Douglas production function:

$$F(k_t, n_t) = k_t^{\alpha} n_t^{1-\alpha}, \quad 0 < \alpha < 1.$$

This specification is widely used in the literature because it generates dynamics that roughly match several important features of business cycles in the U.S. (see, for example, King, Plosser, and Rebelo (1988)). Our experiment assumes the following values: $\alpha = 0.3$, $\theta = 0.3$, $\beta = 0.96$, and $\delta = 0.05$.

In addition to preferences and technology, one needs explicitly to spell out the process of government spending. As mentioned before, the variable g_t, i.e., the ratio of government spending to real output, is assumed to follow a first-order Markov process. In what follows, the autoregressive parameter ρ is assigned either a value of 0 in the case of a temporary government spending or a value of 0.9 in the case of a more persistent government spending. Further, the random variable g_t is assumed to possess a binomial distribution with probabilities concentrated on five distinct points over a bounded interval. The mean and variance of g_t are taken to be 0.3 and 0.005, respectively. These figures imply that g_t will fluctuate around 30 percent (i.e., $g^* = 0.3$), ranging approximately from 15 percent to 45

[10] If the intratemporal equilibrium and the intertemporal equilibrium do not imply the same consumption and saving decisions, then another point along the ray OA must be chosen until the two equilibria are consistent.

percent. Given this specification, the transition probability of g_t, needed to numerically solve the model, is constructed using the method proposed by Rebelo and Rouwenhorst (1989).

V.
Dynamic Effects of Government Spending

Consider the following scenario: Suppose, initially, that the economy has settled at its steady state equilibrium, and that government spending has reached its long-run level relative to the economy's real output such that 30 percent of real output is claimed by the government. At date 1 the government raises taxes and increases spending. Thereafter, the ratio of government spending to real output follows a time path prescribed by the autoregressive process and gradually returns to its steady state. The left- and right-hand sides of Figure 2 plot the mean path of g_t, measured as percentage deviations from the steady state, for $\rho = 0$ and $\rho = 0.9$, respectively. These hypothetical paths are generated by taking an average of 5000 random realizations of g_t, conditional on the given change at the initial date. Notice that the case of $\rho = 0.9$ yields a more lasting displacement of g_t.

Given the displacement of government spending, what would be the dynamic response of quantities and prices in the pure lump-sum versus pure income tax regime? To answer this question, one needs to understand the forces that govern individual behavior. It is instructive to consider a simpler case in which the increase in government spending is financed by a lump-sum tax. Figure 3 shows the shift in the consumer's equilibrium for this case. As in Figure 1, the points E and F represent the initial equilibrium prior to the occurrence of shocks to government spending. As government spending rises, the budget line shifts downward by an amount equal to the increment of government spending, i.e., $-\Delta v_t = \Delta(g_t y_t)$. With lump-sum tax financing, the slope of the budget line or the after-tax wage rate remains unchanged. As a result, the new equilibrium will still lie on the rays OA and OB (recall that the utility function is homothetic). Given the new budget constraint, the intratemporal and intertemporal equilibrium will be achieved at point E′ and F′, respectively. Since there is only an adverse income effect, represented by the downward and parallel shift of the budget line, the new equilibrium displays less consumption for both periods and greater work effort. The individual is willing to work harder because leisure is a normal good and the individual is poorer than before due to tax increases. Because both income and consumption are lower, the effect on saving is indeterminate. In other words, at the initial interest rate, saving or investment may rise or fall, so it appears that the equilibrium interest rate may go either way. In the simulation below, however, we will see that it rises.

How might results differ with income tax financing? Now, substitution effects of changes in the after-tax wage rate and rental rate become potentially important. A change in the income tax rate will induce not only a substitution between consumption and leisure at a given date, but also a substitution of consumption over time. In order to assess the impact of government spending, it is necessary to trace out the equilibrium paths of quantities and prices.

The dynamic responses of the system are displayed below the dotted line in Figure 2. These response functions are calculated by taking an average from 5000 random realizations of the system, conditional on the initial displacement of government spending. To contrast the effects under different tax regimes, each figure contains two transition paths of the same variable; the solid line traces out the dynamic response under lump-sum tax financing; the dotted line traces out the dynamic response under income tax financing. Since the steady state is different for the two tax regimes, these responses are expressed in terms of percentage deviations from the steady state. The following discusses the different implications under the two tax financing schemes.

Lump-Sum Tax vs. Income Tax Financing (Temporary Case)

Consider first the case of a temporary increase in government spending in which g_t jumps from 30 percent to above 40 percent at date 1. Since the shock is temporary, it lasts for about one period (see Figure 2, left-hand side). As the left-hand side of Figure 2 shows, both lump-sum tax financing and income tax financing have negative effects on capital, consumption, and investment. The magnitudes are quite different, however. In the case of lump-sum tax financing, capital falls by 3 percent on impact, while consumption and investment decrease by 2 percent and 70 percent, respectively. The negative effects are much more severe under income tax financing; capital falls by over 9 percent while consumption and investment drop by more than 5 percent and 180 percent, respectively. Two reasons are responsible for these results. First, a rise in the income tax rate decreases the after-tax marginal product of capital. In addition, a decrease in labor

Figure 2

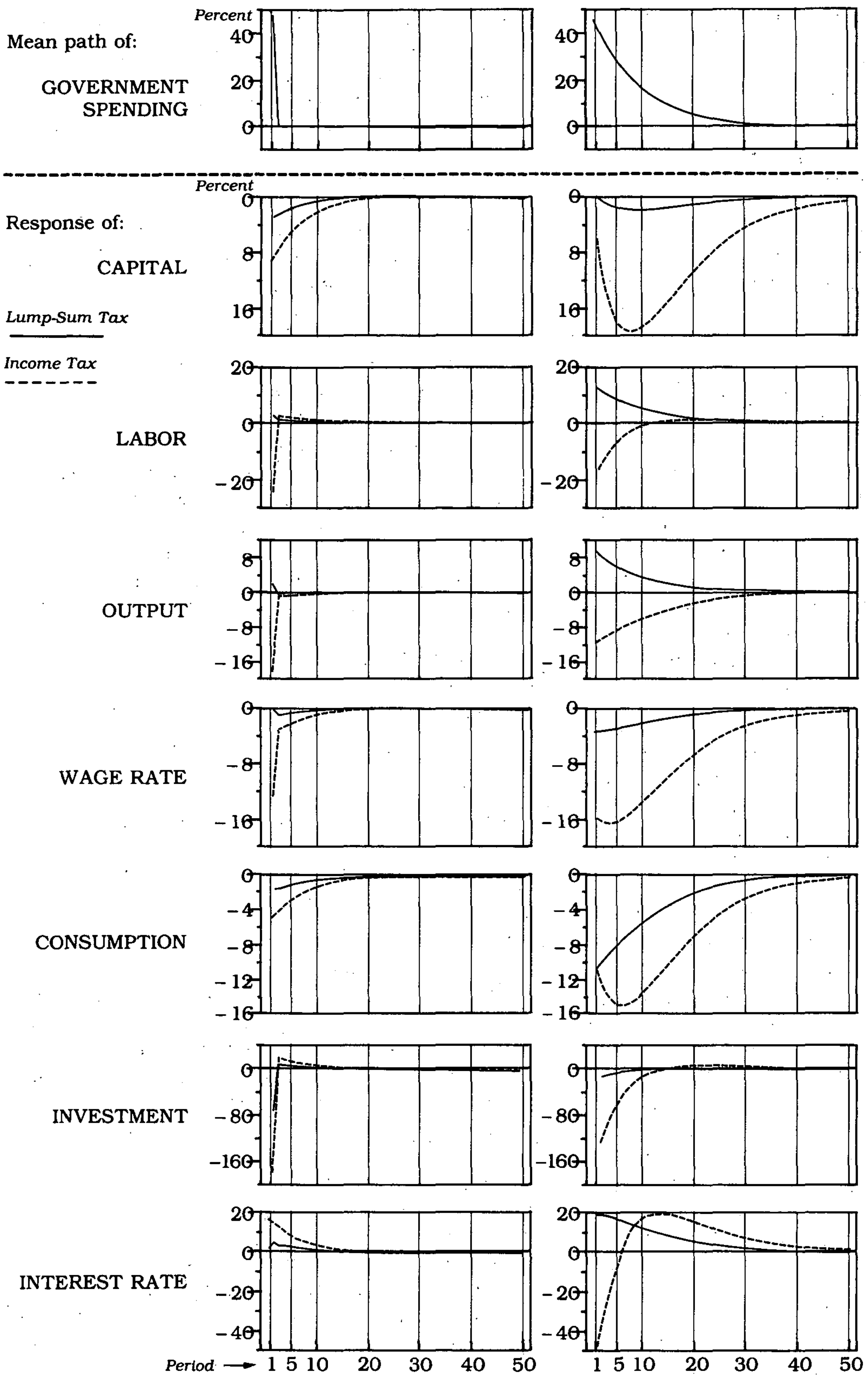
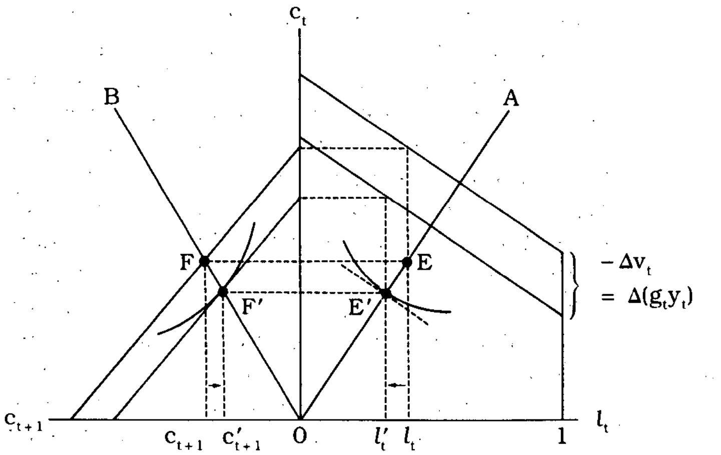

rho = 0
rho = 0.9
Mean path of:
GOVERNMENT SPENDING
Percent
40
20
0
Response of:
CAPITAL
Lump-Sum Tax
Income Tax
0
8
16
LABOR
20
0
−20
OUTPUT
8
0
−8
−16
WAGE RATE
0
−8
−16
CONSUMPTION
0
−4
−8
−12
−16
INVESTMENT
0
−80
−160
INTEREST RATE
20
0
−20
−40
Period
1 5 10 20 30 40 50

Figure 3

CONSUMER'S EQUILIBRIUM: EFFECTS OF AN INCREASE IN LUMP-SUM TAX



hours, which results also from a lower after-tax wage rate, pushes the marginal product of capital even lower.[11] As the productivity of capital falls, agents have less incentive to accumulate capital so that the decrease in investment is larger under income tax financing. Finally, the lower productivity of capital and labor represents an additional loss of income which makes agents poorer than in the lump-sum tax case. Therefore, the decrease in consumption is also larger under income tax financing. In order to induce agents to consume less, the real interest rate will go up to maintain equilibrium in the goods market.

The most visible difference between lump-sum tax financing and income tax financing shows up in their effects on labor effort and real output. In the case of lump-sum tax financing, both labor effort and real output rise on impact by about 2 percent, while income tax financing causes them to decrease by more than 24 percent and 17 percent, respectively. Three forces determine the response of labor supply. First, an increase in government spending leads to the use of real resources and makes agents poorer. This adverse income effect motivates consumers to work harder. However, since the disturbances are temporary, this effect is relatively small. Second, there is a wage effect. As Figure 2 shows, the after-tax wage rate falls by more than 13 percent in the income tax case, as opposed to a tiny 0.6 percent drop in the lump-sum tax case. The larger decrease in the wage rate tends to dampen the response of labor supply. The lower wage rate implies that leisure is less expensive relative to consumption and as a result, consumers are more willing to take leisure instead of consumption. Finally, there is an interest rate effect. According to Figure 2, the real interest rate rises on impact, which largely reflects the increase of aggregate demand associated with an increase in government spending. The rise in the interest rate encourages consumers to work harder due to a higher rate of return. Under lump-sum tax financing, the wage effect is dominated by the income effect and the interest rate effect, resulting in greater labor effort. Since the capital stock is fixed at the beginning of the period, output also increases. Although the interest rate rises even higher in the case of income tax financing, this rise together with the income effect is not sufficient to outweigh the wage effect so that both labor hours and real output decrease.

The initial response of the interest rate and output can be analyzed using the traditional aggregate demand and aggregate supply paradigm. Figure 4a depicts the equilibrium shift in the goods market when a lump-sum tax is used to finance government spending. The real interest rate and output are measured on the vertical and horizontal axis, respectively. The point E is the initial equilibrium point. As government spending rises, the aggregate demand schedule shifts to the right because of the increase in goods demanded by the government. The aggregate supply schedule also shifts to the right because, as explained above, labor supply increases. However, since the increase in government spending is temporary, the shift in aggregate supply will be relatively small due to the negligible income effect. As a result, there is an excess demand at the initial interest rate r*, which must rise in order to restore equilibrium in the goods market. As the real interest rate rises, aggregate supply (labor effort) increases while aggregate demand (consumption and investment) decreases and the new equilibrium is reached at point F. Comparing points E and F reveals that both output and the real interest rate are higher.

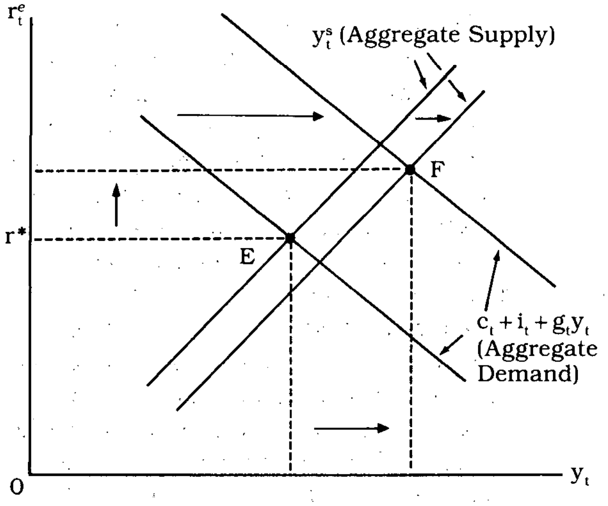
The case of income tax financing can be analyzed in a similar fashion (see Figure 4b). The principal difference here is that the aggregate supply schedule will now shift to the left because of the decrease in labor supply. The shift in aggregate supply will of course depend on the extent to which the marginal

[11] Since capital and labor are complements in production, a decrease in labor input lowers the productivity of capital.

EFFECTS OF TEMPORARY INCREASES IN GOVERNMENT SPENDING

Figure 4a

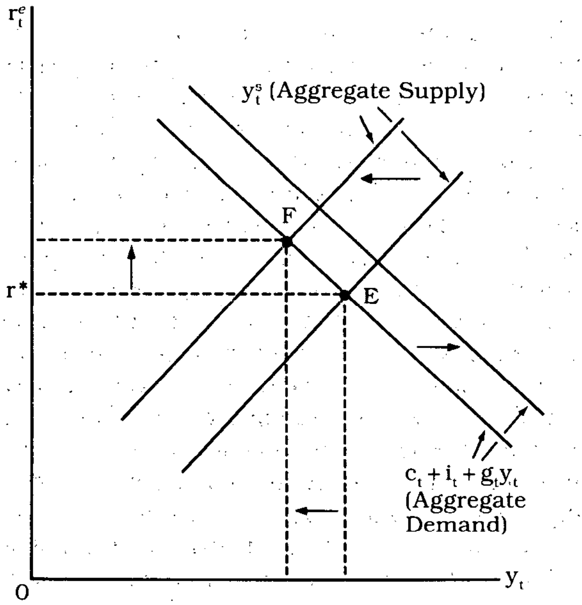
With Lump-Sum Tax



r^e_t
y^s_t (Aggregate Supply)
F
r^*
E
$c_t + i_t + g_t y_t$ (Aggregate Demand)
0
y_t

Figure 4b

With Income Tax



r^e_t
y^s_t (Aggregate Supply)
F
r^*
E
$c_t + i_t + g_t y_t$ (Aggregate Demand)
0
y_t

product of labor is reduced. It turns out that in the case under consideration the shift of aggregate supply outweighs that of aggregate demand so that output decreases while the interest rate rises.

The analysis up to this point has focused on the short-run effects of an increase in government spending. Consider now the transition dynamics of the system after the initial impact. Since the capital stock is lower at date 1, the marginal product of capital increases.[12] As a result, agents begin to accumulate more capital after date 1. As the capital stock or investment increases, the real interest rate (or the marginal product of capital) falls and consumption begins to rise. Consumption rises over the transition period because current consumption becomes less expensive relative to future consumption as the interest rate declines over time.[13] This response applies to the lump-sum tax financing as well as the income tax financing. Figure 2 shows, however, that the transition path of real output and labor effort will depend on the tax regimes. In the lump-sum tax case, both labor hours and real output decrease over time because the real interest rate falls (recall that a lower interest rate implies a lower labor effort). In the case of income tax, the rising wage rate, due to a decrease in the income tax rate and an increase in the capital stock, becomes an overriding force that pushes labor hours up over the transition period. As can be seen from the figure, labor supply will temporarily overshoot the steady state and then decline to the initial equilibrium. As labor supply and the capital stock rise, output also increases until the steady state is reached.

Lump-Sum Tax vs. Income Tax Financing (Persistent Case)

Suppose now that the increase in government spending is more persistent (i.e., $\rho = 0.9$). The right panel shows that the responses are very similar to those of a temporary increase in government spending. The principal difference is the implied wealth effect. Because the shock is expected to persist for a longer period of time, the wealth effect will now play a more important role in the response of quantities and prices.

Consider the case of lump-sum tax financing. Figure 2 shows that labor hours rise by 13 percent and consumption falls by 10 percent on impact. These responses are more than five times the responses in the temporary case. These results occur because consumers are poorer than in the case of a temporary shock. To induce agents to consume less and work harder, the real interest rate will also

[12] Since labor hours rise under lump-sum tax financing, it pushes the marginal product of capital even higher. Under income tax financing, labor effort decreases, but the decrease outweighs that of capital (see Figure 2, left-hand side) and the capital-labor ratio is lower at date 1, implying a higher marginal product of capital.

[13] The negative correlation between current consumption and the real interest rate is sometimes called the effect of intertemporal substitution.

increase by a larger magnitude. Again, since capital is predetermined, real output rises with labor supply. Perhaps the most interesting difference here is that investment does not go down as much as in the temporary case. The principal reason for this result is that the increase in labor hours occurs over a more extended time period and pushes up the marginal product of capital both now and in the future, thus raising the rate of return to investment. It should be noted, however, that investment will still go down on impact as consumers try to smooth out consumption by holding less capital.

The adverse income effect works in a similar fashion under the income tax regime. In particular, consumption drops by more than 10 percent, as opposed to a 5 percent decrease in the temporary case. Because of the income effect, the decrease in labor hours, which is caused by a lower after-tax wage rate, is smaller than that in the temporary case. Consequently, the decrease of real output is also smaller. Because the decrease of labor effort is smaller, the marginal product of capital does not go down as much as in the temporary case, leading to a smaller decrease in investment.

Although the initial effects of a persistent increase in the income tax rate are not as large as those in the temporary case (except consumption), major variables such as output and investment will stay below their steady state for a long period of time. In fact, the shock is so persistent that agents will eat up some existing capital for one period before consumption (and capital) begins to rise over the transition period. This is the case of a severe recession. The reason for this result is that the marginal product of capital is so low in the future that agents have very little incentive to accumulate capital.

A surprising feature of the income tax regime is that the real interest rate declines in response to a persistent increase in government spending. Again, this result can be attributed to the income effect. As noted before, output supply will decline, but the decrease will not be as much as that in the temporary case because the income effect motivates agents to work harder. On the demand side, the income effect and the lower productivity of capital in the future decrease both consumption and investment at the initial real interest rate. The decrease of consumption and investment may reach the point at which it outweighs the increase of government purchases, leading to a decrease of aggregate demand. The extent to which aggregate demand decreases will depend on how long the shock persists. It turns out that in the case under consideration, the decrease in aggregate demand is quite sizable so that at the initial real interest rate there is an excess supply, resulting in a lower interest rate. Clearly, this argument hinges on the persistence of the shock and the intensity of the income effect. If the government spending shock is less persistent, then the interest rate will decline by a smaller amount or even increase as in the pure temporary case.

VI.
Conclusions and Extensions

This paper examines the balanced budget effects of government spending under different tax financing schemes. The results suggest that, in the case of lump-sum tax financing, persistent changes in government spending have larger effects on prices and quantities except investment. This result, due to larger income effect and interest rate effect, is consistent with the findings of King (1989) and others. In general, an increase in government spending under lump-sum tax financing will reduce consumption and investment but raise labor effort and real output. This result is driven by the income and interest rate effects that encourage individuals to work harder. Under income tax financing, however, some of the above results are reversed. In particular, regardless of the persistence of spending shocks, both output and labor effort now decline in response to an increase in government spending. This result occurs because the decline in the wage rate dominates the income and interest rate effects.

There are several features of the model that are oversimplified and can be improved upon. Most notably, the government budget is assumed to be balanced in each period. This assumption prevents one from seriously considering the implications of deficit or debt financing. It is relatively easy to introduce such a financing scheme into the model. Extension along this line will probably yield fruitful results if government debts coexist with some types of distortionary tax such as the income tax considered in this paper. The most important implication of debt financing is that it allows the tax burden to be smoothed out over time. This mechanism reduces the distortionary effect on labor supply, particularly when the increase in government spending is temporary. In this case, real output and labor hours may no longer decline as in the case of a balanced budget.

Another extension worth undertaking concerns the function of government spending. The current paper assumes that government spending is a waste of resources and is not utility- or production-enhancing. This assumption is inappropriate for some types of government spending that may either substitute for private consumption or increase the economy's productivity. These features could be introduced into the model by specifying a more general utility function or production function, such as those employed by Barro (1984). Such refinements would nullify or even reverse some of the negative effects associated with income tax financing.

APPENDIX

This appendix presents a definition of the equilibrium discussed in the text and outlines a numerical method to construct the equilibrium. Formally, the general equilibrium for the model economy consists of a sequence of quantities $\{c_t,k_{t+1},n_t,l_t\}$ and prices $\{w_t,r_t\}$ that satisfy the following two conditions: (1) the sequence $\{c_t,k_{t+1},n_t,l_t\}$ solves the maximization problems of consumers and firms for a given sequence of prices $\{w_t,r_t\}$ and (2) the commodity market clears at each date t such that aggregate demand equals aggregate supply:

$$c_t + i_t + g_t y_t = y_t. \qquad \text{(A1)}$$

Equation (A1) states that the total of consumption, investment, and government purchases must exhaust total output. The government budget constraint, which must also be satisfied in equilibrium, is implied by the market-clearing condition (A1) and the individual budget constraint (1) in the text.

To further characterize the equilibrium one must solve the maximization problems of consumers and firms. The firm's problem is straightforward. It requires, as stated in equation (2), that the rental rate and the wage rate be equal to the marginal product of capital and labor, respectively. This condition defines the equilibrium prices that will clear the labor market and the rental market for the existing capital stock. As discussed in the text, the consumer's equilibrium is characterized by the budget constraint (1) and the time constraint $l_t + n_t = 1$ together with two first-order necessary conditions, which are rewritten as follows:

$$u_l(c_t,l_t)/u_c(c_t,l_t) = (1-\tau_t)w_t. \qquad \text{(A2)}$$

$$u_c(c_t,l_t) = \beta\, E_t\big[u_c(c_{t+1},l_{t+1}) \\ [1 + (1-\tau_{t+1})r_{t+1} - \delta]\big]. \qquad \text{(A3)}$$

The meaning of (A2) and (A3) is discussed in the text.

The approach used to determine the equilibrium of the model economy is as follows. First, substitute the time constraint and equations (1) – (3) and (A1) into (A2) and (A3) to obtain

$$\frac{u_l[(1-g_t)F(k_t,n_t)+(1-\delta)k_t-k_{t+1},1-n_t]}{u_c[(1-g_t)F(k_t,n_t)+(1-\delta)k_t-k_{t+1},1-n_t]} = (1-\tau_t)F_n(k_t,n_t), \qquad \text{(A4)}$$

and

$$u_c[(1-g_t)F(k_t,n_t)+(1-\delta)k_t-k_{t+1},1-n_t] = \\ \beta\, E_t\,\{u_c[(1-g_{t+1})F(k_{t+1},n_{t+1}) \\ + (1-\delta)k_{t+1}-k_{t+2},1-n_{t+1}] \\ \times [(1-\tau_{t+1})F_k(k_{t+1},n_{t+1}) + (1-\delta)]\}. \qquad \text{(A5)}$$

Note that equations (A4) and (A5) are alternative versions of the consumer's equilibrium with quantities and prices replaced by the market-clearing condition and the firm's marginal conditions. These two equations jointly determine the equilibrium level of capital k_{t+1} and labor n_t,[14] which can be used to determine consumption, investment, output and equilibrium prices. Note that given the beginning of period capital k_t, a decision rule for k_{t+1} is equivalent for a saving decision made at time t.

In general, an analytical solution to equations (A4) and (A5) does not exist except for a very few special cases. Numerical methods are therefore required to obtain an approximate solution. The following briefly describes an iterative procedure used to solve the model. Technical details of this method can be found in Coleman (1989) and will not be presented here. Basically, the solution to equations (A4) and (A5) comprises a pair of decision rules for capital k_{t+1} and labor n_t that can be expressed as functions of k_t and

[14] Note that k_{t+2} and n_{t+1} are "integrated out" when (A4) and (A5) are solved.

g_t (i.e., the state of the system). The numerical procedure involves approximation of these decision rules over a finite number of discrete points on the space of k_t and g_t. Starting from an arbitrary capital rule (usually, a zero function), the procedure first solves the labor rule from equation (A4) and then iterates on equation (A5) until the capital rule converges to a stationary point, that is, until capital as a function of k_t and g_t does not change over consecutive iterations. The resulting stationary function is the equilibrium solution for capital and labor.

By construction, the above procedure yields solutions that satisfy both (A4) and (A5) for all contingencies of government spending. These solutions imply three imputed or shadow prices that are consistent with the market equilibrium. Specifically, the equilibrium wage rate w_t and rental rate r_t can be computed from the firm's marginal condition (2), and the real interest rate r_t^e, by definition, is the ratio of the marginal utilities of consumption between time t and time t+1, i.e., $u_c(c_t,l_t)/[\beta E_t u_c(c_{t+1},l_{t+1})]$. In a deterministic equilibrium, the gross real interest rate r_t^e is equal to $(1-\delta)$ plus the capital rental rate r_{t+1}, as can be seen from equations (A3) and (A5). This is the price that will clear the commodity market.

References

Aiyagari, S. R., L. J. Christiano, and M. Eichenbaum. "The Output, Employment, and Interest Rate Effects of Government Consumption," Discussion Paper 25, Institute for Empirical Macroeconomics, Federal Reserve Bank of Minneapolis, March 1990.

Barro, R. J. *Macroeconomics*, 2nd edition. New York: John Wiley & Sons, 1984.

Baxter, M. and R. G. King. "The Equilibrium Approach to Fiscal Policy: Equilibrium Analysis and Review," Manuscript, University of Rochester, May 1990.

Coleman, W. J. "Equilibrium in an Economy with Capital and Taxes on Production," Manuscript, Federal Reserve Board, May 1989.

King, R. G. "Value and Capital in the Equilibrium Business Cycle Program," Working Paper, University of Rochester, February 1989.

King, R. G., C. Plosser, and S. Rebelo. "Production, Growth and Business Cycles I. The Basic Neoclassical Model," *Journal of Monetary Economics* 21, (March/May 1988).

Rebelo, S. and G. Rouwenhorst. "Linear Quadratic Approximations Versus Discrete State Space Methods: A Numerical Evaluation," Manuscript, University of Rochester, March 1989.

Why Do Estimates of Bank Scale Economies Differ?*

David B. Humphrey

I. INTRODUCTION

A number of policy issues turn on whether or not large commercial banks, merely because of their size, are more efficient than small banks. Such scale economies, where average cost declines as bank output rises, would result from spreading fixed costs over a greater volume of output. Scale economies are an important policy consideration in interstate bank branching.

Interstate branching was long prohibited on the grounds that (1) industry concentration and monopoly power would result, and (2) local areas may be less well served by giant banks having little interest in these localities, as more profitable uses for funds would likely be found elsewhere. Cost savings associated with large scale economies, however, might overcome these negatives. As well, interstate branching would allow banks to diversify their portfolios geographically, strengthening the industry. Consumer and business bank customers would likely benefit from lower prices and reduced banking risks which could follow.

In contrast, if scale economies were small, fears of concentration might outweigh any perceived benefits of expansion. It would then be more politically tenable to limit the size and geographical distribution of banks. While there still could be loan risk diversification, this benefit by itself might not justify the concentration of economic power in truly giant banking organizations.

The level of bank scale economies is an empirical question, but one where widely differing results have made it difficult to form a clear and unambiguous conclusion. Fortunately, there are now enough studies to attempt to sort out *why* past results have differed. Such a sorting out is useful in its own right and for the implications it has for policy decisions that depend on scale economies in banking. It also illustrates the benefits a detailed analysis could have for other areas of economics where empirical findings differ and can cloud proper policy formation (such as in the appropriate definition of the money supply).

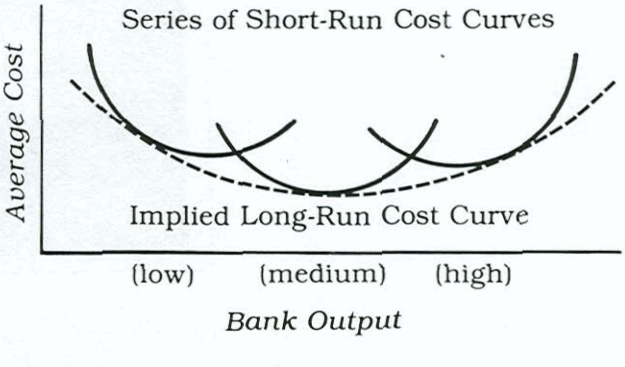
In many sciences, researchers use the same experimental technique to generate new and independent data and then look for consistency in the results. In contrast, economists generally use similar data but vary the experimental technique—that is, the particular specification and definition of variables, functional form, and time period used. Thus robust results are less frequent. If enough studies are performed, however, a pattern to the results may emerge suggesting why they differ. Then we can compare the relative advantages of different experimental techniques. Instead of a single scale economy conclusion that applies in all cases, we obtain a set of different results that illustrate how sensitive our measures are to the research design chosen. From this and from some additional thought on how we best measure scale economies, we develop a general conclusion on the size and significance of scale economies in banking.

II. COMMON DIFFERENCES AMONG STUDIES

Graphically, bank scale economies appear as the slope of an average cost curve indicating how costs vary with output. An example is shown in Figure 1. A series of short-run average cost curves (solid lines) for three different-sized banks, each producing different levels of bank output, trace out an implied long-run average cost curve (dotted line). A downward-sloping long-run average cost curve reflects scale economies. An upward slope reflects diseconomies, since higher average costs are incurred when more output is produced. The assumption is that a cross-

Figure 1

Illustrative Bank Average Cost Curves



[*] Comments by Mike Dotsey, Bob Graboyes, Tom Humphrey, and Dave Mengle are appreciated, although the opinions expressed are those of the author alone. Able research assistance was provided by Bill Whelpley.

section of different-sized banks at a point in time will reveal the appropriate long-run curve; from this is derived a measure of scale economies. Thus as smaller banks expand their output in the future, their costs are likely to "look like" the costs of larger banks today.

The cost curve itself (and the implied scale economies reflected in it) is actually derived from an equation similar to (1), below, where costs (C) are regressed on the level of bank output (Q) and other variables which affect costs but need to be held constant in the cross-section data set:

(1) $C = f(Q, \text{other variables})$.

Other variables, such as the prices of labor and capital factor inputs, need to be held constant in a cross-section in order to statistically separate movements *along* the cost curve (due to changes in output) from *shifts* in the cost curve (due to influences on bank costs which are essentially unrelated to output).

With this background, we now outline the most common differences observed in bank scale economy studies and assess how these differences have affected the results derived from them. More specifically, our purpose is to critically review the literature on bank scale economies, to select a preferred method for estimating these economies, and thereby to determine which empirical result is the most appropriate for policy purposes, as well as defensible on theoretical grounds. The most common research design differences among studies of bank scale economies concern the following:

(1) Cost definition (operating cost versus total cost);

(2) Bank output definition (numbers of accounts versus dollars in these accounts);

(3) Functional form used (linear versus quadratic);

(4) Scale economy evaluation (single office versus banking firm);

(5) Time period used (high versus low interest rate period);

(6) Commingling scale with scope (single versus multiple output); and

(7) Bank efficiency differences (assume all observations are efficient versus only those on the efficient frontier).

In the following sections, each of these differences is discussed in conjunction with one or more published studies. Some other differences occur and, when appropriate, they too are noted.

III.
Operating Versus Total Bank Costs

This section concerns how the dependent variable—cost (C)—is defined in equation (1). Many studies relate only *operating* costs to bank output levels in estimating scale economies (Langer, 1980; Nelson, 1985; Hunter and Timme, 1986; Evanoff, Israilevich, and Merris, 1989). Operating costs include wages, fringe benefits, physical capital, occupancy, and materials cost, along with management fees and data processing expenses paid to the holding company and other entities. On average, operating costs only comprise slightly over 25 percent of total costs. Most other studies have used total costs, which are obtained by adding interest expenses on purchased funds and core deposits to operating costs.[1] The two interest cost categories are large and each exceed operating costs since they comprise around 35 and 40 percent, respectively, of total costs. Clearly, it makes a difference which definition of cost is used to derive an estimate of scale economies.

The difference in cost definitions—operating versus total costs—would not be an issue if all banks had the *same* percentage composition of interest and operating expenses regardless of their size. This is because interest expenses typically have little or no economies associated with them. Therefore, adding these roughly constant cost expenses to operating costs (giving total costs) means that any scale economies or diseconomies found using operating costs alone would only be attenuated, rather than reversed, if the ratio of interest to operating costs were the same across banks. But this ratio is not even close to being stable across banks. The proportion of assets funded with purchased funds rises substantially as banks get larger so that the proportion of purchased funds interest expense in total cost rises while the proportion of core deposit interest expense and operating cost falls.

For example, at small branching banks (those with $50 to $75 million in assets in 1984), purchased funds were 12 percent of the value of core deposits plus purchased money. For medium-sized banks (with $300 to $500 million in assets), the purchased funds proportion rises to 19 percent. And for large banks (with $2 to $5 billion and then over $10 billion in

[1] Purchased funds are purchased federal funds, CDs of $100 thousand or above, and foreign deposits (which are almost always over $100 thousand). Core or produced deposits are demand deposits and small denomination (i.e., less than $100 thousand) time and savings deposits. The costs of equity and subordinated notes and debentures are small and are almost always excluded from bank cost studies.

assets), the proportion rises further to 36 and finally 60 percent. At unit state banks for the same four size groupings, the purchased funds proportions are 16, 31, 61, and 78 percent. Thus the percentage composition of interest and operating expenses varies considerably across banks and is closely related to bank size, which is the key to the problem which arises when operating costs are used.

Purchased funds have very low operating expenses per dollar raised; their only significant cost is interest expense. In contrast, core or produced deposits generate the major portion (49 percent) of all operating (capital, labor, materials) expenses. Since purchased funds are a strong substitute for core deposits, the interest expense of purchased funds is also a substitute for the operating and interest expenses of core deposits. To accurately gauge how bank costs really change with size thus requires that purchased funds and core deposit interest expenses be included with operating costs. Taken together, these components allow one to determine the average cost actually faced by a bank even as its funding mix is altered. In this way, changes in the funding mix do not bias the results.

This point is illustrated by comparing the actual *average operating cost* (operating expenses divided by total assets) for 1984 with the *average total cost* (operating plus interest expenses divided by total assets) for the same year across 13 size classes of banks (see Figure 2). The branching state bank comparison is shown in Panel A with the unit state bank comparison in Panel B.[2] Operating cost per dollar of assets is seen to fall more rapidly than total costs per dollar of assets. Thus if only operating costs are used in a statistical analysis of bank scale economies, as some investigators have done, greater scale economies (or lower diseconomies) will typically be measured when an equation like (1) is estimated and a curve is fitted to these raw data points.[3]

Hunter and Timme, 1986, obtained this result when they alternatively used operating costs and then operating plus interest costs in their statistical estimates of scale economies for 91 large bank holding companies over 11 years (1972-82). They found significant operating cost scale economies (using only operating costs) but no significant total cost scale economies (when interest expenses were included). Their study covered large banks separately and did not include any small or medium-sized institutions.

While operating costs are of some interest in themselves, it would be misleading to conclude that reductions in the ratio of operating costs to assets accurately reflects inherent bank scale economies. If this were true then a bank with a wholesale orientation (large purchased funds, small core deposits) would always experience lower costs solely because of lower operating costs per dollar of assets. But lower operating costs per dollar of assets are typically offset by having greater interest costs per dollar of assets through more intensive reliance on purchased funds instead of core deposits. Thus the proper comparison of costs, and measurement of scale economies, must rely on total costs rather than only on operating costs by themselves. When this is done, then differences in a bank's funding mix will not bias the results.[4]

IV.
Bank Output Measurement: Number of Accounts versus Dollars in the Accounts

Another important difference in published studies concerns the definition of bank output (Q), a key independent variable in equation (1). In most other industries, the measurement of output is not a problem. Output is a flow concept measured in physical terms, either because the physical unit is homogeneous and can be easily observed or because there is a convenient index of the value of the output flow which can be deflated by an appropriate output price index. In banking, neither of these alternatives exists and data availability dictates how bank output is defined. Output flow information is not available for each individual bank so information on the stock of output is used instead. Generally, researchers assume that the unobserved output flow is proportional to the observed output stock. Thus use of stock information in statistical analyses is presumed to give results similar to those obtainable using flow data.

[2] The top line in each comparison is the mean average total cost curve (solid line). To make this comparison clearer, the scale for average operating costs—right side of the figure—has been shifted up so that the two curves will appear to start from the same point for the first size class. The scale for average total costs is on the left side.

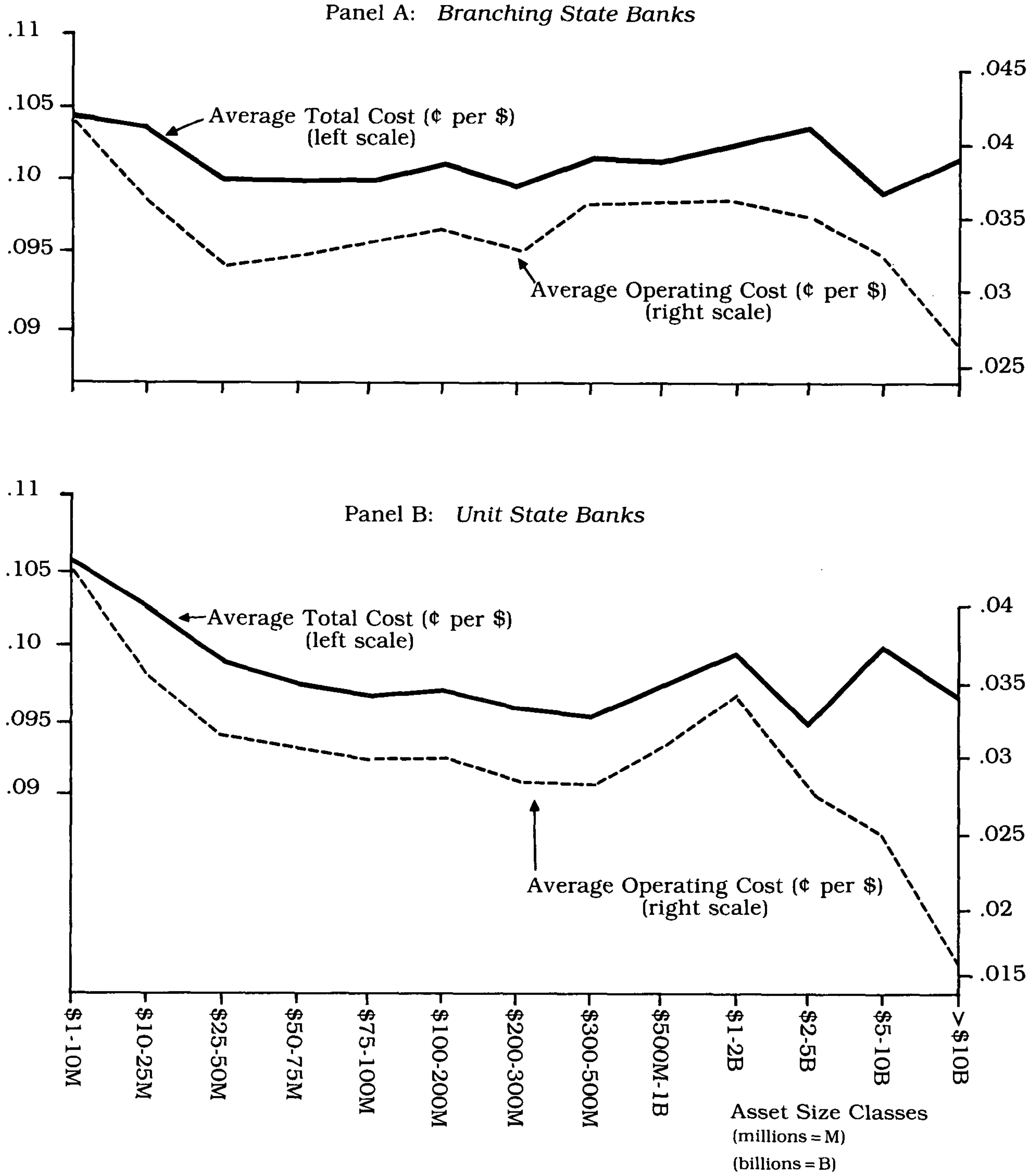
[3] The same sort of bias toward finding scale economies when only operating costs are used also exists for thrift institutions. This can be seen in the raw data presented in Verbrugge, McNulty, and Rochester, 1990, Table 1.

[4] If the U.S. banking system were considerably more consolidated, as could occur if full interstate branching were permitted, then the importance of purchased funds would of course be reduced. Once this occurs, looking at operating cost per dollar of assets could be more revealing. There would be less substitution of purchased funds for produced deposits and the funding mix bias that exists in current studies using only operating cost would be attenuated.

Figure 2

Comparing Actual Average Operating and Average Total Cost

(1984 data points)



Panel A: *Branching State Banks*

Panel B: *Unit State Banks*

Data on the number of deposit and loan accounts, an output stock measure, also are not available for all banks. Nevertheless, some information is given in the Federal Reserve's annual *Functional Cost Analysis* (FCA) survey. This survey covers 400 to 600 banks but typically excludes the very largest (those with more than $1 billion in assets). Also, the same banks are not in the sample each year.[5] Alternatively, the value of dollars in the various deposit and loan accounts, another output stock measure, is publicly available for each individual bank in every year in the *Report of Condition and Income* (Call Report).

Some researchers have made a strong argument for using the number of accounts as an indicator of bank output (Benston and Smith, 1976). Fortunately, it turns out that the scale economy results are reasonably robust to the use of either the number of accounts or the dollar value in the accounts. That is, using both of these alternative representations of bank output in the same model for the same year leads to similar scale economy results (Benston, Hanweck, and Humphrey, 1982; Berger, Hanweck, and Humphrey, 1987). This occurs because these two approximations to bank output, while numerically quite different, are highly correlated, both in the U.S. and elsewhere (see Berg, Forsund, and Jansen, 1990).

A preferable measure for bank output would measure the flow of some physical aspects of bank output rather than just the stock of accounts serviced or their dollar values. While the Bureau of Labor Statistics compiles such a measure annually, it applies only to the aggregate of all banks in the U.S. (BLS, 1989). This aggregate flow measure is a specially weighted index of the number of checks processed (for demand deposit output), the number of savings account deposits and withdrawals (for savings and small-denomination time deposit output), the number and type of new loans made (for various loan outputs), and the number of trust accounts serviced (for trust output).[6]

Over a recent 10-year period (1977-86), the BLS aggregate measure of bank output rose by 40.4 percent. Over the same period, a cost share-weighted index of the *value* of demand deposits, savings and small time deposits, real estate, installment, and commercial and industrial loans (all deflated by the GNP deflator) rose by 43.8 percent (Humphrey, forthcoming). These 5 output stock categories accounted for around 75 percent of bank value-added during the 1980s and so clearly reflect the majority of services produced by banks (in a flow sense). Importantly, the similar growth rates indicate, at the aggregate level at least, that the flow and stock measures of bank output closely correspond to one another. This suggests that use of a stock measure of bank output (the only one available at the individual bank level for all banks) may be a reasonable approximation of the unobserved flow measure for recent time periods. Thus it would seem that little bias has been introduced in past scale economy studies when a stock of output measure is used in place of a flow measure. Also, either the stock of accounts or the stock of dollars in those accounts seems to give qualitatively similar scale economy results (when properly used in the same model).

A related issue, often noted in the literature, concerns the similarity of the survey bank data from the FCA versus that for the population of all banks in the Call Report. The only published study addressing this issue concluded that while there were statistically significant differences between the FCA sample and the Call Report population data (in terms of portfolio composition, capital/asset ratio, and total cost/asset ratio), these differences were quantitatively small. In fact, FCA banks in 1970 experienced mean average costs which were 6 percent lower than the average costs for the mean of the non-FCA bank size-matched sample (Heggestad and Mingo, 1978). Updating this comparison for 1984, but using all banks, we find that the mean difference is now only 3 percent, and most of this arises for banks with the highest costs. Thus, FCA data should not lead to markedly different scale economy results compared to use of data on all banks, or on only large banks not covered in the FCA sample.

V.
A Linear versus a Quadratic Functional Form

Historically, bank scale economies were typically estimated using a linear functional form for equation (1), such as the log-linear Cobb-Douglas form.[7] Such forms were commonly used in cost or production analyses in areas where the research emphasis was

[5] The sample has varied by as much as 15 to 20 percent each year. Also, credit unions and thrift institutions (such as MSBs) can and do participate in the FCA survey. In 1984, the participation rate of thrift and credit unions was almost 17 percent of the total sample.

[6] The FCA data also provide physical flow information, similar to that used by the BLS, but these data are available only for banks in the survey, not for all banks.

[7] Greenbaum, 1967, is an important exception as he used a simple quadratic equation and, as a result, found a U-shaped average cost curve (in contrast to studies using a Cobb-Douglas form).

on factor shares in the distribution of income and on estimating the various sources of output growth over time. Unfortunately, one property of the log-linear Cobb-Douglas form is that the *same* cost economies or diseconomies will be measured for *all* banks in the sample regardless of their size. Put differently, all banks will either have scale economies, scale diseconomies, or constant costs. A U-shaped long-run cost curve, similar to that illustrated in Figure 1, cannot be estimated when only Q enters the regression equation (1). What is needed is a specification that includes Q and Q^2, making (1) a quadratic equation.

Earlier studies, such as the comprehensive analyses of Benston, 1965 and 1972, and Bell and Murphy, 1968, used a Cobb-Douglas form and found that scale economies existed in many banking services.[8] Overall, these economies were relatively small. The average scale economy value was .92.[9] This means that for each 10 percent increase in bank output, costs rise by only 9.2 percent, so average costs would be estimated to fall as a bank gets larger. A scale economy value greater than one—say 1.05—would have suggested a 10.5 percent rise in costs for each 10 percent increase in output (thus reflecting scale diseconomies).

Recently, more flexible functional forms have been developed and used. One of the most common is the translog form, which is a quadratic form. That is, the translog has linear output terms, like the Cobb-Douglas, but also squared output terms. As a result, the translog form can estimate a U-shaped cost curve if one exists in the data. If a U-shaped cost curve were in fact estimated, it would show scale economies at smaller banks and diseconomies at larger ones, like that illustrated in Figure 1. Unlike the Cobb-Douglas form, quadratic forms capture variations of scale economies across different sizes of banks.

Studies using the translog form, such as Gilligan, Smirlock, and Marshall, 1984, Lawrence and Shay, 1986, or Benston, Hanweck, and Humphrey, 1982, generally find that bank cost curves are weakly U-shaped. Scale economies exist in banking but seem to be limited to the relatively smaller banks. Either constant costs (for banks in branching states) or some scale diseconomies (for those in unit banking states) seems to apply to larger institutions. Since under certain restrictions the translog reduces to the Cobb-Douglas form, it is possible to see if these restrictions significantly reduce the ability of the model to fit the underlying data. In these tests, the Cobb-Douglas has been rejected in favor of the more general translog form. That is, the restrictions the Cobb-Douglas form places on the translog model (equal scale economies for all sizes of banks and all elasticities of factor input substitution equal to 1.0) are rejected.

Use of the translog instead of the Cobb-Douglas is one way these restrictions can be relaxed. Another way is through a specialized adjustment (called a Box-Cox adjustment) to the Cobb-Douglas model, as applied by Clark, 1984, and Lawrence, 1989. With such an adjustment, Clark finds only scale economies in his small and medium-sized unit bank data set (the largest bank had only $425 million in assets). In contrast, Kilbride, McDonald, and Miller, 1986, find scale economies at small unit banks but diseconomies at large ones using the same technique as Clark. Since the Kilbride, et al. study differs in two respects—it covered a later time period (1979-83 versus Clark's 1972-77) and added large unit banks up to $10 billion in assets to the unit bank sample—it is not clear which change led to the reversal in Clark's results: the different time period covered, the inclusion of large banks, or both.

Recently, Lawrence, 1989, generalized the Box-Cox adjustment of the Cobb-Douglas model by adding the possibility of multiple outputs—either multiple classes of loans or loans plus certain types of deposits. Both the Clark and the Kilbride, et al., studies had used a single composite measure of bank output. With this adjustment, both the multiple output translog and the single output Cobb-Douglas forms can be tested to see which form best fits the data. The single output Cobb-Douglas form, even with a Box-Cox adjustment, was rejected in favor of the multiple output translog. Thus it appears that both the possibility of U-shaped cost curves and cost complementarities among different bank outputs are important generalizations of the single output Cobb-Douglas form (which cannot reflect either of these more flexible specifications). In sum, a functional form that permits the estimated average cost curve to be U-shaped, rather than monotonic, is preferred. Thus a quadratic form dominates a linear form when

[8] Squared terms of some independent variables were used in Benston's regressions but only rarely applied to the output variables. Thus U-shaped cost curves could not, except in these infrequent cases, be estimated.

[9] Simple averages of Benston's, 1965, direct and indirect expense scale economies were .87 and .98, respectively (Table 26, p.544). As indirect expenses were 43 percent of total operating expenses, this yielded a weighted average scale economy of .87(.57) + .98(.43) = .92. Bell and Murphy obtained an overall scale economy of .93 (Table 4, p.8).

measuring bank scale economies and typically yields different scale economy conclusions as well.

Closely related to the choice of a proper functional form is the assumed constancy of the estimated relationship for all sizes of banks. More precisely, all banks in a particular sample are presumed to lie on the same average cost curve. While some studies estimate scale economies for only large banks and others estimate these economies for small and medium-sized institutions, few have systematically tested to see if all banks lie on the same curve, and therefore face the same technology. This hypothesis has been rejected statistically (Lawrence, 1989), likely due to the large samples which produce a very peaked sampling distribution. However, contrasts of published results for large and small banks separately suggest that scale economy values may not differ much in an economic sense. That is, the relatively flat U-shaped cost curves identified using all banks are replicated when only large banks are used separately (e.g., Noulas, Ray, and Miller, 1990). In either case, it is clear that on average the very largest banks do not appear to have a significant cost advantage due to scale economies compared to most smaller institutions.

VI. Scale Economies at the Office or Banking Firm Level

When only bank-incurred costs are being minimized, scale economies for the average banking office and the average banking firm—both derived from equation (1)—should be the same. But when costs include both the production and the *delivery* of output to the customer, as occurs in banking, these two measures can differ. In effect banks minimize both bank and customer-incurred costs together, but only the bank portion is observed. Some banks will find it profitable to do more delivery—branching—than others. These banks will save customers' transportation and transaction costs (Nelson, 1985, Evanoff, 1988) but will add to bank costs, and so look to be less efficient compared to others which provide less delivery. As customer costs are unobserved, differences in delivery strategies can give the appearance of higher than minimum bank costs, even though profits may be maximized in either case. In this situation, scale economies can be measured at the office level (as seen in the results of Lawrence and Shay, 1986, who only measure office economies) while diseconomies can be measured at the firm level (as found in Hunter and Timme, 1986, and Berger, Hanweck, and Humphrey, 1987).

Some insight into resolving this difficulty, however, may be obtained by observing how banks behave when they have virtually no branches. Here the office *is* the firm. This is the result when scale economies are estimated for banks in unit banking states.[10] Scale diseconomies are regularly observed for the larger unit banks. Because these banks have (except in rare instances) no branch network to provide "convenience" to customers, these diseconomies must therefore be related to production inefficiencies alone, not to the extra expense of providing consumer convenience. In contrast, banks operating in branching states and hence providing customer convenience through a branching network have lower scale diseconomies at the firm level and slight economies at the office level (for all sizes of banks). Thus it appears that permitting a bank to branch will itself lower costs for the larger banks. The implication is that branching, far from being an extra cost of customer convenience, actually *lowers* both bank and customer costs. Branching permits a banking firm to lower costs by producing services in more optimally sized "plants" or offices rather than producing virtually all of the output at a single office, as occurs in unit banking states.[11] Thus the customer convenience aspect of branching would appear to be largely a side effect of a bank's desire to lower scale diseconomies by choosing a more optimal configuration of production facilities.

For banks in branching states, which in 1988 included all but Colorado, Illinois, Montana, and Wyoming, the average number of accounts per banking firm rises steadily with bank size, while the average number of accounts per office remains steady after a certain minimum is reached. This fact implies that branching banks can add output (deposits and loans) in either of two different ways: by adding additional offices in new market areas (which attract new accounts and balances) or by adding new accounts and balances to existing offices. The data indicate that the former method of output expansion, which includes internal growth as well as mergers, dominates the latter (Benston, Hanweck, and Humphrey, 1982, Table 1).

[10] Early on, published studies lumped banks in unit banking and branching states together. This is inappropriate since more recent studies have shown that these two classes of banks are significantly different from one another in terms of how costs vary with size. It should be noted that banks in unit banking states do at times have a limited number of branches while unit banks—those with no branches—exist in branching states.

[11] Two studies which contrast unit and branching bank scale economies are Benston, Hanweck, and Humphrey, 1982, and Berger, Hanweck, and Humphrey, 1987. Other studies generally parallel these results for banks in these two different regulatory environments.

To determine economies at the average banking office, the number of branches is included as an explanatory variable in equation (1) and scale economies at the office level are obtained from a partial derivative of the estimated total cost equation with respect to scale (or output) alone. For economies at the average banking firm, the same model is estimated but the total derivative of the equation with respect to both scale and number of branches is used. Equivalently, the variable measuring the number of branch offices can be deleted from (1) to obtain scale economies at the firm level. The results typically indicate that the average office still has some realizable scale economies whereas for the firm, these economies have either disappeared or have turned into slight diseconomies.

Researchers have in the past estimated scale economies for the average banking office and then conclude that large banks (banking firms) have lower costs. They do so without realizing there can be a difference between the office and firm results. In fact, most of the early studies of bank scale economies are deficient in this regard because they typically specified the number of branches as an independent variable in their estimating equation and then proceeded to derive scale economies as the partial derivative of costs with respect to output. But this derivation only gives scale economies when the number of banking offices is held constant and thus reflects only one of the two ways that bank output expansion can affect costs. A better approach is to compute scale economies both ways, and be clear about what concept is being measured, or to compute only those economies which apply to the banking firm as a whole—the relevant concept for policy purposes. That is, most policy issues in banking, whether relating to interstate banking, foreign bank competition, or bank costs faced by users, are a function of the relation between costs and firm size, not costs and the size of the average office. The prices of banking services necessarily reflect all banking costs, so the former, not the latter, is the appropriate point for scale economy evaluation.

VII.
Time Periods With High Versus Low Interest Rates

The time period chosen for a cross-section study of scale economies can affect the estimated slope of the average cost curve. The reason is that total costs—the appropriate cost concept to use when measuring scale economies—will vary over the interest rate cycle and alter the slope of the estimated cost curve.

Each of the three major components of average cost—purchased funds interest cost, core deposit interest cost, and the prices of factor inputs which comprise operating cost—are influenced by the interest rate cycle in cross-section data sets, but by differing amounts and with different lags. For example, average operating cost rises, with a lag, with the rate of inflation while the average cost of purchased funds rises immediately and fully reflects the level of market interest rates. In contrast, the average interest cost of core deposits almost always rises by less than the rise in market rates and usually with a lag. Since larger banks rely more on purchased funds, it is easy to see that large banks will necessarily have higher average costs than smaller banks when interest rates are high. This holds even if equal average costs would prevail across all banks when interest rates are at their "normal" level. Similarly, the reverse can hold if interest rates are at an exceptionally low level.[12]

Simply put, the slope of the average cost curve and estimates of bank scale economies can differ when they are based on single year cross-section data simply because the level of the market interest rate varies over time. Since the vast majority of scale economy estimates are in fact derived from single-year cross-section studies, interest rate variations can be an important consideration in explaining why some studies show more or less scale economies than others. Such variations are especially important when studies conducted in the 1960s and early 1970s, periods of relatively low interest rates, are contrasted with studies of the late 1970s and early 1980s, periods of unusually high rates. But even over 1980-84 when rates were high there was enough variation in the market interest rate to alter the slope of the average cost curve, shifting around the large banks so that small scale economies became small diseconomies (Humphrey, 1987, Figures 4a and 4b).

To abstract from this problem, time-series studies are needed since they can control for the year-to-

[12] If core deposits could be easily and rapidly substituted for purchased funds when market rates were relatively high, and vice versa when these rates were low, then the slope of the average cost curve would not be dependent on the interest rate cycle in the manner just described. But since such substitution is quite limited in practice, and because core deposits are typically treated as quasi-fixed inputs to the banking firm (Flannery, 1982), the effects of the interest rate cycle on cross-section scale economy estimation are operative.

year variation in the level of interest rates.[13] It turns out that those few time-series studies that do exist show constant costs for large banks—a flat average cost curve—when evaluated using the average interest rate over the sample period (Hunter and Timme, 1986). When a broader sample of banks are used over time, slight economies are measured for small banks (around .95) and slight diseconomies for the largest banks (around 1.05).[14] Overall, these time-series results are quite similar to many, but not all, of the studies that used cross-section data for a single year.[15] Thus, while the time period can affect the slope of the average cost curve and therefore the estimate of the associated scale economy, in practice the bias appears to have been relatively small. In any event, the safest course is to rely on generalizations of a number of single year cross-section results (as Mester, 1987, and Clark, 1988, have done) rather than generalize from only a single one. The close correspondence between many cross-section studies and the few time-series studies which exist supports this conclusion.

VIII. Single Versus Multiple Bank Outputs

Until quite recently, scale economy estimates were based on how costs varied with changes in a single, aggregate (stock) measure of bank output. That is, Q rather than the separate and different bank outputs (Q_i) that make up Q were specified in equation (1). A problem with this approach is that there are at least two quite different reasons why costs may vary with an aggregate measure of output and only one of them reflects scale economies. The other reflects economies of scope, or cost changes related to the number and joint production nature of the different outputs produced. Scope economies occur when costs fall as product mix is expanded, allowing fixed costs to be spread over a larger number of different outputs.

In single-output studies, there is the possibility that economies associated with output levels have been confounded with economies associated with joint production. One may avoid this problem by specifying a multiproduct estimating framework (using a number of different Q_is), rather than relying on an aggregate index of the different outputs (where Q is a weighted sum of the Q_is). In this way, the two separate influences—scale and scope—can be separated.[16]

A number of studies have tested the (functional separability) conditions needed to justify a single index of bank output and have rejected them statistically (Kim, 1986). Even so, as often happens, statistical rejection has not led to economic rejection: the scale economy results from single output studies are quite similar to those found in multiproduct analyses. That is, slight but significant economies are measured at the office level (.96 to .98) for all sizes of banks whereas the average cost curve describes a relatively flat U-shape at the level of the banking firm, this shape indicating significant economies at small banks (around .94) but significant diseconomies at the largest (around 1.06).[17] As a result, biases that could be due to commingling scope economies with scale economies appear in practice to be slight. Banks produce very similar product mixes, on average, so that the importance of measured scope economies using current *observed* production is apparently small enough not to bias the scale economy results obtained specifying single versus multiple outputs.[18] In sum, there are strong theoretical reasons to (1) reject studies of scale economies that have aggregated all bank outputs into a single index and (2) use an explicit multiproduct specification in its place. In practice, however, the overall

[13] Making the average interest rate an independent variable in equation (1) will control for the small variation in this rate across banks in a cross-section analysis but will not control for the bias introduced if the level of interest rates are atypically high or low for the time period studied.

[14] These results are from unpublished work by the author using a panel of almost 700 banks over 1977-88 that accounted for $2 trillion out the $3 trillion in total U.S. banking assets.

[15] A large number of cross-section studies are summarized in the comprehensive surveys of bank scale economies done by Mester, 1987, and Clark, 1988. Their conclusions are similar to those here in that scale economies seem to exist for small banks while constant costs or slight diseconomies are measured at the largest.

[16] Strictly speaking, the relationship between scale and scope economies is $S_{1,2} = (W S_1 + (1 - W) S_2)/(1 - S_c)$ where $S_{1,2}$ is the measure of overall economies of scale (in a two-output situation), S_1 and S_2 are the product-specific scale economies of the two outputs, S_c is the scope economy measure, and W is a weight which is similar to the share of variable costs in total cost for output 1 (See Bailey and Friedlaender, 1982, pp. 1031-32). Thus, the measure of overall economies of scale is related to scope economies in the usual aggregate (single) output situation. Even if S_1 and S_2 show constant costs, the overall scale measure ($S_{1,2}$) can falsely reflect economies or diseconomies depending on the value of scope economies (S_c).

[17] These results hold for both banks in unit banking and branching states, with the exception that the results noted in the text for the firm also apply to the average office in unit states (Berger and Humphrey, 1990).

[18] This result refers to the small expansion path subadditivity results in Hunter, Timme, and Yang, 1988, and Berger, Hanweck, and Humphrey, 1987. Scope economies are a special case of subadditivity and the complete specialization needed to reflect the scope concept is rarely seen in banking.

measure of scale economies is little affected by this adjustment.[19]

IX. All Banks Are Efficient Versus Only Those on the Frontier

A final source of bias in the estimation of bank scale economies is the possibility that the economies exhibited by the set of most efficient or "best practice" banks can differ from those exhibited by all banks, efficient and inefficient. The potential for such bias exists because scale economies measured using all banks may be affected by other inefficiencies, unrelated to scale. These other factors would give a distorted picture of the true scale effects obtainable if all banks were as well managed and efficiently organized as those best practice banks with the lowest average costs.

This possibility arises because substantial cost differences, likely reflecting inefficiencies, seem to exist in banking (Humphrey, 1987). When all banks are stratified by size and then divided up into quartiles based on their levels of average costs for various years during the 1980s, the mean variation in average cost between the highest and lowest average cost quartiles of banks is 34 (31) percent for branching (unit) state banks. Since the mean variation in average cost across size classes was only 8 (12) percent, the variation between quartiles is seen to be 4 (2) times the variation across size classes. This pattern indicates that relative efficiency differences between similarly sized banks far exceed those obtainable by only altering bank size.[20]

To put these results differently, if a $500 million asset bank experienced a drop in its average cost from the mean of the highest to the mean of the lowest average cost quartile, costs would have fallen by 31 to 34 percent. Such a cost reduction would be equivalent to a scale economy value of .69 to .66. Since this figure far exceeds most estimates attributable to scale economies (e.g., .95), it is seen that even the existence of substantial scale economies at higher cost banks will *not* enable them to become competitive with smaller *or* larger banks that happen to be in the lowest cost quartile. Thus the competitive implications of scale economies at large banks are qualified by the existence of offsetting differences in cost levels or relative efficiency for all sizes of banks.[21]

Surprisingly, given the large differences in average costs between low- and high-cost banks, the scale economy results for banks in the lowest cost quartile (and therefore on the efficient cost frontier) are very similar to those obtained when all banks are pooled together (Berger and Humphrey, 1990). Thus while there are considerable differences in cost efficiency across banks, these differences do not significantly affect the scale economy results or conclusions of the previous section. Frontier analyses, which focus on low-cost or efficient banks, give the same results as the more traditional studies which estimate scale economies for all banks in a sample.

X. Summary and Conclusions

There are important economic and political issues related to the size of scale economies in banking. Measurement of these economies is an empirical issue and, when many studies exist, it is possible to sort out the likely reasons for seemingly conflicting results. Such an understanding of the data and the results of different research designs permits the derivation of a consensus position useful for policy purposes.

Seven common differences in existing bank scale economy studies have been identified and discussed. These are summarized in Table I. Of the seven, only three (numbers 1, 3, and 4) led to problems sufficiently serious to warrant discounting the conclusions of studies incorporating them. Analyses which relate operating costs—not total costs—to variations in bank output contain a bias due to differences in the funding mix across banks. As a result, these analyses are typically biased toward finding scale economies when

[19] One benefit of a multiproduct specification, however, is that scale economies for each output can be determined separately and contrasted. The scope economy results derived from a multiproduct specification have, however, been disappointing as there has been a lack of consistency in the value of scope economies estimated. It has been shown that one reason for the markedly different scope economy results in different studies is a limitation in the translog functional form itself (virtually the only form used today in banking studies). When a form that better fits the data is used instead, consistent values for scope economies result regardless of the point of evaluation (Pulley and Humphrey, 1990).

[20] These differences are not due to chance occurrences of high or low costs among banks as they exist for the same banks during different time periods, when chance variations would be expected to average out. As well, low-cost banks consistently have higher profits (and vice versa). Thus whatever is happening on the cost side rolls over to the revenue side as well, rather than being the result of high-cost banks producing a different output which is offset by higher revenues (Berger and Humphrey, forthcoming).

[21] Similar conclusions apply to thrift institutions (Verbrugge, McNulty, and Rochester, 1990).

Table I

Summary of Differences Among Bank Scale Economy Studies

Common Differences:	Bias Found:
1. Cost Definition (operating versus total cost)	Use of operating cost gives bias toward finding scale economies.
2. Output measurement (number of accounts versus dollars in the accounts)	Either output measure gives similar results.
3. Functional form (linear versus quadratic)	Linear (Cobb-Douglas) form gives bias toward finding scale economies.
4. Point of scale economy evaluation (single office versus banking firm)	Evaluation for average banking office not relevant for policy purposes.
5. Time period used (high versus low interest rates)	Bias exists but is minor.
6. Commingling scale with scope (single versus multiple outputs)	Similar scale economy results with either single or multiple outputs.
7. Efficiency differences (average bank versus those on frontier)	No effect on scale economy results.

none may exist after proper account is taken of all costs associated with producing bank outputs. Thus, believable scale economy estimates should be based on models using total costs, not just operating costs. As well, a quadratic functional form such as the translog that permits a U-shaped cost curve to be estimated if it exists in the data, is always favored over a linear function such as the Cobb-Douglas. This eliminates the majority of the earlier studies in which the (log linear) Cobb-Douglas form was used and scale economies were regularly (mis)identified. Lastly, only those scale economies evaluated at the level of the banking firm are pertinent to the policy issues at hand since it is the size of the banking firm, not the size of the average office, which captures the full cost efficiency associated with the two ways that bank output can be expanded. While some problems are encountered in using different measures of bank output, selecting different time periods for estimation, commingling scale with scope economies, and pooling efficient with inefficient banks, the resulting scale estimates obtained in these four cases are reasonably robust to these different treatments.

Overall, a consensus conclusion of the preferred studies on bank scale economies suggests that the average cost curve in banking reflects a relatively flat U-shape at the firm level, with significant economies at small banks (around .94) but small and significant diseconomies at the largest (around 1.06). This relatively flat U-shape also holds even when large banks are viewed separately. The implication is that the slight diseconomies identified for all large banks together represents an average for some of the smaller large banks possessing economies and the very largest which seem to possess diseconomies.

From these results, some practical conclusions may be inferred. First, there would seem to be little benefit of a cost-reducing nature from a marked increase in bank size alone, although significant benefits from loan diversification would exist for giant nationwide banks. Second, the measured scale or cost economies are small in comparison to existing differences in cost levels between similarly sized banks. This finding implies that even if cost economies were pervasive, which they are not, they would have a much smaller competitive impact than has been heretofore presumed. The large and persistent cost differences between banks of a similar size and product mix suggest that greater competition within the banking industry would be beneficial but that this need not be associated with bank size. One way to enhance competition is to permit easier entry into and exit from the industry. A step in this direction will come with full interstate banking during the next decade when geographical restrictions on entry are to be removed.

References

Bailey, Elizabeth, and Ann Friedlaender, "Market Structure and Multiproduct Industries," *Journal of Economic Literature, 20* (September 1982), 1024-48.

Bell, Frederick, and Neil Murphy, *Costs in Commercial Banking: A Quantitative Analysis of Bank Behavior and its Relation to Bank Regulation*, Research Report No. 41, Federal Reserve Bank of Boston, Boston, MA (1968).

Benston, George, "Branch Banking and Economies of Scale," *National Banking Review, 2* (June 1965), 507-49.

————, "Economies of Scale of Financial Institutions," *Journal of Money, Credit, and Banking, 4* (May 1972), 312-41.

Benston, George, and Clifford Smith, Jr., "A Transactions Cost Approach to the Theory of Financial Intermediation," *Journal of Finance, 31* (May 1976), 215-31.

Benston, George, Gerald Hanweck, and David Humphrey, "Scale Economies in Banking: A Restructuring and Reassessment," *Journal of Money, Credit, and Banking, 14* (November 1982), 435-56.

Berg, Sigbjorn, Finn Forsund, and Eilev Jansen, "Technical Efficiency of Norwegian Banks: the Non-Parametric Approach to Efficiency Measurement," Working Paper, Research Department, Bank of Norway, Oslo, Norway (January 1990).

Berger, Allen, Gerald Hanweck, and David Humphrey, "Competitive Viability in Banking: Scale, Scope, and Product Mix Economies," *Journal of Monetary Economics, 20* (December 1987), 501-20.

Berger, Allen, and David Humphrey, "The Dominance of Inefficiencies Over Scale and Product Mix Economies in Banking," Working Paper, Board of Governors of the Federal Reserve System (May 1990).

Berger, Allen, and David Humphrey, "Measurement and Efficiency Issues in Banking," *Output Measurement in the Services Sector*, Conference on Research in Income and Wealth, National Bureau of Economic Research, University of Chicago Press (forthcoming 1991).

Board of Governors of the Federal Reserve System, *Functional Cost Analysis*, National Average Report, Commercial Banks, Washington, D.C. (Various years).

————, *Report of Condition and Income*, Washington, D.C. (Various years).

Bureau of Labor Statistics, U.S. Department of Labor, *Productivity Measures for Selected Industries and Government Services*. Bulletin 2322, Washington, D.C. (February 1989), 170.

Clark, Jeffrey, "Estimation of Economies of Scale in Banking Using a Generalized Functional Form," *Journal of Money, Credit, and Banking, 16* (February 1984), 53-68.

————, "Economies of Scale and Scope at Depository Financial Institutions: A Review of the Literature," Federal Reserve Bank of Kansas City *Economic Review, 73* (September/October 1988), 16-33.

Evanoff, Douglas, "Branch Banking and Service Accessibility," *Journal of Money, Credit, and Banking, 20* (May 1988), 191-202.

Evanoff, Douglas, Philip Israilevich, and Randall Merris, "Technical Change, Regulation, and Economies of Scale for Large Commercial Banks: An Application of a Modified Version of Shephard's Lemma," Working Paper, Federal Reserve Bank of Chicago (June 1989).

Flannery, Mark, "Retail Bank Deposits as Quasi-Fixed Factors of Production," *American Economic Review, 72* (June 1982), 527-36.

Gilligan, Thomas, Michael Smirlock, and William Marshall, "Scale and Scope Economies in the Multi-Product Banking Firm," *Journal of Monetary Economics, 13* (May 1984), 393-405.

Greenbaum, Stuart, "A Study of Banking Costs," *National Banking Review, 4* (June 1967), 415-34.

Heggestad, Arnold, and John Mingo, "On the Usefulness of Functional Cost Analysis Data," *Journal of Bank Research, 8* (Winter 1978), 251-56.

Humphrey, David, "Cost Dispersion and the Measurement of Economies in Banking," Federal Reserve Bank of Richmond *Economic Review, 73* (May/June 1987).

————, "Cost and Technical Change: Effects of Bank Deregulation," *Journal of Productivity Analysis*, (forthcoming 1991).

Hunter, William, and Stephen Timme, "Technical Change, Organizational Form, and the Structure of Bank Productivity," *Journal of Money, Credit, and Banking, 18* (May 1986), 152-66.

Hunter, William, Stephen Timme, and Won Yang, "An Examination of Cost Subadditivity and Multiproduct Production in Large U.S. Banks," Working Paper, Department of Finance, Georgia State University, Atlanta, GA (December 1988).

Kilbride, Bernard, Bill McDonald, and Robert Miller, "A Reexamination of Economies of Scale in Banking Using a Generalized Functional Form," *Journal of Money, Credit, and Banking, 18* (November 1986), 519-26.

Kim, Moshe, "Banking Technology and the Existence of a Consistent Output Aggregate," *Journal of Monetary Economics, 18* (September 1986), 181-95.

Langer, Martha, "Economies of Scale in Commercial Banking," Working Paper, Banking Studies Department, Federal Reserve Bank of New York (December 1980).

Lawrence, Colin, and Robert Shay, "Technology and Financial Intermediation in a Multiproduct Banking Firm: An Economic Study of U.S. Banks 1979-1982," in C. Lawrence and R. Shay (Editors), *Technical Innovation, Regulation and the Monetary Economy*, Ballinger, Boston, MA (1986), 53-92.

Lawrence, Colin, "Banking Costs, Generalized Functional Forms, and Estimation of Economies of Scale and Scope," *Journal of Money, Credit, and Banking, 21* (August 1989), 368-79.

Mester, Loretta, "Efficient Production of Financial Services: Scale and Scope Economies," Federal Reserve Bank of Philadelphia *Economic Review*, (January/February 1987), 15-25.

Nelson, Richard, "Branching, Scale Economies, and Banking Costs," *Journal of Banking and Finance, 9* (June 1985), 177-91.

Noulas, Athanasios, Subhash Ray, and Stephen Miller, "Returns to Scale and Input Substitution for Large U.S. Banks," *Journal of Money, Credit, and Banking, 22,* (February 1990), 94-108.

Pulley, Lawrence, and David Humphrey, "Correcting the Instability of Bank Scope Economies from the Translog Model: A Composite Function Approach," Working Paper, College of William and Mary, Williamsburg, VA (June 1990).

Verbrugge, James, James McNulty, and David Rochester, "For Thrifts, Bigger Doesn't Necessarily Mean Better," *Journal of Retail Banking, 12* (Summer 1990), 23-33.