Economic Review

FEDERAL RESERVE BANK OF KANSAS CITY

April 1984

The Productivity "Slowdown": A Sectoral Analysis

Theories of Price Determination

Deposit Insurance and the Deregulation of Deposit Rates

The *Economic Review* (ISSN 0161-2387) is published ten times a year by the Federal Reserve Bank of Kansas City. Subscriptions and additional copies are available without charge. Send requests to the Research Division, Federal Reserve Bank of Kansas City, 925 Grand Avenue, Kansas City, Missouri 64198. If any material is reproduced from this publication, please credit the source. Second-class postage paid at Kansas City, Missouri. Postmaster: send address changes to the address above.

Economic Review

FEDERAL RESERVE BANK OF KANSAS CITY

April 1984

The Productivity "Slowdown": A Sectoral Analysis

By Charles S. Morris

Widespread concern has surfaced in recent years over the slowdown of labor productivity growth in the United States since 1973. Although productivity grew rapidly in 1983, official measurements of annual productivity growth over the previous nine years averaged only 0.7 percent, far below previous rates. This slowdown is believed by many to be partly responsible for the low economic growth and high inflation in the 1970s and early 1980s. Accordingly, many have advocated governmental action. For example, the President's Council of Economic Advisers stated that "a program to stimulate productivity growth must be a keystone of economic policy."[1]

The advisability of policy actions to reverse the productivity growth slowdown depends on understanding its causes. If, as most economists believe, the slowdown reflects a decline in the underlying trend growth rate of productivity, it might be advisable to implement long-run policies to reverse the slowdown. If, however, there has not been a trend productivity growth slowdown, say, because of errors in official measurements of productivity, no policy action would be necessary.[2]

This article argues that there has been no slowdown in trend labor productivity growth since 1973. The first section presents an overview of productivity behavior in the postwar period, including traditional estimates of trend productivity growth since 1973. The second section discusses alternative explanations of measured productivity growth since 1973. The third section presents empirical evidence that trend productivity growth has not decreased since 1973 either in the economy as a whole or in major business sectors. Policy implications of these findings are discussed in the conclusion.

Charles S. Morris is an economist in the Economic Research Department at the Federal Reserve Bank of Kansas City. Lyle Matsunaga provided research assistance.

Overview of trend productivity growth

Labor productivity measures the amount of goods and services produced by one worker in a given time period.[3] Several factors affect labor

---

[1] Council of Economic Advisers, *Economic Report of the President*, U.S. Government Printing Office, 1981, p. 69.

[2] Recent research by Michael R. Darby has shown that there has not been a trend productivity growth slowdown since 1973 See Michael R. Darby, "The U.S. Productivity Slowdown: A Case of Statistical Myopia," *American Economic Review*, forthcoming June 1984. Preliminary evidence of the lack of a decline in trend productivity growth since 1973 was also found by George L. Perry, "Potential Output and Productivity," *Brookings Papers on Economic Activity*, 1977:1, pp. 11-47.

[3] Although productivity is measured in many ways, the most common measure, output per manhour worked (hourly productivity), is used here.

productivity growth. Some, such as the influences of business cycles, have only temporary effects. Others affect the long-run trend of productivity growth. This section explains the concept of trend productivity growth. Traditional methods of measuring trend productivity growth are then used to show why most economists believe trend productivity growth has slowed since 1973.

Trend labor productivity growth — the growth that would occur if all resources in the economy were fully employed at desired levels — depends on several factors. For every hour worked, more goods and services will be produced if better technological processes are used, if workers have more machinery with which to work, or if the workers themselves have better skills. Thus, trend labor productivity growth increases with increases in the speed of technological progress, the growth of capital relative to labor, and the growth of workers' skills.[4]

Because of cyclical factors, actual productivity growth can differ from trend productivity growth over short periods. The difference is called cyclical productivity growth. Cyclical fluctuations in the economy cause actual resource employment levels to deviate from fully employed desired levels. These cyclical variations in resource employment cause the actual time path of productivity to fluctuate about the trend path. Although rarely equaling trend, actual productivity moves toward the trend as the economy moves toward full employment of resources. Because the cyclical and trend components of productivity cannot be observed separately, studies of trend productivity often lead to different conclusions.

The difference between trend and actual productivity growth can be clarified by examining Figure 1. Suppose that the line AB represents the path of trend labor productivity. Because of cyclical variations in productivity, actual productivity, represented by the S-shaped curve, fluctuates about the trend path. For example, from time t_1 to time t_2 cyclical productivity growth is positive and actual growth exceeds trend growth.

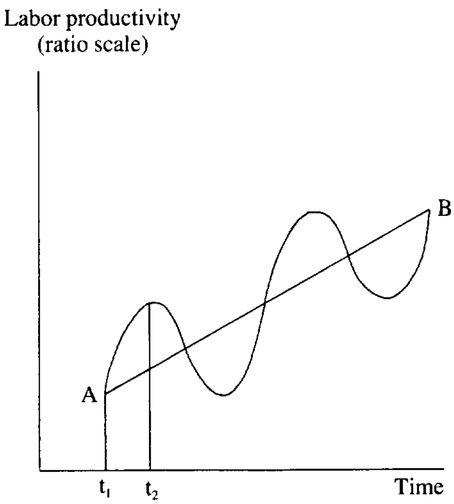
Trend productivity growth traditionally has been measured by actual growth from one year of high resource utilization to another. If trend productivity growth were estimated simply by calculating actual productivity growth over arbitrary periods of time, the estimate would partly reflect the influence of cyclical productivity growth. But if resource utilization rates are the same at the beginning and end of a period, actual productivity growth will be an accurate estimate of trend growth over that period.

Traditional estimates suggest that trend productivity growth slowed in the middle to late 1960s and slowed further after 1973. These estimates are shown in Table 1 for the U.S. private business sector and several subsectors. According to these estimates, which measure actual productivity growth between years of high resource utilization, trend productivity grew slower in every sector from 1968 to 1979 than from 1948 to 1968. The evidence also suggests that a further slowdown in trend growth started in 1974. The annual rate of trend productivity growth in the aggregate private business sector is estimated as having fallen from 3.1 percent over the 1948-68 period to 1.4 percent over the 1968-79 period. Within this later subperiod, it is estimated that trend productivity grew at an annual rate of 2.1 percent from 1968 to 1973, but only 0.8 percent from 1973 to 1979.

Although trend productivity growth first began to slow after 1968, most analysts have focused on the apparent second slowdown that began in

[4] Strictly speaking, trend productivity growth depends on the growth rate of the fully employed and desired ratio of capital to labor. Of course, many other factors also affect the growth of trend productivity. These other factors are not mentioned here because they are either not quantifiable or have not been found to have a statistically or economically significant influence on trend labor productivity growth in previous studies.

FIGURE 1
Hypothetical time paths of actual and trend labor productivity



1974. One reason for focusing attention on the period after 1973 is that policy actions aimed at reversing the earlier slowdown were deemed to be unnecessary. The major reason, however, is that traditional estimates suggest that the apparent slowdown that began after 1973 was more severe than the first slowdown. According to the traditional estimates in Table 1, the decline in the annual rate of aggregate trend productivity growth from the 1948-68 period to the 1968-73 period was only 1.0 percent. Indeed, trend productivity growth in the important manufacturing sector rose in the later period. For these reasons, the productivity growth slowdown that began in the 1960s is generally considered "small, readily explained, and not particularly worrisome."[5] In contrast, the traditional estimates show a more pronounced and more pervasive decline in trend productivity growth after 1973. According to these estimates, aggregate trend productivity growth fell to less than 1.0 percent a year, with all major sectors sharing in the overall decline.

Because the apparent severity of the productivity growth slowdown after 1973 is the basis for recommending government actions to stimulate productivity growth, the remainder of this article concentrates on explaining trend productivity growth since 1973. If it is found that the traditional estimates are misleading in suggesting a second slowdown in trend productivity growth after 1973, the case for government actions would depend on the magnitude and causes of slower productivity growth over the entire 1968-79 period.[6]

Alternative explanations of labor productivity growth since 1973

Many reasons have been given for the measured decline in labor productivity growth beginning in 1974. These explanations fall into two general categories. The conventional explanations hold that there actually has been a slowdown in trend productivity growth. Some of these studies attribute the slowdown in trend productivity growth to a slower rate of technological progress, some to a slower rate of capital accumulation, and others to the unexpected increases in energy prices in 1973-74. Michael R. Darby, on the other hand, attributes the measured decline in trend productivity growth to distortions in measured output resulting from the price controls of the early 1970s rather than to a true decline in trend growth.

---

[5] Edward F. Denison, "The Interruption of Productivity Growth in the United States," *The Economic Journal*, March 1983, p. 56.

[6] There is evidence that policy actions aimed at reversing the productivity growth slowdown over the entire period from the 1960s to the present are unnecessary. For example, Darby found that slow trend productivity growth over the 1965-79 period can be explained entirely by changes in the demographic composition of the labor force in terms of age, sex, place of birth, and education ("The U.S. Productivity Slowdown"). As new workers gain experience and learn new skills, however, trend productivity growth should increase automatically. Therefore, if demographic factors are the major cause of slow trend productivity growth over the 1965-79 period, policy actions are not only unnecessary but also inappropriate.

TABLE 1
Traditional estimates of trend productivity growth
(percent per year)

	U.S. Private Business Sector				
			Goods-Producing*		
	Total	Service-Producing	Total	Manufacturing	Farming
1948-79	2.5	2.0	3.0	2.6	4.7
1948-68	3.1	2.4	3.7	2.7	5.1
1968-79	1.4	1.1	1.7	2.4	4.0
1968-73	2.1	1.5	2.8	3.5	5.0
1973-79	0.8	0.8	0.8	1.5	3.2

NOTE: Based on data constructed by Elliot Grossman for the American Productivity Center.

*The goods-producing industries are manufacturing, farming, mining, and construction. Separate results are not reported for the mining and construction sectors due to data limitations.

Conventional explanations

Some studies, such as those by Edward Denison, attribute the decline in trend productivity growth since 1973 to a reduction in the pace of technological progress.[7] These studies explain trend productivity growth by using all variables other than technological progress that are thought to affect trend productivity growth. The part of trend productivity growth that these variables do not explain is attributed to technological progress. While these studies found a slowdown in trend productivity growth, they could not explain it and, therefore, attribute the slowdown to a reduction in the pace of technological progress.

Other studies claim that slower capital accumulation caused trend productivity growth to decline after 1973.[8] For this factor to explain a permanent decrease in trend productivity growth, it must be assumed that the slower capital accumulation rate reflects a decline in the long-run growth rate of the desired capital stock. The slower growth of desired capital relative to labor would, according to these studies, cause the growth rate of trend labor productivity to decline.

The growth path of trend productivity implied by these two explanations is shown in Figure 2A. Whether because of slower technological progress or slower capital accumulation, trend productivity grew slower after 1973 than before, as indicated by a decline in the slope of the trend productivity path after 1973.

The most popular explanation of the productivity growth slowdown is that it resulted from the energy price increases in late 1973.[9] According to the energy price explanation advocated by Robert Rasche and John Tatom, unexpected increases in the relative price of oil in 1973-74 reduced the amount of energy used in production. For a given level of hours worked, this reduction in energy use caused real output and, therefore, trend productivity to decline. Furthermore,

[7] For example, see Edward F. Denison, *Accounting for Slower Economic Growth: The United States in the 1970s*, Washington: The Brookings Institution, 1979, and Denison, "The Interruption of Productivity Growth," pp. 56-77.

[8] For example, see J. R. Norsworthy, Michael J. Harper, and Kent Kunze, "The Slowdown in Productivity: Analysis of Some Contributing Factors," *Brookings Papers on Economic Activity*, 1979:2, pp. 387-421.

FIGURE 2

Growth paths of trend productivity according to conventional explanations

Figure 2A

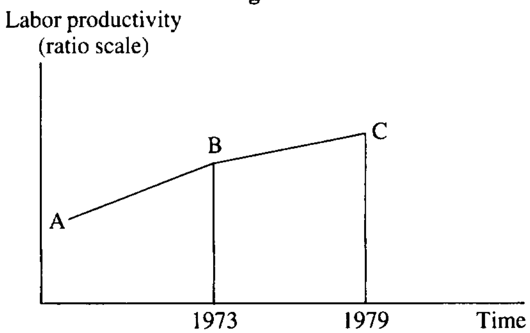
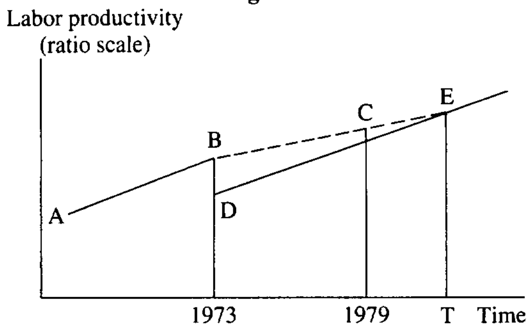


Figure 2B



because energy using capital is more expensive to operate, the desired capital growth path also declined, thereby causing a further decrease in the trend productivity growth path.

If the capital stock fell immediately, productivity would also fall immediately, so that trend productivity growth would not appear to have declined. Immediate adjustment of the capital stock does not occur, however. Instead, capital accumulation slows temporarily, with the result that the actual capital stock slowly approaches the lower desired level. Until the new desired capital stock is reached, the growth of capital relative to labor will decline, causing the growth of labor productivity to fall. Although the decline in productivity growth is only temporary, until the new growth path is reached, trend productivity growth will appear to have declined.

The growth path of labor productivity implied by the energy price shock explanation is shown in Figure 2B. According to this explanation, the trend productivity growth path shifts downward after 1973. This is shown by a parallel shift in 1973 of the trend productivity growth path from AB to DE. Actual productivity, however, does not decline immediately to the lower level. Instead, it gradually approaches DE along the path indicated by the dashed line BCE. Thus, the path of actual labor productivity is indicated by the path of ABE. Until the lower equilibrium level is reached at time T, trend productivity growth will appear to have declined.

The price control explanation

According to the price control explanation developed by Darby, trend productivity growth did not decline further after 1973 but only appears to have declined.[10] The price control pro-

[9] The extensive development of this explanation is due to Robert H. Rasche and John A. Tatom. See John A. Tatom, "Energy Prices and Capital Formation: 1972-1977," Federal Reserve Bank of St. Louis *Review*, May 1979, pp. 2-11, and Robert H. Rasche and John A. Tatom, "Energy Price Shocks, Aggregate Supply and Monetary Policy: The Theory and International Evidence," *Carnegie-Rochester Conference Series on Public Policy*, Spring 1981, pp. 9-93. See also Council of Economic Advisers, *Economic Report of the President*, U.S. Government Printing Office, January 1977, pp. 45-57.

[10] The theoretical argument is developed in Michael R. Darby, "Price and Wage Controls: The First Two Years," and "Further Evidence" in K. Brunner and A. H. Meltzer, eds., *The Effects of Price and Wage Controls, Carnegie-Rochester Conference Series on Public Policy,* Vol. 2, supplement to the *Journal of Monetary Economics*, April 1976, and Michael R. Darby, "The U.S. Economic Stabilization Program of 1971-1974," in M. Walker, ed., *The Illusion of Wage and Price Control*, Fraser Institution, Vancouver, 1976. The price control explanation is tested in Michael R. Darby, "The Price of Oil and World Inflation and Recession," *American Economic Review*, September 1982, pp. 738-51, and Darby, "The U.S. Productivity Slowdown."

gram of the early 1970s, it is argued, led to a measured level of productivity above the true level in 1973. Although the measurement error was eliminated by 1979, the estimate of trend productivity growth over the 1973-79 period was less than the true growth rate because the measured level of productivity was biased upward at the beginning of the period. Therefore, there appeared to be a decline in trend productivity growth even though there was none.

Under the Nixon administration's Economic Stabilization Program, price controls went through four phases lasting from the third quarter of 1971 through the third quarter of 1974. Phases I and II imposed strict controls on wages and prices from the third quarter of 1971 to the first quarter of 1973. During that time, the price of a product could be raised only if it could be shown that the quality of the product had been improved since controls went into effect and that higher costs were incurred in producing the higher quality product. Phases III and IV were periods of decontrol lasting from the second quarter of 1973 through the third quarter of 1974. Over that period it was much easier for producers to raise their prices. By the start of the fourth quarter of 1974, the price control program had been eliminated.

During phases I and II, when prices were essentially fixed, measured productivity growth overstated true growth. Phases I and II took place during a period when aggregate demand growth exceeded growth in real output. In the absence of price controls, this excessively rapid aggregate demand growth would have caused prices to rise. But because prices were fixed, measured prices increasingly understated the prices that would have prevailed. As this gap increased, the incentive for producers to evade the price controls increased. Producers could evade the price controls either by producing lower quality products without informing government officials, or by falsely claiming that they were producing a higher quality product so that prices could be increased somewhat. Either way, measured prices were less than true prices.[11] Although efforts are made to take account of quality changes in computing the price level, it is doubtful that the quality changes hidden from price administrators were adequately taken into account. As a result, measured prices fell progressively further below true prices, while measured real output (the nominal value of output divided by the measured price level) rose progressively above the true level of real output. Accordingly, measured labor productivity progressively overstated the true level of productivity. Therefore, measured productivity grew faster than true productivity over the first half of the price control program.

During phases III and IV, the decontrol period, an opposite sequence of events led to official measurements of productivity growth that understated the true growth. By the fourth quarter of 1974, when price controls were eliminated completely, the measured price level accurately reflected the true price level. As a result, measures of real output and productivity were accurate. Because measured productivity was greater

[11] For a price index to reflect true prices faced by individuals, changes in the quality of component products must be taken into account. This can be seen in the following example. Suppose that the weight of a 20 cent candy bar is reduced from 10 to 5 ounces, but the price is not changed. The quality reduction, in this case, takes the form of a weight reduction. Because only half as much candy is received for the same amount of money, the price of the candy has actually doubled. If such changes in quality are ignored when price indexes are computed, the indexes will be a misleading indicator of the true cost of living. In this example, the measured price understates the true price by 50 percent. Furthermore, the measured real output of candy bars (the nominal value divided by price) will be twice the true quantity of candy produced. To avoid these problems, attempts are made to take account of even very subtle changes in quality when price indexes are computed. For a microeconomic analysis at the firm and industry level of the effects of price controls on the quality of output, see Sung Hee Jwa, "Price Controls: A New Look at Old Matters — Theoretical and Empirical Analysis of the Effects of Price Controls on Quality Offerings," Ph.D. Dissertation, UCLA, Department of Economics, 1982.

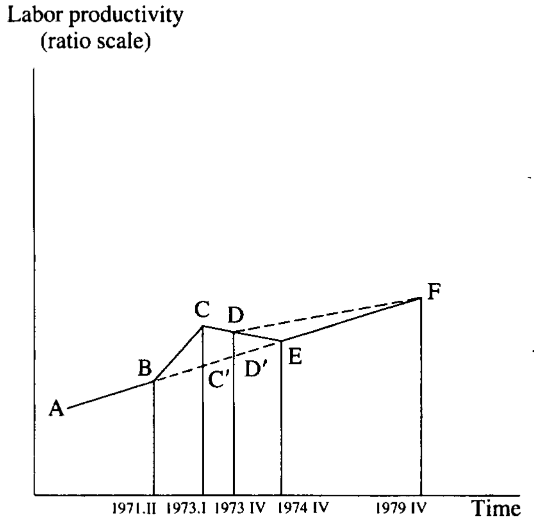
than true productivity at the start of phase III but equal to true productivity at the end of phase IV, measured productivity must have grown slower than true productivity over the decontrol period. Thus, measured productivity growth understated true productivity growth over the last half of the price control program.

The growth path of productivity implied by the price control explanation is shown in Figure 3. The growth rate of trend labor productivity is indicated by the slope of the trend growth path, AF. During phases I and II, measured productivity grew along BC, rising above the true growth path, BC′. During the decontrol phases, measured productivity returned to the true growth path along CE. From then on, labor productivity was measured accurately along the true growth path, EF.

Because the true level of productivity was overstated in 1973, the estimated growth rate of trend productivity over the 1973-79 period is less than the true growth rate. That is, trend productivity growth appeared to decline when, in fact, it did not. In Figure 3, the true growth rate of trend productivity over the 1973-79 period equals the slope of D′F. Trend growth, however, is estimated by the growth of measured productivity from 1973 to 1979, the slope of DF. Because the estimated growth of trend productivity is less than the true growth, trend productivity growth appeared to have declined.

The behavior of labor productivity implied by the price control explanation differs from that implied by the explanations that claim there was actually a further slowdown in trend productivity growth after 1973. To determine which of the explanations most nearly explains the actual behavior of productivity, it is necessary to test the alternatives in a complete model of productivity.

FIGURE 3
Growth path of trend productivity according to the price control explanation



An empirical analysis of productivity growth

This section presents a simple model of labor productivity behavior. The model is estimated by using both aggregate and disaggregated data from the private business sector. The main conclusion of the empirical analysis is that after taking proper account of cyclical factors and price controls, there is little evidence of a further slowdown in trend productivity growth after 1973.

A model of labor productivity

The model of labor productivity growth used in this study, which is similar to the one used by Darby, divides measured productivity growth into cyclical and trend components.[12] The model allows for the possibility that trend productivity growth slowed after 1968 and again after 1973. It also allows for the possibility that measured productivity growth appeared to slow after 1973

[12] See Darby, "The Price of Oil," and Darby, "The U.S. Productivity Slowdown."

because of measurement errors caused by price controls. As shown in Table 2, the model consists of five equations.

While traditional models decompose measured labor productivity growth only into cyclical and trend components, the model presented in Table 2 also accounts explicitly for the possibility that measured labor productivity growth may be systematically mismeasured. Equation 1 defines measured labor productivity growth, π, as the sum of cyclical productivity growth, π^C, trend productivity growth, π^T, and a systematic measurement error, π^E.

Cyclical productivity growth is represented by equation 2.[13] It is assumed that cyclical productivity growth depends on the unemployment rate, the layoff rate, and employment.[14]

The behavior of trend productivity growth is described by equation 3. The first term, α_0, represents the growth rate of trend productivity that would be observed if there were no slowdown in trend productivity growth in the last 20 years. A dummy variable, D69, represents the change in trend productivity growth that began in the middle to late 1960s. The first quarter of 1969 was chosen as the starting point for the first trend productivity growth slowdown.[15] Another dummy variable, D74, allows for a second slowdown in trend productivity growth after 1973. If the coefficient on D74, α_2, is *not* found to be statistically different from zero, it would support the hypothesis that there was no further decline in trend labor productivity growth after 1973.

The systematic measurement error predicted by the price control explanation is represented by equation 4. Unlike models used in most previous studies, the model used here allows for a systematic measurement error in measured productivity growth induced by price controls. The dummy variable DPC allows for the mismeasurement of productivity growth during the price control period from the third quarter of 1971 to the fourth quarter of 1974.[16] Because the price control explanation implies that the overstatement of true productivity growth during phases I and II is completely offset by the understatement during phases III and IV, the price control dummy variable, DPC, sums to zero. According to the price control explanation, the estimated value of the coefficient on DPC, ϕ, should be significantly positive.

The final equation of the model combines all of

[13] It is possible that traditional estimates of trend productivity growth show a further decline after 1973 partly because they do not take adequate account of cyclical factors. The traditional estimates are accurate only if the rate of resource utilization is similar in each cyclical episode. Because the cyclical episodes since 1973 have been very different from previous postwar cyclical periods, it may be the case that the rates of resource utilization are different as well. For example, the recession from 1973:IV to 1975:I and the most recent recession that extended from 1981:III to 1982:IV were the worst of the postwar recessions. Furthermore, 18 of the 96 quarters between 1948:IV and 1973:IV, or 18.8 percent, were recessionary. That percentage almost doubled for the quarters between 1974:I and 1982:IV to 33 3 percent when 12 of 36 quarters were recessionary.

[14] Specifically, cyclical productivity growth is assumed to depend on the change in the civilian unemployment rate, the change in the layoff rate, and the current and lagged growth rate of employment in manufacturing, mining, and construction. These variables are taken from Darby, "The Price of Oil," and Darby, "The U.S. Productivity Slowdown." Note that none of the cyclical variables is a nominal value deflated by a price index, but they are all based on counts of individuals. Darby points out that it is important that the cyclical variables are not nominal variables deflated by price indexes, such as measures of the real money stock. If the cyclical variables were deflated nominal variables, they would suffer from the same mismeasurement problem that real output and productivity suffer from and, therefore, completely explain the mismeasured productivity data.

[15] Because opinions about the beginning date of the middle to late 1960s slowdown are so varied, the following procedure was used to choose the starting date of 1969:I. Using data from the entire private business sector, equation 5 was sequentially estimated with the 1960s slowdown beginning in the first quarter of each year from 1964 to 1970. The initial year of the 1960s slowdown was chosen by selecting the equation with the lowest root-mean-square error. Although 1969:I is chosen as the first quarter of the 1960s slowdown, the results presented in the text are not significantly different from the results obtained using any other year as the starting point.

TABLE 2
A model of labor productivity growth

(1) $\pi = \pi^C + \pi^T + \pi^E$
(2) $\pi^C = \beta z$
(3) $\pi^T = \alpha_0 + \alpha_1 D69 + \alpha_2 D74$
(4) $\pi^E = \phi DPC$
(5) $\pi = \alpha_0 + \alpha_1 D69 + \alpha_2 D74 + \phi DPC + \beta z$

Definitions:

π = measured labor productivity growth rate
π^C = cyclical labor productivity growth rate
π^T = trend labor productivity growth rate
π^E = systematic measurement error in the measured labor productivity growth rate
z = a vector of cyclical variables that includes the change in the civilian unemployment rate, the current and lagged growth rate of employment in manufacturing, mining, and construction, and the change in the layoff rate
D69 = a dummy variable that represents the change in trend productivity growth after 1968
D74 = a dummy variable that represents the change in trend productivity growth after 1973
DPC = a dummy variable that represents the overstatement of labor productivity growth from 1971:III to 1973:I and the understatement of labor productivity growth from 1973:II to 1974:IV as predicted by the price control explanation

the information contained in equations 1 through 4. Empirical estimates of equation 5 can be used to determine whether trend productivity growth declined further after 1973 or whether, instead, the measured decline was an illusion caused by the systematic distortions resulting from price controls. That is, the estimates of the model can be used to answer the following question: after accounting for the effects of variations in cyclical economic activity, the late 1960s trend productivity growth slowdown, and the mismeasurement of productivity during the price control period, is there any evidence of a further decrease in trend productivity growth after 1973?

Empirical estimates of the model

The model was estimated for the private business sector as a whole and for several sectors within the private business sector. Although Darby has used a similar model to explain aggregate productivity behavior, the price control explanation has never been tested through use of productivity data from individual sectors.[17]

[16] See Darby, "The Price of Oil," and Darby, "The U.S. Productivity Slowdown," for a detailed description of DPC. Note that the sum of DPC represents the mismeasurement of the level of labor productivity. Because DPC sums to zero, the model implies that productivity growth was not mismeasured over the entire period from 1969:I to 1981:IV. Also, it should be clear now why the cyclical variables cannot be represented by deflated nominal variables. Because any deflated variable will also be mismeasured over the price control period, it would completely explain the productivity data during the price control period, and there would be nothing left for the price control dummy variable to explain.

[17] See Darby, "The Price of Oil," where international and U.S. data are used to show that the price control explanation explains real gross national product behavior better than the oil price shock explanation. See Darby, "The U.S. Productivity Slowdown," where productivity data for the entire U.S. economy are used to show that the price control explanation explains productivity behavior better than the oil price shock explanation.

Empirical estimates explaining productivity behavior in various sectors — goods-producing, service-producing, manufacturing, and farming — are useful for two reasons. First, policymakers can formulate better policies and implement them more efficiently when provided detailed information. Second, empirical findings would be more persuasive if consistent throughout the various sectors of the economy.

Empirical estimates of equation 5 are presented in panel A of Table 3.[18] (Complete results are reported in the Appendix table.) These empirical estimates show that trend productivity growth did not slow further after 1973. For every sector, the estimated coefficient on the variable D74 is not significantly different from zero. Moreover, the estimates are of the wrong sign (negative) for the farming and service-producing sectors. Finally, the estimate of the change in annual trend productivity growth after 1973 is small, never more than 1 percent.[19]

In addition to showing that trend productivity growth did not decline further after 1973, the results support the price control explanation of the decline in measured productivity growth. The estimated coefficient on the price control dummy variable DPC is positive for all sectors and statistically significant for all but the farming and manufacturing sectors.[20] Furthermore, the estimates of the price control coefficients are large enough to be significant in an economic sense. For example, for the private business sector as a whole, the measured productivity growth rate overstated the true growth rate by 2.4 percentage points a year during phases I and II of the price control period.[21] Thus, the empirical estimates confirm that the apparent slowdown in trend productivity growth beginning in 1974 results from mismeasuring true productivity growth over the period of price controls.[22]

Estimates of trend productivity growth implied by the empirical estimates of equation 5 of the model are presented in panel B of Table 3.[23] The estimated growth rate of trend productivity for each sector from the third quarter of 1948 to the

---

[18] The data consisted of quarterly observations from 1948:III to 1981:IV. The productivity data are from a database constructed by Elliot Grossman for the American Productivity Center. All other data series are from the Citibank Economic Database. The estimation period ends in 1981:IV because the layoff rate series is not available after that date. Dropping the layoff rate variable and extending the period of estimation to 1983:II produced no significant changes from the results reported in the text.

[19] Because the equations are estimated with quarterly data, all estimated coefficients are multiplied by 400 to get annual growth rates.

[20] Note in the Appendix that no coefficient is significantly different from zero in the farming sector except for the constant term and the autoregressive parameter. The price control dummy variable is marginally significant in the manufacturing sector. Apparently, there is not even an illusory trend productivity growth slowdown in manufacturing after 1973 because the estimates of both α_2 and ϕ are insignificantly different from zero. Furthermore, the data indicate that there was not even a late 1960s trend productivity growth slowdown in manufacturing. That is, all productivity changes in the manufacturing sector can be explained by cyclical factors alone.

[21] The measured overstatement is calculated by multiplying the price control coefficient estimate of 0.0422 by one-seventh and then annualizing that product by multiplying by 400.

[22] Some may argue that the statistical significance of the DPC dummy variable results from the "end-of-expansion effect" (EOE) as described in Robert J. Gordon, "The 'End-of-Expansion' Phenomenon in Short-Run Productivity Behavior," *Brookings Papers on Economic Activity*, 1979:2, pp. 447-61. The EOE effect refers to the observation that productivity growth tends to decline at the end of the expansion phase of the business cycle. However, reestimating equation 5 with aggregate data after Gordon's EOE effect variable was included produced no significant changes from the results reported in the text. In particular, the estimated coefficient on DPC remained statistically significant, and the estimated coefficient on D74 remained statistically insignificant. Thus, a major result of this section — that after accounting for price controls, trend productivity growth did not decline further after 1973 — appears to be robust.

[23] Because the hypothesis of a further slowdown in trend productivity growth after 1973 was rejected, the estimates of trend productivity growth were calculated after dropping D74 from equation 5 and reestimating the model. The estimates of trend productivity growth are the only results reported because the results from these regressions were so similar to those reported in the Appendix.

TABLE 3
Results from estimates of equation 5

	(A) Coefficient Estimates[1]		(B) Estimated Trend Productivity Growth[2]	
	D74[3]	DPC[4]	1948-68	1968-81[5]
Private Business Sector	0.00048	0.0422[6]	3.3	1.8
	(0.24)	(3.16)		
Service-Producing	-0.00043	0.0460[7]	3.0	1.5
	(0.18)	(2.77)		
Goods-Producing	0.0015	0.0397[8]	3.6	2.1
	(0.47)	(1.80)		
Manufacturing	0.0028	0.0257	2.9	2.8
	(0.85)	(1.23)		
Farming	-0.0024	0.0124	6.0	6.3
	(0.17)	(0.13)		

NOTE: Initial ordinary least squares estimates of equation 5 indicated the presence of first-order autocorrelation. The Prais-Winsten two-step procedure was used in estimating equation 5 to correct for the presence of a first-order autoregressive error term.
[1]t-statistics are in parentheses.
[2]Units are percent per year continuously compounded.
[3]1974 Dummy: D74 = 0 from 1948:III to 1973:IV and -1 thereafter.
[4]Price Control Dummy: DPC = 0 from 1948:III to 1971:II, +1/7 from 1971:III to 1973:I, -1/7 from 1973:II to 1974:IV, and 0 thereafter.
[5]Because the growth rate of trend labor productivity from 1968 to 1981 is constant, it equals the trend growth rate from 1968 to 1979. Therefore, the numbers in this column can be compared directly with the trend productivity growth rates for the period from 1968 to 1979 in Table 1.
[6]Significant at 0.1 percent level.
[7]Significant at 1 percent level.
[8]Significant at 5 percent level.

fourth quarter of 1968 is the estimate of the constant term, α_0. From then on, the estimate of trend productivity growth is the difference between the constant term and the coefficient on D69, $\alpha_0 - \alpha_1$.

The evidence indicates that both the service-producing and goods-producing sectors contributed to the slowdown in aggregate trend productivity growth after 1968. However, there is no evidence of a slowdown in trend productivity growth after 1968 in either of the two goods-producing sectors for which the model was estimated. Indeed, trend productivity growth in the farming sector seems to have increased slightly.[24]

The combined results from both panels of Table 3 tell a very different story about postwar trend productivity growth from those in Table 1.

[24] Although the point estimate of trend productivity growth in the farming sector after 1968 is greater than the point estimate before 1969, the two estimates are insignificantly different in a statistical sense.

According to the traditional measures in Table 1, trend productivity growth began to slow in 1969 and slowed further after 1973 in all sectors. In contrast, the empirical estimates of the model in panel A of Table 3 show that trend productivity growth did not slow further in any sector after 1973. Moreover, the estimates in panel B indicate that aggregate trend productivity growth after 1973 is 1.8 percent a year, more than twice the traditional estimate of 0.8 percent. Although the estimates of trend productivity growth in panel B indicate that trend productivity growth declined in the goods-producing sector after 1968, neither the farming nor the manufacturing sectors contributed to the slowdown. Thus, the entire slowdown in the goods-producing sector resulted from a slowdown in mining and construction, the only other industries in the goods-producing sector. More important, though, the overall results indicate that trend productivity growth did not slow further after 1973 in any sector. Thus, traditional estimates of trend productivity growth such as those in Table 1 give a misleading impression of the source of slow productivity growth since 1973.

Summary and conclusions

This article presents evidence that the apparent second slowdown in trend productivity growth beginning in 1974 is an illusion resulting from the wage and price control program of the early 1970s. An empirical analysis of postwar labor productivity shows that after accounting for the 1969 shift in trend productivity growth and the cyclical and price control effects on measured productivity, there was no further slowdown in trend productivity growth after 1973. Because the results were consistent across major sectors, the findings strengthen the case for the price control explanation, which had previously been tested only with aggregate productivity data.

One implication of these findings is that any policy designed to reverse the "slowdown" in trend productivity growth that appeared to begin in 1974 should be reevaluated. Although a slowdown in trend productivity growth began in the middle to late 1960s, there is no evidence of a second slowdown in the aggregate private business sector or any of its subsectors that were studied. Productivity growth was slow from 1979 to 1983, but this was due to the cyclical behavior of the economy. Viewed in this way, slow productivity growth is more the result of sluggish economic growth in recent years than the cause. As a result, macroeconomic policies designed to return the U.S. economy to balanced, noninflationary economic growth may well cause productivity growth to return to more normal rates in the years ahead.

A further implication of these findings is that for future research on trend productivity growth to be useful for policy analysis, efforts should be made to understand why productivity growth has slowed over the *entire* period from the late 1960s to the present. By failing to account for the effects of price controls, previous research—and therefore the resulting policy prescriptions — has incorrectly focused on the decline in productivity growth that appeared to begin in 1974. Instead, attention should be given to the entire period from the late 1960s to the present. To the extent that this slow productivity growth is found to result from such factors as regulations and tax laws, structural policies designed to increase productivity growth might be appropriate. On the other hand, if it is found that slow productivity growth is due to factors not amenable to policy actions, such as demographics, structural policies designed to boost productivity growth would be unnecessary.

APPENDIX

Estimates of equation 5

Coefficient Estimate[1]	Private Business Sector Total	Service-Producing	Goods-Producing Total	Manufacturing	Farming
Constant	0.0083 (10.79)	0.0074 (5.74)	0.0090 (5.18)	0.0072 (5.66)	0.0149 (2.66)
D69[2]	0.0035 (2.05)	0.0041 (1.79)	0.0026 (0.86)	-0.0016 (0.56)	0.0007 (0.06)
D74[3]	0.00048 (0.24)	-0.00043 (0.18)	0.0015 (0.47)	0.0028 (0.85)	-0.0024 (0.17)
DPC[4]	0.0422 (3.16)	0.0460 (2.77)	0.0397 (1.80)	0.0257 (1.23)	0.0124 (0.13)
U	-0.0096 (2.89)	-0.0088 (1.36)	-0.0080 (0.95)	-0.0112 (2.65)	-0.0112 (0.44)
E	-0.0055 (0.05)	-0.0848 (0.44)	0.3874 (1.57)	0.3177 (2.69)	-0.5316 (0.63)
E_{-1}	-0.2778 (3.45)	-0.1855 (1.37)	-0.6879 (3.94)	-0.5195 (5.53)	-0.3421 (0.52)
LR	-0.0029 (0.90)	0.0015 (0.27)	-0.0015 (0.21)	-0.0042 (1.18)	0.0058 (0.22)
ρ[5]	-0.1766 (2.08)	-0.2304 (2.31)	-0.1937 (1.92)	0.1562 (1.83)	-0.3270 (4.01)
Summary Statistics					
R^2	0.41	0.21	0.40	0.50	0.03
Degrees of Freedom	125	86	86	125	125
Root-Mean-Square Error	0.0076	0.0097	0.0126	0.0094	0.0620

NOTE: Initial ordinary least squares estimates of equation 5 indicated the presence of first-order autocorrelation. The Prais-Winsten two-step procedure was used in estimating equation 5 to correct for the presence of a first-order autoregressive error term. U is the change in the civilian unemployment rate, E and E_{-1} are the current and lagged growth rate of employment in manufacturing, mining, and construction, and LR is the layoff rate.

[1] t-statistics are in parentheses.
[2] 1969 Dummy: D69 = 0 from 1948:III to 1968:IV and -1 thereafter.
[3] 1974 Dummy: D74 = 0 from 1948:III to 1973:IV and -1 thereafter.
[4] Price Control Dummy: DPC = 0 from 1948:III to 1971:II, + 1/7 from 1971:III to 1973:I, -1/7 from 1973:II to 1974:IV, and 0 thereafter.
[5] First-order autoregressive parameter.

Theories of Price Determination

By George A. Kahn

The rising inflation rates of the 1970s and the falling inflation rates of the early 1980s have provided a challenging laboratory for testing and designing theories of price determination. While the laboratory has produced widespread agreement on the long-run determinants of the aggregate price level, disagreement over theories of the short run still persists. These differences have led to a variety of policy prescriptions designed to reduce inflation while maintaining an "acceptable" level of production and employment. Policy differences result, despite a common model of long-run price behavior, because of different assumptions about short-run aggregate supply and demand.

While no theory of price determination can claim to explain recent price behavior completely, Keynesian, monetarist, and new classical theories have emerged as the leading contenders.[1] These theories fall into two categories based on their policy recommendations. Keynesians or, more generally, policy activists recommend that policy respond to current or expected economic conditions. By "leaning against the wind" or accommodating adverse supply shocks, policy activists believe they can reduce fluctuations in real output. Monetarists and new classical economists, on the other hand, recommend a policy that is invariant to the current state of the economy. By adhering to simple policy rules, monetarists and new classical economists maintain that uncertainty will be reduced and economic performance improved. To understand how these diverse policy prescriptions arise and how they relate to underlying theories of price determination requires a closer examination of the prescribing theories.

This article uses an aggregate supply and demand analysis to illustrate common features of models of price determination and to highlight important differences. The first section

George A. Kahn is an economist with the Economic Research Department of the Federal Reserve Bank of Kansas City. Craig Hakkio, William Keeton, Raymond Lombra, Gordon Sellon, and Stuart Weiner provided helpful discussions.

[1] While most readers are probably familiar with the Keynesian and monetarist schools, they may not be familiar with the new classical school. The principal features of the new classical economics are perfectly competitive markets and rational expectations Other names that have been given to this school of thought include rational expectations, monetarism Mark II, and, in tribute to its main authors, the Lucas-Sargent-Wallace approach. The somewhat stylized distinctions made in this article between theories of price determination are meant to reflect the general views of leading economists. Prominent representatives of the Keynesian school include Franco Modigliani, Paul Samuelson, and James Tobin. Representatives of the monetarist school include Karl Brunner, Milton Friedman, Allan Meltzer, and the staff of the Federal Reserve Bank of St. Louis. Representatives of the new classical school are Robert Lucas, Thomas Sargent, and Neil Wallace.

presents the supply and demand framework. It describes the generally accepted downward sloping shape of the aggregate demand curve and the vertical shape of the long-run aggregate supply curve. It also establishes the role of price expectations in the derivation of short-run aggregate supply. How quickly prices and price expectations adjust to changes in demand is shown to be an important point of disagreement among theories of price determination. The second section focuses on other factors affecting aggregate supply and demand and discusses the importance placed on each factor by different theories of price determination. It also describes the different policy recommendations associated with proponents of Keynesian, monetarist, and new classical schools of thought.

Prices, price expectations, and output

A useful framework for studying price determination is the aggregate supply and demand model. This model divides the macroeconomy into a demand side and a supply side. The demand side represents the spending of consumers, firms, and the government. The supply side represents the production of goods and services by firms using labor and other inputs. Together, demand and supply determine the level of prices and real output.[2]

Long-run theory of price determination

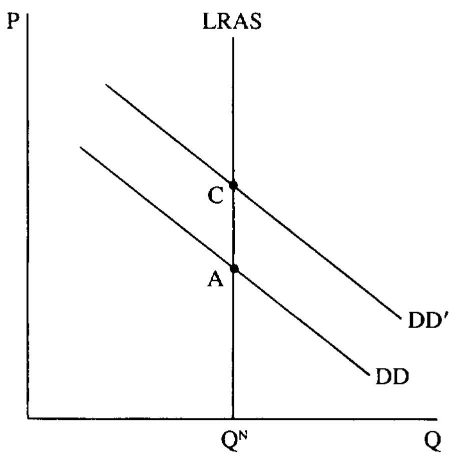
In the long run, prices are determined by the interaction of aggregate demand and long-run aggregate supply. Two common features of most theories of price determination are a downward sloping demand curve and a vertical long-run supply curve. Aggregate demand is an inverse relationship between the price level and the level of real output. It represents combinations of prices and real output that clear both the product market and the money market under specific behavioral relationships for private spending and given levels of the money supply and fiscal policy variables. A higher price level, other things equal, requires a higher interest rate to keep the demand for money equal to the supply of money. The higher interest rate, in turn, reduces interest-sensitive spending and causes real output to decline. Because a higher price level is associated with a lower level of real output, the demand curve slopes downward.

The other common feature of models of price determination is the shape and position of the long-run aggregate supply curve. This curve indicates the amount produced at each price level when all resources are fully employed and price expectations are correct. The level of output determined by the long-run supply curve is called the natural rate. The size of the labor force, the structure of the labor market, the state of technology, and other factors not generally explained in simple macroeconomic models determine the level of real output that corresponds to the natural rate. Because it is generally agreed that this level of output does not depend on the price level, the long-run aggregate supply curve is drawn as a vertical line. In fact, some economists prefer to define the position of the long-run aggregate supply curve as the level of output at which there is no tendency for inflation to rise or fall.[3]

Figure 1 shows two possible aggregate demand curves, DD and DD′, and one possible long-run aggregate supply curve, LRAS.

[2] While this article analyzes theories of price level determination, many of the results extend to theories of inflation determination. Because inflation can be represented by a continual upward shifting of the supply and demand curves described in this article, growth in the variables that influence the price level can be thought of as determining the inflation rate.

FIGURE 1



The price level, P, is on the vertical axis, and real output, Q, is on the horizontal axis. If DD represents the initial level of aggregate demand, then DD′ represents an increase in aggregate demand. At each possible price level, DD′ implies greater real output, and at each level of real output, DD′ implies a price level higher than DD. The greater aggregate demand thus associated with DD′ could be the result of a more stimulative fiscal or monetary policy or greater consumer or business confidence. Figure 1 also shows a long-run aggregate supply curve. Reflecting the natural-rate hypothesis, LRAS is drawn as a vertical line at the natural rate of output, Q^N. The slope of the long-run aggregate supply curve indicates that there is no long-run relationship between the price level and output. Any price level is consistent with the natural rate of output.

The long-run price level is determined by the intersection of DD and LRAS, as in Figure 1. When aggregate demand increases as a result of, say, an increase in the money supply, the demand curve shifts up and to the right to DD′ along the fixed LRAS curve. In Figure 1, the economy moves from point A to point C. The long-run result of the increase in aggregate demand is an increase in the price level, with no change in real output. In fact, anything that shifts aggregate demand will affect only prices in the long run. Output will be left equal to its natural rate. It is impossible, therefore, in most modern theories of price determination for government policies affecting aggregate demand to have any lasting effects on real output.[4]

Theories of short-run aggregate supply

The short-run aggregate supply curve shows how firms alter production and employment for given price expectations in response to changes in the price level. Because production and employment decisions depend on the price level and must often be made before the price level is known, expectations about prices play an important role in short-run aggregate supply determination. Any deviation between the actual and expected price level is an unantici-

[3] More technically, the natural rate of output is defined as the amount of real GNP produced when the economy is operating at the natural rate of unemployment. The natural rate of unemployment, in turn, is defined as the unemployment rate consistent with a stable inflation rate and no supply shocks. The natural rate of unemployment is not necessarily the optimal rate of unemployment. Reducing the natural rate requires microeconomic policies that increase the efficiency of labor markets. For a discussion of "structural" unemployment and ways to reduce it, see Stuart Weiner, "Enterprise Zones as a Means of Reducing Structural Unemployment," *Economic Review*, Federal Reserve Bank of Kansas City, March 1984, pp. 3-16.

[4] Strictly speaking, macroeconomic policies can affect the natural rate by influencing the allocation of income between investment and consumption. If increases in government spending on consumption goods "crowds out" private investment spending, the future capital stock will fall. A reduction in the capital stock will then reduce production possibilities.

pated price change that causes economic agents to reevaluate past decisions. In the long run, as agents gain information about prices and incorporate that information into their decisions, actual and expected prices are the same. Only when expectations turn out to be correct is the economy operating on its long-run aggregate supply curve at the natural rate of output. At any other level of output, actual prices deviate from expected prices, output deviates from its natural rate, and economic agents are led to revise their inaccurate price expectations.

Two explanations are widely given to explain why deviations of the actual price level from its expected value cause the aggregate supply curve to slope upward. One explanation focuses on the product market. If prices turn out to be greater than expected, producers will attribute at least some of the deviation of actual from expected prices to an increase in the relative price of their particular product.[5] As a result, firms will increase employment and production, and real output will rise. If, on the other hand, prices turn out to be lower than expected, producers will attribute part of the deviation to a decline in the relative price of their product, and real output will fall. Therefore, for given price expectations, a higher price level leads to a higher level of real output.

The other explanation focuses on the labor market. If wages are set before product prices and based on the expected price level, then an unexpectedly high price level will lead to an unexpectedly low real wage rate (the nominal wage rate divided by the price level). Because labor costs are lower, firms will want to hire more labor and increase production.[6] Similarly, a lower than expected price level leads to a higher than expected real wage and a reduction in employment and real output. Thus, for given price expectations, a higher actual price level is associated with a higher level of real output. The short-run aggregate supply curve slopes upward.

Figure 2 illustrates the effect of an increase in aggregate demand under an upward sloping short-run aggregate supply curve. When DD shifts to DD′, the economy moves initially from A to B. Both real output and the price level increase, provided price expectations remain unchanged. In the long run, as prices and expectations adjust to the higher level of aggregate demand, the economy moves back along DD′ to point C. In other words, the

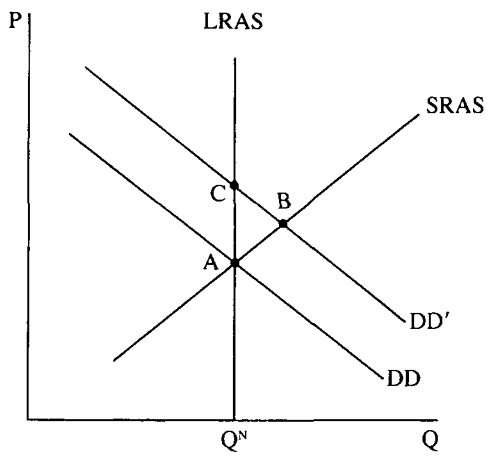
---

[5] Robert Lucas uses the confusion between local and aggregate price shocks to derive an upward sloping supply curve in "Expectations and the Neutrality of Money," *Journal of Economic Theory*, Vol. 4, April 1972, pp. 103-24. For a more accessible version of the same model, see Robert Lucas, "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, Vol. 63, June 1973, pp. 326-34.

[6] This is not the complete story, however, since workers must be induced to supply the additional labor at a lower real wage rate. Suppose workers evaluate the purchasing power of their nominal wage in terms of expectations about average consumer prices and producers evaluate their real labor costs in terms of the actual price received for products produced. As long as workers' expectations are slow to adjust and lag behind reality, a higher price level can lead to increased employment and production. Workers' expectations may be slow to adjust if workers sample prices only during infrequent shopping trips. Milton Friedman proposed this explanation for the upward sloping short-run aggregate supply curve in "The Role of Monetary Policy," *American Economic Review*, Vol. 58, March 1968, pp. 1-17. He also coined the phrase "natural rate" in that article. About the same time Edmund Phelps also described the natural rate hypothesis in "Phillips Curves, Expectations of Inflation, and the Optimal Unemployment Over Time," *Economica* (NS), Vol. 34, August 1967, pp. 254-81.

Another way to induce workers to supply more labor at a lower real wage would be to have them sign wage contracts at the beginning of the period. These contracts specify the nominal wage rate and require workers to supply whatever labor is demanded by firms. When prices go up, real wages go down. Firms demand and workers supply more labor. For a discussion of why workers and firms would find it mutually advantageous to sign long-term contracts, see Arthur Okun, "Inflation: Its Mechanics and Welfare Costs," *Brookings Papers on Economic Activity*, 1975:2, pp. 351-90.

FIGURE 2



short-run aggregate supply curve shifts up as expectations change. Consequently, there is no long-run output gain associated with the increase in aggregate demand. The economy has a higher price level at the natural rate of output. Old-fashioned aggregate price theories that predicted a long-run price-output tradeoff and a permanent shift from point A to point B as a result of an increase in aggregate demand are no longer credible theories of price determination.[7] In their place are natural-rate theories that posit only short-run tradeoffs between prices and output.

Keynesians, monetarists, and new classical economists agree that unanticipated price changes cause changes in real output. They disagree, however, about the speed of adjustment of prices and expectations. Keynesians and monetarists claim that shifts in aggregate demand affect real output for several periods. This is because they believe prices and expectations adjust slowly. Price and wage rigidities introduced by such institutions as labor market contracts and government regulations cause prices to respond gradually to changes in aggregate demand. If, for example, three-year staggered wage contracts are prevalent, as they are in the U.S. economy, the adjustment of wages to price changes can take up to three years. Specifically, if nominal wages are determined by long-term contracts between workers and firms and if workers agree to supply the labor demanded at the contract wage, then price increases are associated with lower real wages and greater output and employment.[8] The resulting gain in real output will last until all existing contracts have been rewritten to incorporate a higher price level. In the long run, when all contracts have been renegotiated, nominal wages will fully reflect the higher price level. Real wages and real output will return to their initial levels, but prices will remain higher. Thus, the long-run aggregate supply curve is vertical.

New classical models of price determination combine perfect price flexibility with rational expectations to arrive at a strikingly different set of conclusions. In these models, prices and wages are determined by auction markets, and price expectations are based on all available information and an accurate understanding of

[7] The inverse statistical relationship between inflation (rather than the price level) and unemployment (rather than output) is called the Phillips curve. It was first described by Irving Fisher in "A Statistical Relation Between Unemployment and Price Changes," *International Labour Review*, Vol. 13, June 1926, pp. 785-92. A. W. Phillips "discovered" and popularized a similar inverse relationship between wage growth and unemployment in "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861-1957," *Economica*, Vol. 25, November 1958, pp. 283-99. Note that the supply and demand model derived in the text actually implies a short-run price *level*-output tradeoff, not a Phillips curve relationship.

[8] For a labor market-based derivation of an upward sloping short-run aggregate supply curve, see George Kahn, "Wage Behavior in the United States: 1907-80," *Economic Review*, Federal Reserve Bank of Kansas City, April 1983, pp. 16-26.

the economy. The assumed flexibility of prices and wages implies that perceived movements in aggregate demand lead to proportionate, contemporaneous movements in the price level. In terms of Figure 2, a shift in aggregate demand from DD to DD′ due to, say, an increase in the money supply, causes a movement along LRAS from A directly to C as long as the increase in demand is fully anticipated. In other words, SRAS shifts upward and to the left at the same time as the shift in DD, so that the economy comes to rest immediately at the long-run equilibrium point C. In effect, the SRAS curve is vertical.

The distinction between anticipated and unanticipated changes in the new classical models of aggregate demand is important. If anticipated, a change in demand will be perceived as economywide and will result in agents immediately adjusting prices and wages so that all relative prices (real wages, for example) remain the same. If unanticipated, agents will think at least part of the increase in demand is unique to their market and will increase production and employment in order to exploit a perceived temporary increase in the relative price of their product.[9] Once agents figure out that the increase in aggregate demand affected all prices proportionately so that relative prices actually remained the same, they will reduce production to its previous level. Thus, in Figure 2, the effect of an unanticipated increase in aggregate demand is a movement from point A to point B to point C. In contrast to Keynesian and many monetarist models, the associated output gain lasts only one period. This lack of persistence results from the new classical assumptions of price flexibility, rational expectations, and a one-period information lag.[10]

Both rational expectations and perfect price flexibility are crucial to the new classical model. The assumption of rational expectations—that people form expectations on the basis of all information available and an accurate model of the economy—has had wide influence on macroeconomic theory. It is not, however, the only assumption necessary for the new classical result that only unanticipated demand shocks affect real output. Not only must expectations be rational, but prices must be flexible, as they would be in a world where all markets worked like auctions. Because models closely resembling Keynesian theories can be constructed under the assumption of rational expectations,[11] the primary issue is not whether expectations are formed rationally but how quickly prices adjust to changes in aggregate demand.

The degree of price and wage flexibility has become an important area of ongoing research, at least among members of the non-

[9] The terms "anticipated" and "unanticipated" usually refer to expectations formed in the previous period. Thus in the Keynesian model (and many monetarist models) an increase in aggregate demand that is fully anticipated as of the previous period can have real effects because some agents are locked into contracts signed in even earlier periods.

[10] Early models from the new classical economics were criticized for their inability to explain the persistence of output fluctuations (Milton Friedman's 1968 model in which workers are "fooled" about the real wage is also subject to this criticism.) Subsequently, these models have been modified to include lags in the transmission of information and accelerator effects. See, for example, Robert Lucas, "An Equilibrium Model of the Business Cycle," *Journal of Political Economy*, Vol. 83, December 1975, pp. 1113-44.

[11] See, for example, Stanley Fischer, "Long-term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, Vol. 85, February 1977, pp. 191-206; John Taylor, "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, Vol. 88, February 1980, pp 1-23; and Edmund Phelps and John Taylor, "Stabilizing Powers of Monetary Policy Under Rational Expectations," *Journal of Political Economy*, Vol. 85, February 1977, pp. 163-90. These theories differ from traditional Keynesian theories in that they imply the optimality of policy rules. They differ from new classical theories in that they assume price stickiness and imply the superiority of rules with feedback over constant growth rate rules.

monetarist, policy activist school. Monetarists have shown surprisingly little interest in research on the short-run dynamics of wage and price behavior. Their lack of interest could reflect the greater weight they place on the long run or their belief that short-run fluctuations cannot be explained. Nevertheless, preliminary evidence on short-run price and wage responsiveness suggests a wide variation of flexibility across time and place.[12] Flexible price models seem appropriate in some instances. In other instances, sticky price models dominate. In the postwar U.S. economy, overlapping three-year wage contracts seem to imply that the sticky price model is more appropriate. Thus, for the United States, it is difficult to accept the price flexibility assumption explicit in the new classical proposition that only unanticipated demand shocks affect real output.

Scope for demand management policies

Proponents of the three major theories of price determination differ considerably on the causes of shifts in aggregate demand and supply. Keynesians or, more generally, policy activists stress the instability of the private sector and the need for countercyclical policy. Monetarists claim that the private sector is basically stable and focus on the destabilizing effects of historical demand management policies. New classical theorists stress the distinction between anticipated and unanticipated policy. While both monetarists and new classical economists recommend nonactivist policies, they do so for somewhat different reasons. An examination of factors thought to shift aggregate demand and supply highlights the differing assumptions underlying alternative policy recommendations.

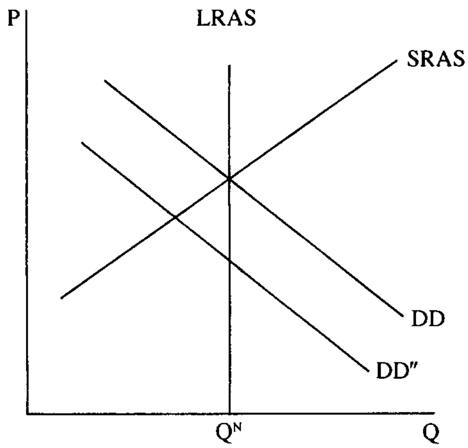
Factors shifting aggregate demand and recommended policy responses

Any factor that causes the aggregate demand curve to shift causes prices, and possibly output, to fluctuate. Increases in government spending, the money supply, net exports, or consumer or business confidence, and decreases in taxes or money demand are among the factors that cause the demand curve to shift to the right. Changes in these variables in the opposite direction shift the aggregate demand curve to the left. Keynesians, monetarists, and new classical economists have differing views on the importance of these factors.

Keynesians believe that shifts in aggregate demand result from the inherent instability of the private sector as well as from policy changes. They think, for example, that money demand can shift as a result of factors other than changes in real output and interest rates. If money demand increases and money supply remains unchanged, the aggregate demand curve will shift down and to the left. Figure 3 shows the price and output effects of a decline in aggregate demand from its initial level, DD, to a lower level, DD″. Under the assumption of an upward sloping short-run aggregate supply curve, real output and prices will fall. Policy activists believe that to keep output and prices from falling, the Federal Reserve should accommodate the increase in money demand by increasing the money supply. Increasing the money supply will shift aggregate demand back up and to the right, keeping the price level stable and output equal to its natural rate. In general, Keynesians believe that by manipulating monetary and fis-

[12] See, for example, George Kahn, "Wage Behavior in the United States: 1907-80," and Jeffrey Sachs, "Wages, Profits, and Macroeconomic Adjustment: A Comparative Study," *Brookings Papers on Economic Activity*, 1979.2, pp. 269-332.

FIGURE 3



cal policy, the impact of both inflationary and recessionary shocks to aggregate demand can be at least partially offset. Through the use of these policy tools, output can be kept closer to its natural rate.

Monetarists believe that aggregate demand would be fairly stable but for perverse policy actions. They believe, for instance, that money demand is a stable function of interest rates and real income. Fluctuations in aggregate demand and the associated short-run deviations of real output from the natural rate result largely from changes in the money supply. Consequently, monetarists recommend stable money supply rules, such as a constant growth rate, to minimize fluctuations in prices and real output.

Furthermore, because monetarists believe discretionary policy affects output and prices with long and variable lags, they do not generally believe the Federal Reserve should react to exogenous shocks. The monetarists' focus on the long run gives further support to their reluctance to lean against the wind. They claim that knowledge of short-run behavior is not sufficient to permit discretionary policy and that knowledge of long-run behavior implies the optimality of fixed money growth rate rules. These results follow even though monetarists believe the short-run supply curve is upward sloping with respect to anticipated demand shifts.

New classical economists, who believe in flexible prices and rational expectations, conclude that any shock to aggregate demand that is fully anticipated will affect only prices and not real output. Unanticipated shocks to aggregate demand will have the same short-run effects in the new classical model as demand shocks in the monetarist or Keynesian models.

The new classical model calls for stable, predictable policy. This recommendation follows from the belief that private agents will "see through" any effort by policymakers to cause real output to deviate systematically from its natural rate. The best policymakers can do is minimize random fluctuations in policy variables so as to minimize unexpected aggregate demand fluctuations. The way to do this, according to the new classical model, is with policy rules. While one rule is as good as another, new classical economists prefer simple rules, such as constant money growth.

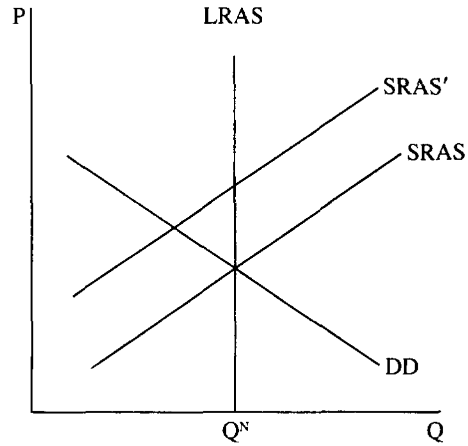
An important implication of the anticipated-unanticipated demand shock dichotomy for policymakers is the need for credibility when changing policy. In a move from an inflationary policy to a disinflationary policy, for example, credibility is important if real output is not to be sacrificed. If a policy change is announced and carried out by the Federal Reserve but not believed by private agents, the effect will be the same as an unanticipated demand shock. In a move to disinflationary policy, aggregate demand will shift down and to the left along a short-run supply curve that is upward sloping with respect to the unexpected shift.[13] Inflation will decline, but so will real output. As agents come to expect the new disinflationary policy, the short-run supply curve effectively becomes steeper, and further reductions in inflation can be made less costly.

Factors shifting aggregate supply and recommended policy responses

In addition to changes in price expectations, supply shocks shift the aggregate supply curve. A supply shock is any event that changes the output firms are willing to produce at a given price level.[14] Adverse supply shocks shift the short-run aggregate supply curve up and to the left. In Figure 4, this effect is shown as a shift from SRAS to SRAS′. Examples of adverse supply shocks are a sudden large increase in the price of energy such as those imposed by OPEC in 1973-74 and 1979-80, an increase in the price of food resulting from a crop failure, a spontaneous demand by workers for higher wages, and a worsening in the terms of international trade.[15] For a given demand curve, these adverse supply shocks raise prices and lower real output, as in Figure 4. The effect on output can be temporary or permanent, depending on the nature of the supply shock. If the shock permanently reduces production possibilities, the long-run aggregate supply curve shifts to the left.

Keynesians or policy activists are distinguished not only by their belief that policy should lean against the wind but also by their belief that policy should accommodate adverse supply shocks. If, for example, an oil shock shifts the short-run aggregate supply curve up and to the left, policy activists believe that the money supply should be increased in order to reduce the fluctuation of real output from its natural rate. Increasing the money supply shifts the aggregate demand curve to the right

FIGURE 4



---

[13] Here, the results derived for the price level are assumed to extend to *changes* in the price level.

[14] See Robert Gordon, *Macroeconomics*, Third Edition, Little-Brown, Boston, 1984, or Glenn H. Miller, Jr., "Inflation and Recession, 1979-82: Supply Shocks and Economic Policy," *Economic Review*, Federal Reserve Bank of Kansas City, June 1983, pp. 8-21, for more extensive, yet accessible discussions of supply shocks and inflation.

[15] The terms of trade worsen when, other things constant, the nominal exchange rate increases or the foreign price level increases relative to the domestic price level. If the terms of trade worsen as a result of, say, an increase in the price of imports relative to the price of exports, the short-run aggregate supply curve will shift up and to the left. This assumes that producers face increased costs of production or try to increase profit margins. Given a downward sloping demand curve, output will fall and prices will rise. Under the same worsening in the terms of trade, however, the aggregate demand curve will shift up and to the right if aggregate spending shifts away from foreign goods toward domestic goods. This shift in aggregate demand tends to increase the price level and real output given an upward sloping short-run aggregate supply curve. The net effect of a deterioration in the terms of trade after allowing for shifts in both supply and demand will be an increase in the price level and an ambiguous change in real output.

and raises the price level above what it would have been without a policy response. An alternative policy response would reduce the money supply to extinguish the tendency for the supply shock to raise prices. In this case, the aggregate demand curve would shift to the left, and output would fall below what it would have been had no policy response been made.

Monetarists and new classical economists agree that policy should not respond to supply shocks, but for different reasons. Monetarists believe that while supply shocks may reduce real output, any policy response is likely to do more harm than good. Long and variable lags in the transmission of policy changes to the economy are likely to influence output at the wrong time and create additional problems. Thus, monetarists believe a constant growth of the money supply should be maintained even in the face of supply shocks. New classical economists, on the other hand, argue that supply shocks affect only relative prices and evoke optimal responses from the free market. Policy intervention, if anticipated, will only exacerbate the price-level effects of supply shocks. Thus new classical economists also argue for an unchanging policy in the face of supply shocks.

Summary and conclusions

Monetarists and Keynesians basically share an underlying model of price determination. They differ, however, regarding the stability of aggregate demand, the emphasis placed on the short run, and the desirability of discretionary policy. Monetarists and new classical theorists share a common philosophy that the economy will operate most efficiently if left on its own under stable, predictable policy rules. They differ, however, regarding predictions about the effects of anticipated policy. These issues reflect not merely different assumptions about economic behavior but, in fact, different approaches to the study of macroeconomics. While some of these issues are empirical questions, it is doubtful that all the issues can be resolved by data analysis alone.

Selected Readings

Asterisks mark relatively nontechnical readings that are accessible to a wide audience. Other readings represent important contributions to the economics of price or inflation determination. They vary in difficulty.

Arrow, Kenneth J., "Toward a Theory of Price Adjustment," in Moses Abramowitz, *The Allocation of Economic Resources*, Stanford: Stanford University Press, 1959, pp. 41-51.

*Barth, James R., "The Costs of Slowing Inflation: Four Views," *Economic Review*, Federal Reserve Bank of Atlanta, January 1982, pp. 39-49.

Barro, Robert J., "A Theory of Monopolistic Price Adjustment," *Review of Economic Studies*, 39, January 1972, pp. 17-26.

________, "Rational Expectations and the Role of Monetary Policy," *Journal of Monetary Economics*, 2, January 1976, pp. 1-32.

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

________, "Second Thoughts on Keynesian Economics," *American Economic Review*, 69, May 1979, pp. 54-9

*Branson, William, *Macroeconomic Theory and Policy*, New York. Harper and Row, 1972.

Bronfenbrenner, Martin, and Frederick D. Holzman, "Survey of Inflation Theory," *American Economic Review*, 53, September 1963, pp. 593- 661.

Brunner, Karl, and Allan H. Meltzer, "Institutional Arrangements and the Inflation Problem," *Journal of Monetary Economics*, 3, Supplementary Series, 1976

________, eds., *The Problem of Inflation*, Carnegie-Rochester Conference Series on Public Policy, 8, 1978.

Drazen, Allan, "Recent Developments in Macroeconomic Disequilibrium Theory," *Econometrica*, 48, March 1980, pp. 283-306.

Eckstein, Otto, ed., *The Econometrics of Price Determination Conference*, Washington: Federal Reserve Board, 1972.

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

Fisher, Irving, "A Statistical Relation Between Unemployment and Price Changes," *International Labour Review*, 13, June 1926, pp. 785-92.

*Friedman, Milton, *Inflation: Causes and Consequences*, New Delhi: Asia Publishing House, 1963.

*________, "Nobel Lecture: Inflation and Unemployment," *Journal of Political Economy*, 85, June 1977, pp. 451-72.

*________, "The Role of Monetary Policy," *American Economic Review*, 58, March 1968, pp. 1-17

Goldfeld, Stephen, "Rules, Discretion, and Reality," *American Economic Review*, 72, May 1982, pp. 361-66.

Gordon, Robert J., "The Demand for and Supply of Inflation," *Journal of Law and Economics*, 18, December 1975, pp. 807-36.

*________, *Macroeconomics*, Third Edition, Boston: Little-Brown, 1984.

________, "Output Fluctuations and Gradual Price Adjustment," *Journal of Economic Literature*, 19, June 1981, pp. 493-530.

________, "Recent Developments in the Theory of Inflation and Unemployment," *Journal of Monetary Economics*, 22, April 1976, pp. 185-219.

________, "The Theory of Domestic Inflation," *American Economic Review*, 67, February 1977, pp. 128-34.

Hall, Robert E., "The Rigidity of Wages and the Persistence of Unemployment," *Brookings Papers on Economic Activity*, 1975:2, pp. 301-35.

*Heller, Walter W., "Shadow and Substance in Inflation Policy," *Challenge*, 23, January/February 1981, pp. 5-13.

*Kahn, George A., "International Differences in Wage Behavior: Real, Nominal, or Exaggerated," *American Economic Review*, May 1984, forthcoming

*________, "Wage Behavior in the United States: 1907-80," *Economic Review*, Federal Reserve Bank of Kansas City, April 1983, pp. 16-26

Keynes, John Maynard, *The General Theory of Employment, Interest, and Money*, New York. Harcourt Brace, 1936.

Laidler, David, "Monetarism: An Interpretation and an Assessment," *Economic Journal*, 91, March 1981, pp 1-29.

________, and Parkin, Michael, "Inflation: A Survey," *Economic Journal*, 85, December 1975, pp. 741-809.

Lucas, Robert E., Jr., "An Equilibrium Model of the Business Cycle," *Journal of Political Economy*, 83, December 1975, pp. 1113-44.

________, "Expectations and the Neutrality of Money," *Journal of Economic Theory*, 4, April 1972, pp. 103-24.

________, "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, 63, June 1973, pp. 326-34

________, and Rapping, Leonard A., "Real Wages, Employment, and Inflation," *Journal of Political Economy*, 77, September/October 1969, pp. 721-54. An expanded version appears in Phelps (1970).

Mayer, Thomas, ed., *The Structure of Monetarism*, New York and London. Norton, 1978.

McCallum, Bennett T., "Price-Level Stickiness and the Feasibility of Monetary Stabilization Policy with Rational Expectations," *Journal of Political Economy*, 85, June 1977, pp. 627-34.

________, "Rational Expectations and Macroeconomic Stabilization Policy: An Overview," *Journal of Money, Credit, and Banking*, 12, November 1980, pp. 716-46.

*Miller, Glenn H., Jr., "Inflation and Recession, 1979-82: Supply Shocks and Economic Policy," *Economic Review*, Federal Reserve Bank of Kansas City, June 1983, pp. 8-21.

*________, "Slowdown in Economic Activity and the Rate of Inflation," *Economic Review*, Federal Reserve Bank of Kansas City, September/October 1981, pp. 18-27

Muth, John F , "Rational Expectations and the Theory of Price Movements," *Econometrica*, 29, July 1961, pp. 315-35.

Nordhaus, William D., "Recent Developments in Price Dynamics," in Eckstein, 1972.

Okun, Arthur M., "Inflation: Its Mechanics and Welfare Costs," *Brookings Papers on Economic Activity*, 1975.2, pp. 351-90.

*________, "The Invisible Handshake and the Inflationary Process," *Challenge*, 22, January/February 1980, pp. 5-12.

*________, *Prices and Quantities: A Macroeconomic Analysis*, Washington: Brookings, 1981.

Perry, George L., "Inflation in Theory and Practice," *Brookings Papers on Economic Activity*, 1980:1, pp. 207-41.

________, "Slowing the Wage-Price Spiral. The Macroeconomic View," *Brookings Papers on Economic Activity*, 1978.2, pp. 259-91.

*________, "Understanding World Inflation," *American Economic Review*, 65, May 1975, pp. 120-24

Phelps, Edmund S., "Phillips Curves, Expectations of Inflation, and the Optimal Unemployment Over Time," *Economica* (NS), 34, August 1967, pp. 254-81

________, et al., *Microeconomic Foundations of Employment and Inflation Theory*, New York: Norton, 1970.

________, and Taylor, John, "Stabilizing Powers of Monetary Policy Under Rational Expectations," *Journal of Political Economy*, 85, February 1977, pp 163-90.

Phillips, A. W., "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861-1957," *Economica*, 25, November 1958, pp. 283-99.

Sachs, Jeffrey, "Wages, Profits, and Macroeconomic Adjustment: A Comparative Study," *Brookings Papers on Economic Activity*, 1979:2, pp. 269-332.

Sargent, Thomas J., "A Note on the 'Accelerationist' Controversy," *Journal of Money, Credit, and Banking*, 3, August 1971, pp. 721-25.

________, and Wallace, Neil, "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *Journal of Political Economy*, 83, April 1975, pp. 241-57.

*Solow, Robert M., "The Intelligent Citizen's Guide to Inflation," *The Public Interest*, Winter 1975, pp. 30-66.

Stein, Jerome, ed., *Monetarist, Keynesian, and New Classical Economics*, Oxford: Blackwell, 1982.

Taylor, John, "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, 88, February 1980, pp. 1-23

Tobin, James, "Are New Classical Models Plausible Enough to Guide Policy?" *Journal of Money, Credit, and Banking*, 12, November 1980, pp. 788-99.

________, "Inflation and Unemployment," *American Economic Review*, 52, March 1972, pp 1-18

*Weiner, Stuart E., "Enterprise Zones as a Means of Reducing Structural Unemployment," *Economic Review*, Federal Reserve Bank of Kansas City, March 1984, pp. 3-16.

Deposit Insurance and the Deregulation of Deposit Rates

By William R. Keeton

The federal deposit insurance system has long been regarded as a total success. Since the establishment of federal deposit insurance in 1933, following the worst banking panic in the country's history, the number of bank failures has fallen dramatically. Though banks have continued to fail—because of fraud, unsound investments, or plain bad luck—the bank runs and banking panics that once plagued the financial system seem to have been eliminated.

In the past few years, however, doubts about the federal deposit insurance system have begun to emerge. The number of banks that have been closed or merged with healthier institutions has increased since the early 1970s, and for the first time failures have included some very large banks, such as Franklin National in 1974. In addition, questions have been raised about the quality of some loans made by larger banks in recent years, particularly to developing countries. Finally, as a result of the financial deregulation mandated by the Depository Institutions Deregulation and Monetary Control Act of 1980, the Garn-St. Germain Depository Institutions Act of 1982, and other legislation, there has been considerable concern that banks will take more risks than before and cause the incidence of bank failures to increase.

This article examines the effect on bank risk-taking of a particular aspect of deregulation, the removal of deposit-rate ceilings.[1] Even before deregulation, deposit insurance gave banks an incentive to take risk by shifting part of the costs of their risk-taking to the Federal Deposit Insurance Corporation. It is argued that the deregulation of deposit rates will increase this distortion in bank behavior, both by expanding the opportunities for risk-taking and by increasing the benefits to be derived.

The first section of the article provides an overview of the current federal deposit insurance system. The next section explains how the deposit insurance system distorted banks' behavior even before the recent financial deregulation. The third section considers how

William R. Keeton is a senior economist in the Economic Research Department at the Federal Reserve Bank of Kansas City.

[1] Although this article focuses on deposit-rate deregulation, it should be noted that the recent financial deregulation may also affect banks' risk-taking behavior by liberalizing their lending and investment powers.

the deregulation of deposit rates is likely to increase these distortions. Possible policy responses are discussed in the last section.

The current federal deposit insurance system

Federal deposit insurance was established by the Banking Act of 1933 in response to three years of widespread banking failures. Over that time, many people became fearful about the safety of their deposits and withdrew their funds to hold in the form of currency. In an effort to meet such withdrawals, banks called in loans and liquidated assets, often at substantial losses. Banks that could not satisfy withdrawals in this way were forced to close.

In the period from 1930 through 1932 about 5,100 banks failed, more than one out of five of the commercial banks in existence at the beginning of 1930.[2] These bank failures brought direct losses to depositors and bank shareholders and to businesses suddenly deprived of a long-standing source of credit. The massive bank failures of the period also were partly responsible for a severe decline in the nation's money supply. To protect themselves against sudden deposit withdrawals, banks that remained in operation held more of their assets in the form of idle reserves. Also, with the increased fear over the safety of bank deposits, the public held more of their liquid assets in the form of currency, which reduced the amount of reserves available to the banking system. Because of these two factors, the money supply fell almost 25 percent from 1929 to the end of 1932, even though the total monetary base—the sum of bank reserves and currency held by the public—did not decline at all.

[2] Lester Chandler, *America's Greatest Depression, 1929-1941*, Harper & Row, New York, 1970, p. 82.

Banking panics were not new. There had been several in the late 1800s and one in 1907. One of the original purposes of the Federal Reserve System had been to prevent such panics, by lending reserves to banks through the discount window. However, the Federal Reserve's discount lending failed to prevent the banking panic of the early 1930s. One reason for this failure was that the Federal Reserve kept its discount rate relatively high. Another reason was that there were restrictions on the collateral that member banks could use in borrowing from the Federal Reserve. Finally, more than 15,000 banks were not members of the Federal Reserve and were therefore not eligible to borrow.

Before the Banking Act of 1933, the public's fear over the safety of bank deposits had the nature of a self-fulfilling prophecy. Once withdrawals began at a bank, it was in the interest of every depositor to withdraw his funds, no matter how sound he considered the bank's loans and investments. Even a bank with loans that were certain to be repaid on schedule could be forced to close if enough of its depositors withdrew their funds and if enough of its assets were illiquid. One objective of federal deposit insurance was to prevent such runs by giving every depositor the assurance that his funds would be safe, regardless of whether other depositors withdrew their funds.

The 1933 act established the Federal Deposit Insurance Corporation (FDIC) and provided for insurance of deposits up to $2,500. All commercial banks that were members of the Federal Reserve System were required to take part in the plan. Banks that were not members could join if approved by the FDIC. In 1934, the National Housing Act extended deposit insurance to savings and loan associations by establishing the Federal Savings and Loan Insurance Corporation (FSLIC), a subsidiary of the Federal Home Loan Bank Board.

The maximum deposit insurance coverage has been increased several times since 1933 and now equals $100,000. From the beginning, an insured bank has been required to pay a premium to the FDIC equal to one-twelfth of 1 percent of its total deposits, both insured and uninsured. Since 1950, however, the FDIC has refunded part of its annual surplus to insured banks at the end of each year in the form of a credit against their future premiums. Thus, the effective premium paid by banks has been somewhat less than one-twelfth of 1 percent of their deposits.

Along with its obligation to provide deposit insurance, the FDIC has certain regulatory authority over banks. It has the right to examine all insured banks, though it generally examines only banks that are not members of the Federal Reserve System, relying on the Comptroller of the Currency or the Federal Reserve to examine member banks. If the FDIC determines that a bank is in unsound condition or has engaged in unsound or illegal practices, it can issue a cease-and-desist order against the practices or even terminate the bank's insurance. However, the FDIC has rarely had to resort to such extreme measures to get a bank to make changes.

Effects of deposit insurance on banks' risk-taking behavior

The fundamental dilemma of any deposit insurance system is that it cannot protect depositors against bank failures caused by a sudden withdrawal of funds—that is, failures due to illiquidity—without also protecting depositors against bank failures caused by poor performance of a bank's loans and investments—that is, failures due to basic insolvency. Because deposits are guaranteed against the second type of failure as well as the first, the current deposit insurance system distorted the behavior of banks even before deregulation.[3]

The anatomy of bank failure

Banks raise funds from two sources: deposits and capital. To attract deposits, a bank can offer an explicit return in the form of interest or an implicit return in the form of gifts or services priced below cost, such as free check-clearing. To increase its capital, a bank can either issue new equity or retain some of its profits. The funds obtained from deposits and capital are used to acquire two kinds of assets: noninterest-bearing reserves held largely to meet reserve requirements and loans and investments held to earn income.

A bank's portfolio of assets usually has some risk, in that the total return on the portfolio can vary.[4] One reason the total return varies is that some of the bank's borrowers may default on their loans. Another reason is that changes in market interest rates may cause capital gains or losses on some of the bank's holdings of marketable securities. Of course, the variability of the total return on the bank's assets will usually be somewhat less than the variability of returns on its individual loans and investments, because low returns on some

---

[3] Although the discussion here focuses on risk-taking by commercial banks, much of what is said also applies to other depository institutions — in particular to S&L's, mutual savings banks, and credit unions.

[4] The total return on a bank's assets is the change in the value of the assets from the beginning of the period to the end of the period, including any interest income earned during the period and subtracting any costs incurred in making loans or buying securities. There also will be occasion later to refer to the total expected return on the bank's assets. This is simply a weighted sum of all the possible returns on the bank's assets, with each possible return weighted by its probability of occurrence. The total return to depositors and the total expected return to depositors will be defined analogously.

assets will often be offset by high returns on other assets. Despite this diversification effect, however, most banks have asset portfolios with some risk.

Whether a bank fails depends in part on what happens to the total return on its assets. If the total return on its portfolio of assets exceeds the total return promised on its deposits, the bank earns profits that can either be distributed to shareholders as dividends or retained to increase its capital. However, if the total return on the bank's portfolio falls below the total return promised to depositors, the bank incurs losses that have the effect of reducing its capital. If the losses are great enough to eliminate the bank's capital—and the shareholders are unwilling to contribute new funds—the bank will be forced to close, because it will not have enough resources to repay depositors in full.

When a bank does fail, some parties lose more than others. Who gets what depends partly on whether the FDIC chooses the "payoff" option or the merger option. Under the payoff option, the bank is placed in receivership, the FDIC pays insured depositors in full, and the proceeds from liquidation of the bank's assets are divided up among uninsured depositors, creditors, and the FDIC in proportion to the claim of each on the bank. Under the merger option, the FDIC arranges for another bank to assume the failing bank's liabilities and in return purchases the "bad" assets of the failing bank. In both cases, the shareholders of the failing bank lose their entire investment but enjoy limited liability in that they are not required to dip into their other assets to cover the bank's obligations. Also, in both cases, the failing bank's insured depositors receive the entire amount due to them. The only real difference between the payoff and merger options is that in payoffs the FDIC covers only insured deposits, while in mergers the FDIC in effect covers all deposits, including those that were nominally uninsured.[5]

The moral hazard problem

Under the current system of fixed-rate deposit insurance, a bank has too much incentive to take actions that increase its probability of failure. The reason is that an increase in the probability of failure raises the expected cost to the FDIC of insuring the bank's deposits but does not increase the premium the bank has to pay for the coverage. In other insurance markets, this distortion in risk-taking behavior is sometimes referred to as "moral hazard."[6]

A bank makes two important choices that affect its probability of failure and are therefore subject to moral hazard: how much risk to assume in choosing the composition of its assets, and how much capital to seek relative to deposits. If the total rate of return on the bank's assets is highly variable, the bank has a high probability of earning high profits but also a high probability of incurring large losses.[7] Also, if the bank does not have much

---

[5] The fact that the FDIC did not merge Penn Square Bank when it failed in the summer of 1982 has probably increased uninsured depositors' concern over the safety of their funds. In the past, banks that size had always been merged with healthier banks.

[6] Moral hazard arises whenever the premium a policyholder is charged for insurance fails to reflect the effect of his actions on either the probability or potential magnitude of his loss. For example, if the premium for fire insurance did not vary inversely with the number of smoke detectors or sprinklers in a building, policyholders would have too much incentive to do without such devices, just as a bank that faces a fixed deposit insurance premium has too much incentive to take actions that increase the probability of its failure.

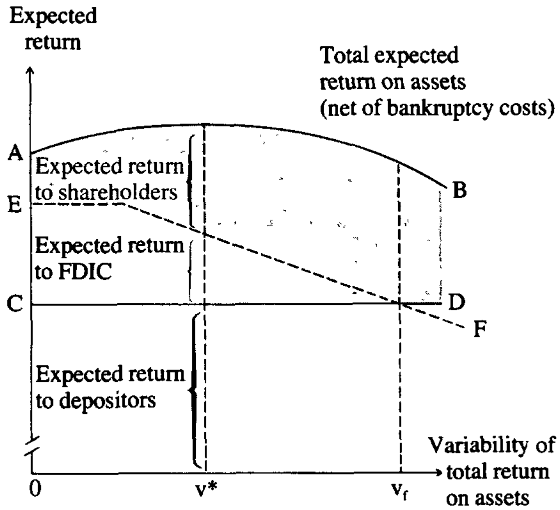
[7] In this article, an increase in the variability of the total return on assets will refer to a shift in density from the center of the probability distribution toward the tails. For an explanation of this concept, see Michael Rothschild and Joseph E. Stiglitz, "Increasing Risk: I. A Definition," *Journal of Economic Theory*, September 1970, and William R. Keeton, *Equilibrium Credit Rationing*, Garland Publishing, New York, 1979, Ch. 3, Sec. II.

capital, a relatively small loss may be sufficient to wipe out its capital and force it to close. Thus, other things being equal, the probability of failure is greater the riskier the bank's assets and the lower the ratio of capital to deposits.

A bank's decision about how much risk to take in choosing the composition of its assets is illustrated by Figure 1. (The decision about how much new capital to raise is discussed in the Appendix.)[8] Assumed as given in Figure 1 are the bank's insured and uninsured deposits, its capital, and its set of investment opportunities. The horizontal axis of the diagram measures the level of asset risk—that is, the degree of variability of the total return on assets. Each point on the horizontal axis corresponds to a different mix of assets.

The bank will choose the level of asset risk that is in the best interests of its shareholders. For convenience, it is assumed that the bank's shareholders care only about the expected return on their investment and not about the variability of the return.[9] Under this assumption, the bank will seek to maximize the total expected return to its shareholders. The total expected return to shareholders equals the total expected return on the bank's assets minus the total expected return to all other parties. The other parties are the bank's insured depositors, the bank's uninsured depositors, and the FDIC.

FIGURE 1
Deposit insurance and asset risk



For every degree of asset risk, the curve AB in Figure 1 represents the maximum total expected return the bank could earn on its assets. This return is net of any bankruptcy costs that would have to be incurred if the bank failed, such as legal and administrative costs of liquidating the bank or losses from distress sale of the bank's assets. The reason these bankruptcy costs must be subtracted is that they use up part of the bank's assets.

Up to a point, increasing the level of risk makes it possible for the bank to earn a higher total expected return on its assets. Thus, the curve AB initially slopes upward. Some of the bank's loan applicants may be new firms having investment projects with a high potential return but also a significant chance of failure. By making loans to these firms instead of firms with safe investment projects, the bank

[8] For more difficult treatments, see John H. Kareken and Neil Wallace, "Deposit Insurance and Bank Regulation: A Partial Equilibrium Exposition," *Journal of Business*, July 1978, and William F. Sharpe, "Bank Capital Adequacy, Deposit Insurance, and Security Values," *Journal of Financial and Quantitative Analysis*, November 1978.

[9] In other words, the bank can act as if its shareholders were "risk-neutral." Even if a shareholder cared about the variability of the return on his total portfolio, he would be indifferent to the variability of the return on his bank shares to the extent that those shares represented a small fraction of his total portfolio and the returns on the other assets in his portfolio were uncorrelated with the return on the bank shares. It is possible, of course, that a bank's managers will act in their own best interests rather than the interests of shareholders. This could lead them to choose a different level of asset risk than the bank's shareholders would prefer.

will increase the riskiness of its asset portfolio. This increase in risk may raise expected bankruptcy costs by making it more likely that the bank fails. However, because the bank will be able to charge a relatively high interest rate on the riskier loans, its expected loan revenues will increase. As long as the increase in expected loan revenues outweighs the increase in expected bankruptcy costs, the total expected return on assets will rise.

Further increases in risk eventually lower the total expected return on assets, causing the curve AB to turn downward. As the bank shifts the composition of its loans toward borrowers with investment projects having still higher potential returns and still higher chances of failure, there will come a point at which the bank cannot fully compensate for the higher probability of default by charging a higher loan rate. The chance of these highly risky projects failing is so great that the bank could not earn as high an expected return on loans made to finance them as on loans made to finance safer projects, even if the bank could receive the entire return from the projects when they succeeded. This reinforces the tendency for increases in risk to lower the total expected return on the bank's assets by raising expected bankruptcy costs. Thus, at point v* in Figure 1, the total expected return on the bank's assets begins to fall.

The total expected return to depositors is represented by the horizontal line CD in Figure 1. For convenience, it is assumed that uninsured depositors can observe exactly how much risk the bank is taking and, like shareholders, do not care about the variability of the return on their investment. This means that whenever the bank increases its probability of failure by choosing a riskier portfolio of assets, uninsured depositors will demand an increase in the deposit rate just large enough to prevent the expected return on their investment from falling.[10] The return to insured depositors is also constant because it is guaranteed by the FDIC. Thus, the total expected return to depositors must be independent of the level of asset risk chosen by the bank. This is why CD is horizontal.

The combined expected return to depositors and the FDIC is represented by the curve EF in Figure 1. The expected return to the FDIC equals the total insurance premium paid by the bank minus the expected cost to the FDIC of compensating insured depositors. If a bank does not fail, the cost of compensating insured depositors is zero. On the other hand, if the bank fails, the cost of compensating insured depositors is the total amount due to insured depositors minus the FDIC's share of whatever assets remain after bankruptcy costs. Under the current deposit insurance system, the total insurance premium depends only on the level of deposits and not on the level of asset risk chosen by the bank. If the level of asset risk is sufficiently low, the bank has no chance of failing and the expected return to the FDIC equals the fixed insurance premium. However, as the level of asset risk is increased, the probability of failure eventually becomes positive. At that point, the expected cost of compensating insured depositors rises above zero and the expected return to the FDIC begins to fall. This is why EF starts out as a horizontal line above CD and then turns downward.

To serve the best interests of its shareholders, the bank will try to maximize the difference between the total expected return on its assets and the combined expected return to depositors and the FDIC. In Figure 1, this difference is represented by the gap between the

---

[10] In contrast to insured deposits, large uninsured deposits were not subject to deposit rate ceilings even before the recent deregulation.

curves AB and EF, or the height of the shaded area. The bank chooses the level of asset risk at which this gap is largest. This is v_f, the point where the curves AB and EF have equal slopes.[11]

The level of risk chosen by the bank in Figure 1 exceeds the socially optimal level of risk. Since the bank's shareholders and uninsured depositors do not care about the variability of the return on their investment, it is in society's interest for the bank to choose the asset portfolio with the highest total expected return. In Figure 1, this is v*, the point where the curve AB attains its maximum value. Although this point is optimal for society, it cannot be optimal for the bank's shareholders. Increasing risk beyond v* reduces the expected return to the FDIC more than the total expected return on assets and thus increases the total expected return to shareholders. In other words, because AB has zero slope where it reaches a maximum while EF has negative slope, the gap between AB and EF can always be increased by moving at least a little bit to the right of v*.[12] The loss to society from the bank's choice of excessive asset risk equals the difference between the total expected return on the bank's assets at v* and the total expected return at v_f. Part of the social loss is due to an increase in expected bankruptcy costs. The rest is due to a shift in composition of the bank's loans toward borrowers with less productive investment projects.

It should not be inferred from the example above that a moral hazard problem must exist for all banks. Other banks may differ from the bank in Figure 1 in two important respects.

First, other banks may not face the same investment opportunities as the bank in Figure 1. As a result, the curve AB indicating the tradeoff between the total expected return on their assets and the variability of the return on their assets may look different. For example, if a bank faces relatively safe investment opportunities, the point v* where the curve AB attains its maximum value may lie further to the left. Conversely, if the bank faces highly risky investment opportunities, v* may lie further to the right. To some extent, these differences in investment opportunities are due to restrictions on interstate and intrastate branching. Even without such restrictions, however, the differences would be likely to exist because the costs of investigating borrowers and monitoring their investment projects make it efficient for banks to specialize in a particular kind of lending or a particular geographical market.

Second, other banks may have shareholders with different attitudes toward variability in the return on their investment. For example, a bank's equity may be concentrated in the hands of a few people.[13] These shareholders may dislike variability in the return on their bank shares because they dislike variability in the return on their total portfolios and because the bank shares represent a large proportion of

---

[11] The diagram assumes that the fixed premium happens to be just high enough to allow the FDIC to break even at the level of risk actually chosen by the bank. This is the only reason the curve EF crosses the line CD at the same point where the gap between EF and AB is largest. With a different premium, the bank could end up being either overcharged or undercharged for insurance. In general, however, it would still end up choosing more risk than was socially optimal.

[12] Although the bank shown in Figure 1 chooses more risk than is socially optimal, it does not choose the highest possible level of asset risk. There are two reasons this might be the case. First, to keep increasing risk, the bank may have to shift the composition of its loans toward borrowers with investment projects having a much higher chance of failure but almost the same potential return. Second, because some deposits are uninsured and because uninsured depositors demand a large enough increase in the deposit rate to keep their expected return from falling, the rise in expected bankruptcy costs that occurs as the level of asset risk is increased will fall partly on the bank's shareholders rather than entirely on the FDIC. Both factors will tend to make the curve AB fall more sharply than the curve EF as the level of risk gets very high.

these portfolios. In terms of Figure 1, this would mean that the socially optimal level of asset risk would be less than v*. Alternatively, shareholders may prefer variability in the return on their investment because they like to gamble. In this case, the socially optimal level of risk would exceed v*.

Whether a moral hazard problem exists for a particular bank depends on both the nature of its investment opportunities and the attitude of its shareholders toward the variability in their return. If the bank faces relatively safe investment opportunities or has shareholders who are highly averse to variability in their return, the socially optimal level of asset risk may be low enough that the bank has no chance of failing at that level of risk. In such cases, a moral hazard problem may not exist — that is, the degree of risk that is optimal for society may also be optimal for the bank's shareholders. This is because the bank may have to increase risk significantly beyond the socially optimal level to shift some of the expected return on its assets from the FDIC to shareholders. An increase in risk that large may entail too great a reduction in the total expected return on assets or too large an increase in the variability of the return to shareholders to leave shareholders with a net gain.[14] However, if the bank faces relatively risky investment opportunities or has shareholders who are not highly averse to variability in their return, the socially optimal degree of risk will be high enough that the bank has at least some chance of failing at that level of risk. In such cases, a moral hazard problem will necessarily exist, because even a very small increase in risk beyond the socially optimal level will tend to benefit shareholders at the expense of the FDIC.

Regulatory and legislative limits on risk-taking

To some extent, the moral hazard problem that arises under the current system of fixed-rate deposit insurance may have been offset through regulatory and legislative limits on risk-taking by banks. These limits have taken two forms.[15]

One way risk-taking has been curbed is through explicit limits on the types of loans and investments banks can make. For example, national banks have always been prohibited from purchasing equities and investing more than a specified percentage of their capital in loans to the same borrower. Also, both national and state banks have faced a number of explicit restrictions on the amount and terms of their real estate loans and their margin loans for the purchase of securities.

The other way risk-taking has been limited is through regulatory supervision aimed at ensuring that each bank remains in sound condition. With the cooperation of other regulatory agencies, the FDIC has tried to limit the amount of risk a bank can take so that the total premiums it collects from all banks will be adequate to cover the total expected costs of insuring deposits. In examining a bank, regulators assess both the quality of the bank's

---

[13] In some cases, it may be more efficient for shares to be owned by a few people who can closely control the bank's management than by a large number of investors who cannot exercise such control. See Michael Jensen and William Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure," *Journal of Financial Economics*, October 1976.

[14] In terms of Figure 1, the point at which the curve EF turns downward may lie too far to the right of the socially optimal level of asset risk.

[15] For up-to-date descriptions of the bank regulatory system, see Carter H. Golembe and David S. Holland, *Federal Regulation of Banking, 1983-84*, Golembe Associates, Washington, D.C., 1983, and Kenneth Spong, *Banking Regulation: Its Purposes, Implementation, and Effects*, Federal Reserve Bank of Kansas City, 1983.

assets and the adequacy of its capital. If the bank is found to have too many doubtful loans or too little capital, the FDIC or other agency making the examination requests the bank to enter a written agreement to correct the problem.

Although regulation may have alleviated the problems associated with fixed-rate deposit insurance, it has clearly not eliminated them. Limits on risk-taking cannot be perfectly enforced. And even if limits on risk-taking could be perfectly enforced, they would not result in every bank choosing the correct amount of risk. Because the socially optimal level of risk differs across banks, a uniform upper limit on risk-taking will necessarily be too low for some banks and too high for others. In other words, banks for which the socially optimal level of risk is relatively low will be allowed to take too much risk, while banks for which the socially optimal level of risk is relatively high will be forced to take too little risk.[16] These limitations of regulation are illustrated in the accompanying box.

Effects of deposit-rate deregulation on banks' risk-taking behavior

The only deposits not subject to interest rate ceilings in the 1970s were certificates of deposit in denominations of $100,000 or more. During this period, however, the maximum deposit insurance coverage remained well below $100,000. Thus, all fully insured deposits were subject to rate ceilings. Two developments have recently altered this situation. First, in March 1980, the maximum deposit insurance coverage was increased from $40,000 to $100,000, making available for the first time a deposit that was not subject to rate ceilings and on which the principal was fully insured—$100,000 CD's.[17] Second, the Depository Institutions Deregulation and Monetary Control Act was passed in 1980, calling for interest-rate ceilings to be gradually phased out on all deposits except demand deposits. The only deposits other than demand deposits that are still subject to rate ceilings are regular NOW accounts, passbook savings accounts, and small time deposits that mature in seven to 31 days. Even ceilings on these deposits are scheduled to be removed soon, and legislation has been introduced in Congress to eliminate the prohibition of interest on demand deposits. Thus, the range of insured deposits for which banks are free to bid has increased dramatically since 1980 and will increase still further in the next few years.

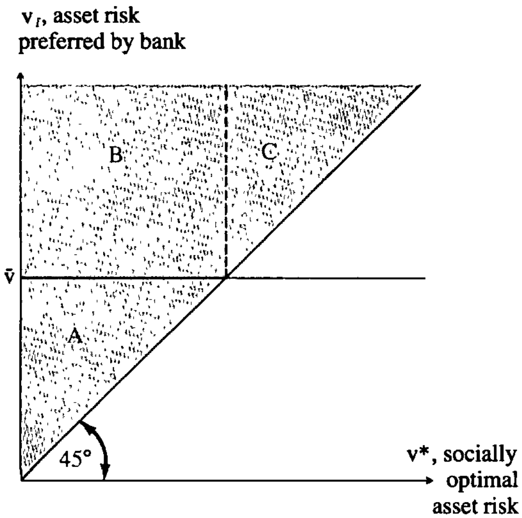
To some extent, banks were able to circumvent deposit-rate ceilings by paying their depositors an implicit return in the form of gifts, convenient locations, free checking, and other services priced below cost. However, the degree to which banks were able to evade the ceilings was limited by the range of services they could provide that were of value to depositors. For business demand deposits, the implicit rate of return paid by banks in the 1970s probably approached the competitive rate of return — the rate that would have been paid in the absence of ceilings — because businesses tended to use a large number of bank services. In the case of household

---

[16] The latter effect is emphasized in Kenneth E. Scott and Thomas Mayer, "Risk and Regulation in Banking: Some Proposals for Federal Deposit Insurance Reform," *Stanford Law Review*, May 1971, pp 872-73, 888.

[17] It has long been the practice of the FDIC to cover both the depositor's initial investment and the accumulated interest at the time the bank closes, as long as the total does not exceed the maximum coverage limit. Thus, on a $100,000 deposit only the principal would be insured, while on smaller deposits both the principal and interest would be insured. This practice was recently formalized in the FDIC's official regulations. See 48 Federal Register 52030-31, November 16, 1983.

Effect of regulation on bank risk-taking



The diagram above illustrates the limitations of regulation. Every bank is represented by a point in the diagram. The horizontal axis measures a bank's socially optimal degree of asset risk, corresponding to point v* in Figure 1. The vertical axis measures a bank's preferred amount of asset risk, corresponding to point v_f in Figure 1. Under the assumptions made earlier, every bank will choose at least the socially optimal degree of risk. Thus, all banks will fall in the shaded area in the diagram, with v_f greater than or equal to v*.

The principal purpose of bank supervision and regulation can be viewed as the imposition of an upper limit on risk-taking. In the diagram, this limit is $\bar{v}$. To the extent that regulation is effective, it will tend to alleviate the moral hazard problem by forcing banks in region B in the diagram to reduce their risk-taking closer to the socially optimal level. However, regulators may not be able to enforce the upper limit perfectly, so that some banks in region B continue to choose a level of risk greater than $\bar{v}$. Furthermore, even if the upper limit is perfectly enforced, banks falling in region B will take more risk than is socially optimal and banks falling in region A will not be affected at all. Finally, banks for whom the socially optimal degree of risk is relatively high—those falling in region C—will be forced to reduce their risk-taking below the socially optimal level.

demand deposits and time deposits, however, banks probably did not pay close to the competitive rate because households used relatively few bank services. Thus, despite the fact that banks had been able to pay some implicit interest, the deregulation of deposit rates that began several years ago should increase competition for insured deposits and result in banks paying a higher total return on insured deposits than if deregulation had not occurred.[18]

The controversy over deposit rates and risk-taking

The relationship between deposit rates and

[18] One study has estimated that implicit interest on demand deposits averaged one-third to one-half of the competitive rate in the 1970-74 period. See Richard Startz, "Implicit Interest on Demand Deposits," *Journal of Monetary Economics*, October 1979. A useful survey of the empirical evidence on implicit interest on demand deposits is John P. Judd and John L. Scadding, "Financial Change and Monetary Targeting in the United States," in Federal Reserve Bank of San Francisco, *Interest Rate Deregulation and Monetary Policy*, November 1982, pp. 85-89.

bank risk-taking behavior was a controversial issue among policymakers and economists long before the recent financial deregulation. When deposit-rate ceilings were originally imposed in 1933, one of the reasons given was to prevent a recurrence of the widespread bank failures of the early 1930s. Without ceilings, it was argued, banks would engage in "ruinous competition." In particular, it was felt banks would bid up deposit rates in competition for funds and then try to cover the increased cost of funds by acquiring risky assets with high potential returns. This kind of behavior was thought to be partly responsible for the more than 5,000 bank failures from the end of 1929 to the end of 1932.

The argument that higher deposit rates would induce banks to invest in riskier assets has been widely disputed, on both empirical and theoretical grounds. Influential empirical studies by George Benston and Albert Cox based on the period before ceilings were first imposed found no evidence that banks paying higher deposit rates also took more risk.[19] Furthermore, such behavior has been alleged to be inconsistent with profit maximization by banks. For example, Benston claimed:

> The willingness of a banker to invest in assets bearing any perceived degree of risk is a function of the expected returns from the investment and the inclination of the banker toward risk-taking. Thus, the interest rate on deposits offered by a banker is a function of the investment possibilities (and their associated risks) available to the banker, rather than the reverse.[20]

Some years later, Carl Gambs attempted to refute these claims. Drawing on the theory of portfolio behavior developed by Harry Markowitz and James Tobin, Gambs showed that a bank that cared about the variability of the return to its shareholders might respond to an increase in deposit rates by choosing a mix of assets with higher risk but also higher expected return. However, his argument was subsequently shown to hold only under special assumptions about shareholders' attitudes toward the variability of their return.[21] As a result, support for the notion that the removal of deposit-rate ceilings will increase risk-taking by banks must be found elsewhere.

In the remainder of this section, two alternative reasons are suggested for why bank risk-taking should increase. First, the removal of ceilings on insured deposits should exacerbate the distortion that already exists as a result of the moral hazard problem. Second, it should create a relatively new distortion by making it much easier for risky banks, which are currently undercharged for deposit insurance, to bid deposits away from safe banks, which are currently overcharged for deposit insurance.

Increase in the existing distortion from moral hazard

It was shown earlier that a moral hazard problem exists under the current system of fixed-rate deposit insurance. After ceilings are removed and interest rates on insured deposits

[19] See George J. Benston, "Interest Payments on Demand Deposits and Bank Investment Behavior," *Journal of Political Economy*, October 1964, and Albert M. Cox, Jr., "Regulation of Interest Rates on Bank Deposits," *Michigan Business Studies*, Vol. 17, No. 4, 1966.

[20] Benston, p. 433. For another expression of this view, see Milton Friedman and Anna Jacobson Schwartz, *A Monetary History of the United States, 1867-1960*, Princeton University Press, Princeton, N.J., 1963, p. 444.

[21] See Carl M. Gambs, "Interest-Bearing Demand Deposits and Bank Portfolio Behavior," *Southern Economic Journal*, July 1975, and the comments on Gambs' article by Philippe Caperaa and Louis Eeckhoudt and by Perry D. Quick in the October 1977 issue of that journal.

are bid up, this moral hazard problem should become worse. One reason for this worsening is that the rise in interest rates on insured deposits will directly increase the total payment the FDIC has to make to a bank's insured depositors if the bank fails. Another reason is that the percentage of deposits that are insured will increase.

Direct increase in the FDIC's potential liability. Under current arrangements, the FDIC covers both the principal and interest on insured deposits as long as the total due to the depositor does not exceed the coverage limit of $100,000. This means that an increase in the interest rates a bank pays on insured deposits smaller than $100,000 will increase the FDIC's total potential liability to depositors even if the quantities of both insured and uninsured deposits remain unchanged.

Because the rise in rates on insured deposits will increase the FDIC's total potential liability, many banks will have a greater incentive to take risk.[22] When a bank chooses a riskier mix of assets or a lower level of capital, it increases its chance of failure and thus also increases the chance that the FDIC's potential liability will be an actual liability. As a result, the increase in the FDIC's potential liability will enhance the tendency for an increase in asset risk or a decrease in capital to shift some of the total expected return on the bank's assets from the FDIC to shareholders. Of course, the rise in rates on insured deposits will also tend to reduce the expected return to the bank's shareholders. If shareholders cared about the variability of the return on their investment, this decline in expected return could make them more conservative, just as a reduction in a person's wealth could make him less willing to gamble. Unless this effect is large, however, the greater tendency for increases in risk to reduce the expected return to the FDIC will mean that it is in shareholders' best interests for the bank to choose a higher level of risk.[23]

The potential increase in the moral hazard problem can be illustrated by Figure 1, the diagram used earlier to explain a bank's choice of asset risk. An increase in interest rates on insured deposits shifts up the curve EF representing the combined expected return to depositors and the FDIC. This tends to reduce the expected return to the bank's shareholders at each level of asset risk by narrowing the gap between EF and the curve AB representing the total expected return on the bank's assets. However, at v_f, the level of asset risk initially chosen by the bank, EF also becomes steeper relative to AB because of the greater tendency for increases in risk to reduce the expected return to the FDIC. As a result, the bank can widen the gap between the two curves and reduce the adverse impact of the higher deposit rates on the expected return to its shareholders by increasing risk.[24]

Increase in the percentage of insured deposits. As interest-rate ceilings on insured deposits are removed and rates on those deposits are bid up relative to rates on unin-

[22] The tendency for an increase in the rate at which funds are borrowed to exacerbate the moral hazard problem has been used in a different context to explain why banks that are unable to monitor borrowers' investment projects might refrain from raising their loan rates to market-clearing levels See Keeton, *Equilibrium Credit Rationing*, Ch. 3.

[23] Because banks were able to partially circumvent the ceilings by paying an implicit return on deposits, the removal of ceilings should lead to a decrease in losses from services priced below cost as banks substitute explicit interest for implicit interest. This will work in the opposite direction from the increase in the FDIC's potential liability to depositors. In particular, it will tend to reduce the effect of increases in risk on the expected return to the FDIC by increasing the amount of the bank's assets that will be available to the FDIC to help pay insured depositors if the bank fails. However, because there were limits to the amount of implicit interest banks could pay, it is unlikely that the decrease in losses from below-cost services would be great enough to offset the increase in the FDIC's potential liability to depositors

sured deposits, many investors will shift from uninsured deposits to insured deposits, causing the percentage of insured deposits to rise.[25] Since there is a $100,000 limit on the size of each insured deposit, some large investors might still prefer to hold uninsured deposits, even if banks pay the same rate on uninsured and insured deposits. However, individuals with large amounts of money to invest can obtain full insurance by splitting up their funds into smaller deposits at different banks. In the last couple of years, the cost of doing this has been significantly reduced by the growth of money brokers acting as intermediaries between investors and banks.[26] Thus, under the current system, the percentage of insured deposits could well approach 100 percent.

If uninsured depositors are able to monitor banks' risk-taking, the increase in the percentage of insured deposits should make the moral hazard problem worse. Earlier in the article it was assumed that uninsured depositors could determine exactly how much risk a bank was taking and that they would respond to any increase in risk by demanding a large enough increase in the deposit rate to prevent their expected return from falling. In this extreme case, the effect of an increase in the percentage of insured deposits is clear. A rise in the percentage of insured deposits increases the total amount the FDIC has to pay insured depositors if the bank fails. Thus, given total deposits and total assets, an increase in risk that raises the bank's probability of failure will shift more of the expected return on the bank's assets from the FDIC to shareholders and still leave the expected return to uninsured depositors unchanged. This means banks will have more incentive to increase asset risk and less incentive to raise new capital.[27]

In practice, uninsured depositors cannot observe all increases in risk and thus cannot always demand a large enough increase in deposit rates to keep their expected return from falling. As a result, increases in risk have a tendency not only to shift the expected return on the bank's assets from the FDIC to shareholders but also to shift the expected return on the bank's assets from uninsured depositors to shareholders.

In these circumstances, an increase in the percentage of insured deposits should still make the moral hazard problem worse, but not as much as when uninsured depositors can monitor risk perfectly. If the percentage of insured deposits rises but total deposits and total assets remain unchanged, an increase in

---

[24] If there are bankruptcy costs, the AB curve will shift down at the same time the EF curve shifts up, because with higher deposit rates the bank will have a greater chance of failure at every level of asset risk. At the bank's initial choice of asset risk, it is also possible that increases in risk will now have a greater tendency to raise expected bankruptcy costs. Although this would make the AB curve steeper, it would also tend to make the EF curve steeper because any increase in expected bankruptcy costs will be borne in large part by the FDIC. Thus, the new EF curve should still be steeper than the new AB curve at the bank's initial choice of asset risk.

[25] This is especially likely since the premium a bank has to pay for deposit insurance under the current system depends on its total deposits, including those that are uninsured. Under this arrangement, there is no reason for a bank operating in a highly competitive deposit market without ceilings to pay more on its uninsured deposits than on its insured deposits, assuming they cost the same amount to service. If the bank did pay more on its uninsured deposits, it could always obtain the same total funds at lower cost by taking fewer uninsured deposits and raising its rate on insured deposits slightly so as to bid away insured deposits from other banks.

[26] Under current regulations, the $100,000 coverage limit does not apply to the total amount of funds placed by a money broker at a bank, but instead to an investor's share of the total. In an effort to limit the brokering of insured deposits, the FDIC and FSLIC have recently proposed changing the regulations to make the $100,000 limit apply to the total amount of funds placed by the broker. See *American Banker*, January 17, 1984, p. 1.

[27] This assumes that uninsured depositors do not already regard their deposits as effectively insured because the FDIC chooses the merger option rather than the payoff option in the event of failure.

risk that raises the bank's probability of failure will produce a larger total shift in expected return from the FDIC to shareholders but a smaller total shift in expected return from uninsured depositors to shareholders. If uninsured depositors had no ability to monitor risk, these two effects would cancel out, leaving the bank with the same incentive to take risk as before. In most cases, however, uninsured depositors probably do have some ability to monitor risk and protect the expected return on their investment. As a result, the somewhat smaller tendency for increases in risk to benefit shareholders at the expense of uninsured depositors should be outweighed by the greater tendency for increases in risk to benefit shareholders at the expense of the FDIC.

In terms of Figure 1, the curve EF representing the combined expected return to the FDIC and depositors will become steeper. As in the case of a direct increase in the FDIC's potential liability to depositors, this means the bank will be able to widen the gap between AB and EF—and thus increase the total expected return to its shareholders—by choosing a higher level of asset risk than before.

Creation of a new distortion from cross-subsidization

The deregulation of deposit rates would exacerbate the moral hazard problem even if all banks were identical. However, banks are not identical. Some will prefer to take more risk than others because they face riskier investment opportunities or because their shareholders are less averse to variability in the return on their investment. Furthermore, some banks that are basically insolvent because their past loans have little chance of being repaid may be able to escape the attention of regulators and remain in operation. These banks will be especially willing to take risks because their only hope of earning a positive return for their shareholders is to acquire risky assets with potential returns high enough to make up for previous losses. Despite these differences in risk-taking among banks, the FDIC charges all banks the same premium per dollar of deposits, resulting in subsidization of relatively risky banks by relatively safe banks. In other contexts, this phenomenon is often referred to as "cross-subsidization."[28]

Although cross-subsidization between risky banks and safe banks might be considered unfair, the existence of interest-rate ceilings on insured deposits at least helped keep it from affecting the distribution of deposits. While ceilings were binding, both risky and safe banks would probably have been willing to incur greater costs to obtain insured deposits. In other words, most banks probably could not pay enough implicit interest on insured deposits to circumvent completely the limits on explicit interest. However, because risky banks were receiving deposit insurance below cost and safe banks were receiving deposit insurance above cost, risky banks would have been willing to pay even more than safe banks for insured deposits. By making it more difficult for all banks to bid for

[28] Like the moral hazard problem, the cross-subsidization of deposit insurance has close analogies in other insurance markets. For example, it has been proposed that drivers be charged a fixed premium for automobile insurance on every gallon of gasoline consumed. Although a driver's probability of having an accident depends on the amount of driving he does and thus indirectly on the amount of gasoline he consumes, there are many other factors that also influence his chance of having an accident, such as his skill as a driver or the kind of traffic conditions in which he does most of his driving. Thus, if all drivers paid the same premium per gallon of gasoline consumed, risky drivers would be undercharged for insurance and safe drivers overcharged, just as risky banks are currently undercharged for deposit insurance and safe banks overcharged. In this example, a price ceiling on gasoline would play the same role as interest rate ceilings on insured deposits.

deposits, interest-rate ceilings also made it more difficult for risky banks to outbid safe banks for deposits.

The deregulation of deposit rates will remove this constraint on competition for funds and distort aggregate risk-taking behavior. Now banks can not only compete for insured deposits within their own regions but also compete for insured deposits from other regions by selling large CD's and making use of money brokers. From society's point of view, the increased competition for funds will have the advantage of allowing banks with highly productive lending and investment opportunities—those for which the curve AB in Figure 1 is relatively high—to increase their share of total deposits. However, it will also have the disadvantage of allowing banks with highly risky lending and investment opportunities—those for which the curve AB in Figure 1 lies relatively far to the right—to increase their share of total deposits.[29] This is a disadvantage from society's point of view because risky banks may end up expanding their lending and investment to a point where the expected return on the last unit of their assets is significantly less than the expected return on the last unit of safe banks' assets. This could happen even if there were no moral hazard problem in the choice of asset risk—that is, even if every bank chose the level of asset risk that was socially optimal given its total assets.[30]

Alternative policy responses

This article has argued that the deregulation of deposit rates will increase the distortion in bank risk-taking behavior that already exists under the current system of fixed-rate deposit insurance. Although a detailed analysis of reforms in the deposit insurance system is beyond the scope of the article, some of the possible policy responses to the problems posed by deposit-rate deregulation can be briefly discussed.

Since deposit-rate deregulation will increase the distortion in bank risk-taking, the question naturally arises whether an appropriate response to the problem would be to reimpose ceilings that have already been removed and maintain those ceilings that are still in effect. Although the removal of ceilings will reduce economic efficiency to some extent by increasing the distortion in banks' risk-taking, it is important to realize that it will increase economic efficiency in other ways. For example, the removal of the ceiling on checkable deposits will eliminate the waste of resources resulting from households and firms trying to economize on their holdings of transactions balances. By making it possible for banks in aggregate to attract more deposits, deposit-rate deregulation will also enable the banking industry to increase its total lending and investment, financing some high-return investment projects that might not otherwise be undertaken. Finally, the removal of deposit-

[29] A related problem is that the owners or managers of a bank may seek insured deposits with the intention of diverting the funds to their own uses — either legally in the form of higher salaries and perquisites or illegally through outright theft The removal of deposit-rate ceilings enables these banks to increase their share of total deposits along with banks that intend to invest depositors' funds in highly risky assets.

[30] For the removal of interest-rate ceilings on insured deposits to shift the distribution of deposits toward risky banks, it is sufficient that risky banks desire more insured deposits than safe banks at every level of deposit rates and that risky banks be unable to obtain as much insured deposits as they would like with the ceilings. However, because the degree of cross-subsidization between safe banks and risky banks increases with the level of interest rates on insured deposits, the demand of risky banks for insured deposits will not only exceed the demand of safe banks at every level of deposit rates but will do so by a greater amount the higher the deposit rate. This phenomenon, often referred to as "adverse selection," makes the problem described here even worse by causing the distribution of deposits to shift toward risky banks by an even greater amount after ceilings are removed.

rate ceilings will tend to improve the distribution of total deposits to the extent that it enables banks with highly productive investment opportunities to bid funds away from banks with investment opportunities that are equally risky but less productive. For all these reasons, deposit-rate deregulation should on balance increase economic efficiency. Reversing the deregulation of deposit rates would amount to throwing the baby out with the bathwater.[31]

A second response to the increased distortion in bank risk-taking would be to reduce FDIC insurance coverage. For example, the FDIC could lower the maximum coverage below $100,000 and make less use of the merger option when banks fail. Because uninsured depositors often cannot observe exactly how much risk a bank is taking or would like to take, both a moral hazard problem and a cross-subsidization problem would remain even if deposit insurance were eliminated altogether. However, because uninsured depositors do in most cases have some ability to monitor risk, it is likely that reducing FDIC coverage would at least partially offset the adverse effects of deposit-rate deregulation on risk-taking.[32] The major problem with this approach is that it would revive the danger of banking panics by giving depositors more reason to worry about the possibility that other depositors would withdraw their funds from the same bank. In addition, efforts to reduce the FDIC coverage limit could well be frustrated by increased brokering of insured deposits.

A third response to the problem would be to adopt some form of variable-rate deposit insurance so that the premium a bank paid for deposit insurance depended on the amount of risk it was taking. To the extent it could be implemented, such an approach would reduce both the moral hazard problem and the cross-subsidization problem. Also, unlike a reduction in FDIC coverage, it would not increase the danger of banking panics. The only problem with this approach—but a major one—is the difficulty of measuring risk. The FDIC can determine how much capital a bank has with reasonable accuracy, but it cannot easily determine how risky a bank's loans and investments are. Nevertheless, the fact remains that the FDIC already collects some information about the amount of risk banks are taking. Instead of using that information to impose an upper limit on risk-taking, it could just as well use it to set premiums that varied with risk.[33]

Some observers have argued that private insurers would do a better job of measuring

[31] Although it would be unwise to reimpose deposit-rate ceilings, the FDIC should at least make sure that deposit rates do not end up excessively high because deposit insurance is underpriced. In other words, if deregulation leads to an increase in aggregate risk-taking by banks at every level of total deposits, the FDIC should raise the insurance premium on each dollar of deposits to help cover the increase in the expected cost of compensating depositors. If the premium per dollar of deposits were left unchanged, the removal of deposit-rate ceilings would not only misallocate deposits between safe banks and risky banks but also lead to too high a level of deposits — and thus too high a level of lending and investment — in the banking industry as a whole.

[32] A possibility that has not received much attention would be to eliminate FDIC coverage of the interest on insured deposits — that is, restrict coverage to the principal. If depositors could determine exactly how much risk a bank was taking, this change would eliminate the tendency for deposit-rate deregulation to exacerbate the moral hazard problem by directly increasing the FDIC's potential liability on every dollar of insured deposits. Deposit-rate deregulation would still tend to exacerbate the moral hazard problem by increasing the percentage of insured deposits and would still make it easier for risky banks to outbid safe banks for insured deposits. However, both effects would be weaker.

[33] The FDIC has recently come out in favor of some modest variation in insurance premiums. See Federal Deposit Insurance Corporation, *Deposit Insurance in a Changing Environment*, April 15, 1983. In this report, which was submitted to Congress in compliance with the Garn-St. Germain Act, the FDIC also recommended other changes in the deposit insurance system.

and pricing risk than the FDIC because the profit motive gives them a stronger incentive. If this were true, another way of getting banks to bear the true costs of their risk-taking would be to make deposit insurance compulsory for deposits of $100,000 or less but allow banks to obtain some or all of their coverage from private insurers rather than the FDIC, if they so chose. Because all deposits would still have to be insured up to $100,000, the danger of banking panics would be much less than it was before the FDIC was established. However, that danger would still be significantly greater than it is now. This is because every depositor at a privately insured bank would have to worry about the insurer's ability to pay claims in the event that withdrawal by other depositors forced the bank to close. Also, even if there were no danger of banking panics, private companies might be reluctant to insure deposits on a large scale because bank failures are not independent risks. Since a nationwide recession increases the probability of failure for all banks simultaneously, private insurers cannot rely on the law of large numbers to reduce risk through diversification, as a life insurer or automobile insurer can.

The final and least dramatic response to the increased distortion in bank risk-taking would be to strengthen bank supervision and regulation—that is, do a better job of enforcing existing limits on risk-taking and perhaps tighten those limits as well. However, if more resources were devoted to bank examinations, it might be better to use those examinations to collect information suitable for setting variable insurance premiums than to use them to enforce an upper limit on the amount of risk banks can take. Also, while a reduction in the upper limit on risk-taking might prevent some banks from taking more risk in response to the deregulation of deposit rates, it would also have the disadvantage of forcing other banks to reduce their risk-taking still farther below the socially optimal level.

Summary and conclusions

Even before the recent financial deregulation, fixed-rate deposit insurance distorted banks' risk-taking behavior by creating a moral hazard problem. Because the FDIC charges a fixed premium per dollar of insured deposits, banks are not forced to bear the full expected costs of their risk-taking. As a result, banks have had a greater incentive to choose a risky mix of assets and a smaller incentive to raise new capital to provide a cushion against losses.

The removal of deposit-rate ceilings and the subsequent bidding up of interest rates on insured deposits is likely to exacerbate this moral hazard problem by enabling banks to shift more of the costs of their increased risk-taking to the FDIC. The increase in deposit rates should do this in two ways, by directly increasing the potential liability of the FDIC on every dollar of insured deposits and by leading to an increase in the percentage of total deposits that are insured. Besides increasing the moral hazard problem, the deregulation of deposit rates should allow a relatively new distortion in aggregate risk-taking behavior to arise as a result of the cross-subsidization of deposit insurance. In particular, the removal of ceilings should make it much easier for risky banks, which are currently undercharged for deposit insurance, to bid deposits away from safe banks, which are currently overcharged for deposit insurance.

There are no easy ways to prevent this increased distortion in bank risk-taking behavior. Reimposing deposit-rate ceilings would reduce economic efficiency in a number of important ways, while reducing FDIC coverage would increase the danger of banking panics like those experienced before 1933. A more promising approach would be for the FDIC to use the information it already collects on banks' risk-taking to introduce some variation in insurance premiums, to collect more information about banks' risk-taking, and to encourage a limited degree of competition from private insurers.

Appendix

This appendix explains the moral hazard problem that exists with respect to a bank's choice of capital under the current system of fixed-rate deposit insurance. This distortion is explained in terms similar to those used in the text to explain the distortion in the choice of asset risk.

In deciding how much new capital to raise, a bank will act in the best interests of its present shareholders, the owners of the shares already outstanding. As in the choice of asset risk, it can be assumed that the bank does this by maximizing the total expected return on their investment and not worrying about the variability of the return. In other words, the bank issues that amount of new equity that maximizes the gap between the total expected return on its assets, net of bankruptcy costs, and the combined expected return to all other parties. Those parties now include not only the FDIC and depositors but also the bank's new shareholders, the investors buying the new equity.

Consider the effect of increasing the amount of new capital the bank raises while holding constant both the amount of deposits and the degree of relative asset risk — that is, the degree of variability in the actual return on assets relative to the expected return on assets. This will change the total expected return to the bank's present shareholders in three ways. First, because the bank uses the extra funds to acquire more assets, the total expected return on its assets will increase. As the amount of capital is increased, this effect should eventually diminish in size because, to increase its total lending, the bank will have to make loans to borrowers with less productive investment projects. Second, because each new share must be sold at a price low enough to guarantee the buyer a positive expected return on his investment, the total expected return to new shareholders will rise. This effect will not diminish in size as the amount of new capital is increased. Third, because the extra capital provides a greater cushion against losses on assets, the bank's probability of failure will fall. Since the insurance premium is fixed, this raises the expected return to the FDIC.

Barring imperfections in the capital markets in which the bank sells its equity, it will be in society's interest for the bank to increase the level of new equity sales to the point where the difference between the total expected return on assets and the total expected return to new shareholders is highest. As the level of new equity sales approaches this point, the first and second effects described above will just offset each other. However, as long as the bank has some chance of failing, the third effect will continue to operate. That is, increases in capital will still have a tendency to reduce the total expected return to the bank's present shareholders by increasing the expected return to the FDIC. Thus, with fixed-rate deposit insurance, it will be in the interest of present shareholders for the bank to stop short of the socially optimal level of new equity sales.

With a few modifications, the same diagram used to illustrate the distortion in the choice of asset risk can also be used to illustrate the distortion in the choice of capital. Let the horizontal axis of Figure 1 now measure the amount of new capital foregone, so that a rightward movement in the diagram corresponds to a reduction in the amount of new capital raised. Also, let the curve AB now represent the total expected return on the bank's assets minus the total expected return that must be offered to investors to get them to buy the new shares. Finally, let the curve EF continue to represent the combined expected return to the FDIC and the bank's depositors. Under these conditions, the total expected return to the bank's present shareholders equals the gap between the curves AB and EF.

The social optimum occurs at the point where the curve AB reaches its highest value. Reducing the amount of new capital and moving to the right of this point decreases the expected return to the FDIC, as indicated by the downward slope of the curve EF. Thus, the bank maximizes the total expected return to its present shareholders—the gap between curves AB and EF—by raising less than the socially optimal amount of new capital.

Economic Review
Federal Reserve Bank of Kansas City
Kansas City, Missouri 64198
April 1984, Vol. 69, No. 4