What Can Price Theory Say about the Community Reinvestment Act?

Robert L. Lacy and John R. Walter

The Community Reinvestment Act (CRA or Act) is credited with funneling billions of dollars in loans and investments to distressed U.S. communities over the last two decades. Passed in 1977 in response to concerns that banks were failing to make loans in declining communities, the Act in essence requires federal bank regulators to assess banks' lending and investment activities and encourage expanded lending and investment in lower-income communities.[1] Its passage was viewed as an important step in curbing lending practices that might discriminate against borrowers in low-income communities and assuring that banks would provide much-needed funding for urban and rural development. And CRA has indeed changed the way bankers think about lending in low-income communities in the United States. With federal regulators looking over their shoulders, bankers today are more careful about their lending and investment practices, particularly to the extent that such practices differ across the often-diverse communities they serve. While it is difficult to quantify the contribution of CRA to community development, some lower-income neighborhoods undoubtedly have benefited from the Act. The percentage of low-income loans in the typical bank's loan portfolio has grown, albeit modestly, in recent years, and many cities have received sizeable grants from banks in support of their community development efforts.

The authors benefited greatly from discussions with Huberto Ennis, Marvin Goodfriend, Thomas Humphrey, Edward S. Prescott, Daniel Tatar, and John Weinberg. The views expressed herein are not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] All depository institutions except credit unions are subject to CRA regulations—i.e., CRA is applicable to commercial banks, savings banks, and savings and loans. Throughout this article the simpler term *banks* will be substituted for *depository institutions*.

Yet, 25 years after its passage, CRA continues to draw criticism and generate controversy. The controversy stems in part from differences of opinion regarding the need for CRA as a tool to counter discriminatory lending practices given that comprehensive fair-lending laws are also in place. Critics of the Act maintain that there is little evidence that banks discriminate by neighborhood or community when lending and that therefore there is little need for legislation that goes beyond the protection offered by laws preventing discrimination based on minority status. Proponents of CRA, however, argue that the Act is needed because in its absence some banks would indeed discriminate based on a borrower's location, and perhaps even completely refuse to serve some neighborhoods, essentially redlining those communities. Many proponents believe that CRA offers minorities additional protection against discrimination even if banks don't redline.

Some proponents of CRA also argue that the Act improves lending-market performance by increasing the information available about lending in low-income areas. If so, then CRA can be beneficial for both borrowers and lenders. Canner and Passmore (1995, 77–79) discuss this view of CRA as a mechanism for correcting for market failure due to externalities and compare this perspective to other views of how CRA affects bank lending. If CRA partially corrects for market failure, then lending markets can become more efficient.

But critics of CRA have suggested that such government intervention in the banking industry reduces the efficiency of credit markets. They argue that this intervention, which in the case of lending allows some low-income borrowers to secure loans at below-market interest rates or on better-than-market terms, can distort credit markets and cause resources to be wasted. There are those who maintain that CRA lending represents a substantial subsidy to low-income borrowers and suggest that direct government aid or government loan programs provide a more efficient means of assisting low-income communities. Lacker (1995, 15–18), for example, is skeptical of the externalities argument, wondering why lending in low-income neighborhoods should be any more susceptible to market failure than lending in more affluent neighborhoods. He argues that funding community development directly out of general tax revenues seems more promising than employing CRA to redistribute income to alleviate the problems of the nation's low-income neighborhoods.

Without denying the benefits of CRA, we identify unique lending costs the Act imposes and analyze how these costs affect credit markets. We argue that CRA enforcement requires some banks to expand their low-income lending to such an extent that unprofitable loans are likely to be made, creating inefficiencies in credit markets. These inefficiencies are created as banks modify lending in response to CRA regulations and examinations. We also explore why CRA continues to survive in today's banking environment, despite the above costs that it imposes on banks but not on banks' competitors. Requirements that

impose higher costs on only one segment of an industry are notoriously difficult to maintain in a competitive environment, and since the passage of CRA in 1977, the financial services industry has clearly become more competitive. Banks must charge customers in some markets higher prices in order to sustain lower, unprofitable prices in other markets. As competitors successfully target these higher-priced markets, however, one would expect the source of funding to eventually dry up. A number of industries—telecommunications, electric utilities, and airlines—that have also been subject to deregulation and increased competition over the last 20 years have seen similar requirements erode as markets became more competitive.

We believe that CRA has survived for 25 years for two reasons: (1) a portion of CRA's costs can be shifted to banking products for which there are fewer alternatives offered by banks' competitors, and (2) CRA's costs are relatively small. Looking ahead, while many community activists support CRA and would like to see it play an even larger role in funding community development, we consider it unlikely that CRA will take on an expanded role because competition imposes very real limits on banks' ability to draw funds from customers to support CRA lending.

1. CRA AND COST SHIFTS

In evaluating the effects of CRA on bank lending, we model CRA as a requirement to expand output of one product of a multiproduct firm. More specifically, this requirement is that a bank's low-income lending equal a fixed proportion of total lending.[2] We also assume that CRA requires banks to expand low-income lending beyond the level that would prevail in the absence of the Act. We begin by discussing a proportion-based requirement in general terms, and then illustrate some of the consequences for costs and output using a hypothetical example of a lawn mower manufacturer. A detailed presentation of the consequences of a proportional CRA lending requirement concludes the section.

A Requirement to Expand Output

Requirements meant to encourage expanded output of one product sold by a multiproduct firm are not unique to banking. For example, regulations adopted

[2] The CRA regulations define four income categories (see, for example, Board of Governors Regulation BB). A low-income individual or family is one with income less than 50 percent of area median income. Moderate income is income that is at least 50 percent and less than 80 percent of area median income. Middle income is at least 80 percent and less than 120 percent of median income. Upper income is 120 percent or more of median income. In our model the phrase *low-income* corresponds to the low- and moderate-income categories in the regulations, while *middle- to high-income* (MHI) corresponds to the two higher-income categories.

in California in 1990 encourage expanded production of environmentally benign motor vehicles. These regulations require that specific percentages of passenger cars and certain light-duty trucks produced for sale in the state by each of the seven largest auto manufacturers be zero-emission vehicles (ZEV). The objective was that ZEVs compose 2 percent of production by model year 1998 and 10 percent by 2003.[3]

In a competitive industry, a requirement to expand output beyond the level chosen by a firm will generally mean that the additional output is produced at a loss. A producer in such an industry normally chooses to produce the quantity at which its marginal costs just equal the price buyers are willing to pay. Since marginal costs of production generally rise with increasing output, and because in a perfectly competitive market a single price exists for the firm's entire output, profits are greatest at a level of output where marginal costs equal the market price. If production is below this level, the cost of producing an additional unit is below the price buyers are willing to pay, so the firm will continue to expand output in order to increase profits. The producer will not choose to expand output beyond the point at which marginal costs equal price because it would suffer losses on this additional output.

If the firm produces two goods and a regulation requires that the production of one good be expanded to be at least a certain percentage of total production of both goods, then a penalty cost is incurred for the production of the second good. The penalty cost consists of the losses from additional sales of the first good, for these losses must be incurred when additional quantities of the second good are sold. The imposition of a penalty cost matters little if all firms are subject to the same regulation. But if some firms are free of regulation, such firms will gain market share in the second good. Unregulated firms do not sustain a penalty cost, so they can charge a lower price and acquire some of the regulated firms' customers.

An Example

To illustrate how a market might respond to a regulation requiring expansion of output, consider a simple hypothetical example of a lawn mower manufacturer that sells both electric- and gasoline-powered lawn mowers in a competitive market. Legislators in this example impose a requirement that all manufacturers of lawn mowers produce relatively more electric-powered models in order to protect the environment.

The supply curve for electric lawn mowers identifies the quantity of such mowers the firm will produce at various prices. As the market price increases,

[3] California regulations have since been modified to allow manufacturers to meet the percentage requirements with a combination of ZEVs and very low emission vehicles. ZEV mandates were adjusted to allow manufacturers to receive partial credit for extremely low emission vehicles that were not pure ZEVs. See California Environmental Protection Agency (2001).

the manufacturer is able to recover higher manufacturing costs, and the number of electric-powered mowers it is willing to produce increases. The number of electric mowers produced at a given price depends on the firm's marginal cost of production. This cost rises as the firm produces more. The demand curve is horizontal because the firm in our example operates in a perfectly competitive market—it is only a small manufacturer in a large market for electric mowers. It cannot affect the market price for electric mowers because its portion of the market is almost insignificant. Given its demand and supply curves, the firm will produce a quantity of mowers identified by the intersection of supply and demand curves, or for our example, 100 electric mowers. The firm earns its highest profit by manufacturing 100 electric lawn mowers.

The manufacturer also chooses to produce the quantity of gas mowers determined by the intersection of supply and demand curves for this product. The equilibrium price and output quantities differ, however, in the gas and electric mower markets. In equilibrium, we will assume the manufacturer produces 100 electric mowers and 900 gas mowers, for a total output of 1,000 lawn mowers.

Suppose legislators decide that electric lawn mowers are more environmentally friendly and that their use should be encouraged. Legislation is therefore passed requiring that electric mowers account for 15 percent of total mower production rather than the current 10 percent. The legislative requirement leads the manufacturer to increase its output of electric mowers beyond 100. The manufacturer's marginal cost exceeds marginal revenue when it produces more than 100 mowers. With a regulation requiring the manufacturer to expand production of electric mowers to 150, the manufacturer suffers losses on each of the last 50 electric mowers produced.

Because of this requirement, the firm's supply curve for gas mowers will shift leftward to a position above what it would be without the requirement. Why does the requirement raise the manufacturer's cost of producing gas mowers? Because the number of electric mowers produced is determined by the manufacturer's production of all lawn mowers, both electric and gas. As a result, when the manufacturer wishes to expand its production of gas mowers, it must also expand its production of electric mowers in order to comply with the electric mower output requirement. So when a firm increases gas mower production, it is required to produce electric mowers at a loss.

The output requirement will affect manufacturers to different degrees because their cost curves vary. Some firms are especially skillful or well equipped for producing electric mowers; others are more skillful at gas mower production. For example, a manufacturer located in a city with battery and electric-motor suppliers can be expected to have relatively low costs for electric mower manufacture because shipping costs for these mower components will be low. For manufacturers especially equipped to make electric mowers, one would expect that production of electric mowers as a percentage of all mowers would

be well above average in the absence of an output requirement, or above 10 percent in our example. For such manufacturers, electric mowers might represent 25 percent of their total mower output. The legislative requirement that electric mowers compose 15 percent of all mowers would not result in a cost increase for these manufacturers—the requirement is nonbinding for them.

CRA in a Competitive Banking Environment

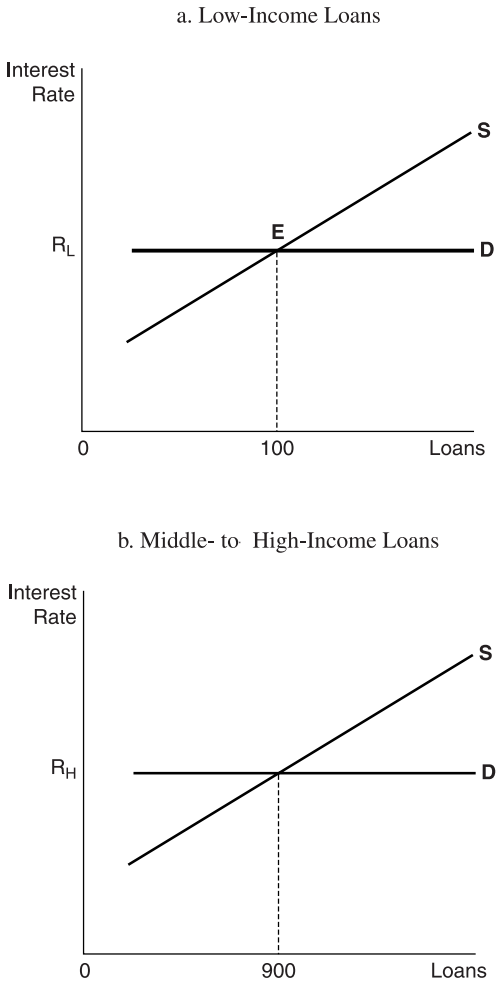
CRA requirements impose a mandate to expand lower-income community lending analogous to an automobile industry requirement promoting electric vehicles and our hypothetical example of a requirement to expand electric mower production. While CRA also promotes community development investments and service, banks focus their CRA compliance efforts on lending; therefore, our analysis of CRA does so as well. We begin with a look at the effect of CRA on an individual bank assumed to be operating in a competitive banking environment. A requirement to expand lower-income lending will produce an increase in the effective cost of other types of lending. This increase in banks' effective lending costs means that nonbank lenders gain a competitive advantage. We then broaden our analysis to cover the financial services industry as a whole, including nonbank competitors. As will be shown, federal bank regulators lose much of their ability to influence the number of loans made to low-income communities when financial institutions not subject to CRA lending requirements emerge.

CRA and One Bank

When a bank operates in a perfectly competitive market, its supply and demand curves for loans made in low-income communities appear as shown in Figure 1a. The bank's supply curve identifies the number of loans it is willing to make at various interest rates. As the market rate of interest increases, the bank is able to recover higher costs of lending, and the number of loans the bank is willing to make increases. The number of loans it will make at a given interest rate, say R_L, depends on the bank's marginal cost of lending.[4] The bank's marginal cost of lending rises because to increase its lending it must, for example, spend more to attract qualified loan officers from other fields or make increasingly higher outlays for marketing to attract additional borrowers. The demand curve D in Figure 1a is horizontal since the bank is only a small part of a huge lending market. Given demand curve D and

[4] While in Figure 1a, and throughout the article, the supply curve is represented as equivalent to the marginal cost curve, this is a simplification. A firm's output also depends on its average variable cost curve. The supply curve would be equivalent to the marginal cost curve only at market prices above the intersection of the marginal and average variable cost curves.

Figure 1 Single Bank, without CRA



supply curve *S*, the bank will make 100 low-income loans—where the supply and demand curves intersect at point *E* in Figure 1a.

Supply and demand curves for lending in other markets will be similar and are represented in Figure 1b. We define other lending as loans to all borrowers except those in low-income communities. We will call them middle- to high-income (MHI) borrowers.[5] Note that in this market too the bank chooses to

[5] For expository purposes, we represent the supply and demand curves of all other loan markets by the supply and demand curves of Figure 1b, even though each type of loan has its own supply and demand curves.

make the number of loans determined by the intersection of supply and demand curves. The equilibrium interest rates and loan quantities differ, however, in the two markets. In our example, in equilibrium the bank makes 100 low-income loans and 900 other loans for a total of 1,000 loans.

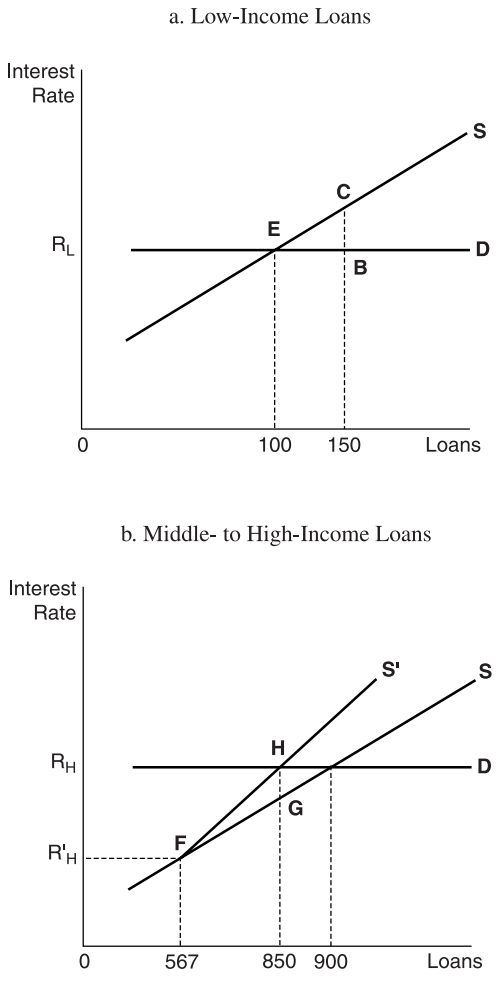
What effect do CRA requirements to expand lending have on the bank in our model? If binding, CRA requirements oblige the bank to make more low-income loans than it would if there were no requirements. Without CRA requirements, the bank chooses to make 100 low-income loans representing 10 percent of its total loans. Binding regulations, however, require the bank to extend more low-income loans in order to merit a satisfactory CRA rating. While its profit-maximizing output is 100 low-income and 900 other loans, binding CRA regulations require the bank to make low-income loans equal to, say, 15 percent of all loans. Although CRA examiners do not have explicit minimum percentage guidelines that banks must meet, they do expect bank lending to low-income individuals, in the absence of extenuating circumstances, to be roughly proportional to the low-income population. More will be said about examiner expectations later in Section 2.

Figures 2a and 2b illustrate the case in which CRA requirements are binding; in our example, the bank in Figure 1a must make more than 100 low-income loans. In Figure 2a the bank's marginal cost exceeds marginal interest earnings when it extends more than 100 low-income loans. With a regulation requiring low-income loans equal to 15 percent of total loans, the bank makes 150 loans, but it suffers losses equivalent to area BCE on low-income loans between 100 and 150. The bank would like to charge borrowers an interest rate sufficient to cover its higher costs of making each additional loan. The bank's low-income customers, however, are only willing to borrow at the interest rate of R_L.

Supply and demand curves for the bank's other lending market are shown in Figure 2b. The S curve in Figure 2b is the bank's cost of making MHI loans, without CRA requirements. The S' curve is the bank's cost of making MHI loans with a binding CRA requirement. The S' curve lies above the S curve for most of its range, reflecting the additional cost of a requirement to make at least some unprofitable low-income loans. We define unprofitable loans as those for which marginal cost exceeds marginal revenue. The S' curve begins to diverge from S at the level of MHI lending for which CRA requirements become binding.

In our example, with a CRA requirement that low-income loans account for 15 percent of all loans, the point of divergence is 567 MHI loans. If the rate of interest is such that the bank wishes to make only 567 loans—R'_H in Figure 2b—it will not be required to make any low-income loans for which costs exceed interest earnings. At 567 MHI loans the bank can make 100 low-income loans [$100/(567 + 100) = 15$ percent], exactly the number of low-income loans it would make without a CRA requirement.

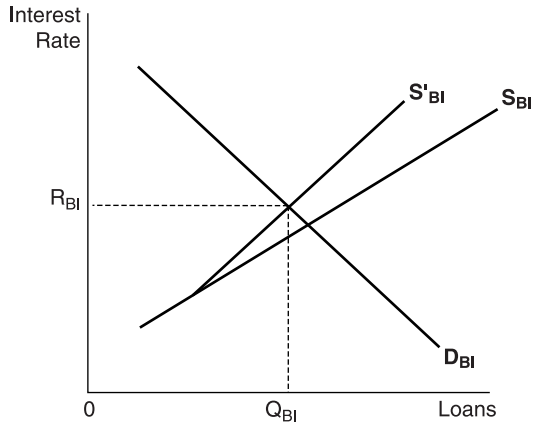
Figure 2 Single Bank, with CRA



If the market rate of interest for MHI loans is above R'_H, the bank will find its most profitable combination of MHI and low-income lending to include more than 567 MHI loans and, because of CRA, more than 100 low-income loans. Since the S' curve represents not only the costs of making MHI loans, but also the losses generated by whatever quantity of low-income loans are called forth by the profit-maximizing level of MHI loans, it represents the bank's most profitable levels of high-income lending, given CRA.[6]

[6] In Figures 2a and 2b we have assumed that the bank will continue to make a total of 1,000 loans (150 low-income and 850 MHI) after CRA is imposed. By imposing an additional

Figure 3 Middle- to High-Income Loans: Banking Industry with CRA



Banking Industry and Nonbank Competition

Banks today compete with nonbanks in most banking-product markets (see *Increased Competition from Nonbanks* in Table 1). For example, securities brokers offer money market mutual funds, for some customers an attractive alternative to holding a deposit account at a bank. Likewise, finance companies offer consumer and business loans, and nonbank mortgage lenders offer residential mortgage loans, all in competition with similar loan products offered by banks. These nonbank competitors are free from CRA requirements. With nonbank competitors present, regulators face limits on their ability to employ CRA requirements to expand low-income lending; limits are present but less severe in the absence of nonbanks.

Figure 3 depicts the banking industry's cost curves for MHI loans—S_{BI} without CRA and S'_{BI} when subject to CRA, with no nonbanks present. These curves represent the horizontal summation of individual bank MHI cost curves like those shown in Figure 2b. Bank cost curves vary and are a function of many factors, including bank size.[7] The demand curve for MHI loans is given by D_{BI} in Figure 3. Unlike the single bank's demand curves in Figures 1 and 2, the industry demand curve is downward sloping. The market interest rate,

cost on banks, CRA may instead result in a decline in total bank lending below 1,000 loans. Such a decline in total lending does not change the conclusion that a low-income requirement raises the effective cost of high-income lending.

[7] Only banks with minimum average cost at or below the market price can earn a profit, so only such banks will be present.

given CRA, is R_{BI}, determined where the S'_{BI} and D_{BI} curves intersect. The quantity of MHI loans made is Q_{BI}.

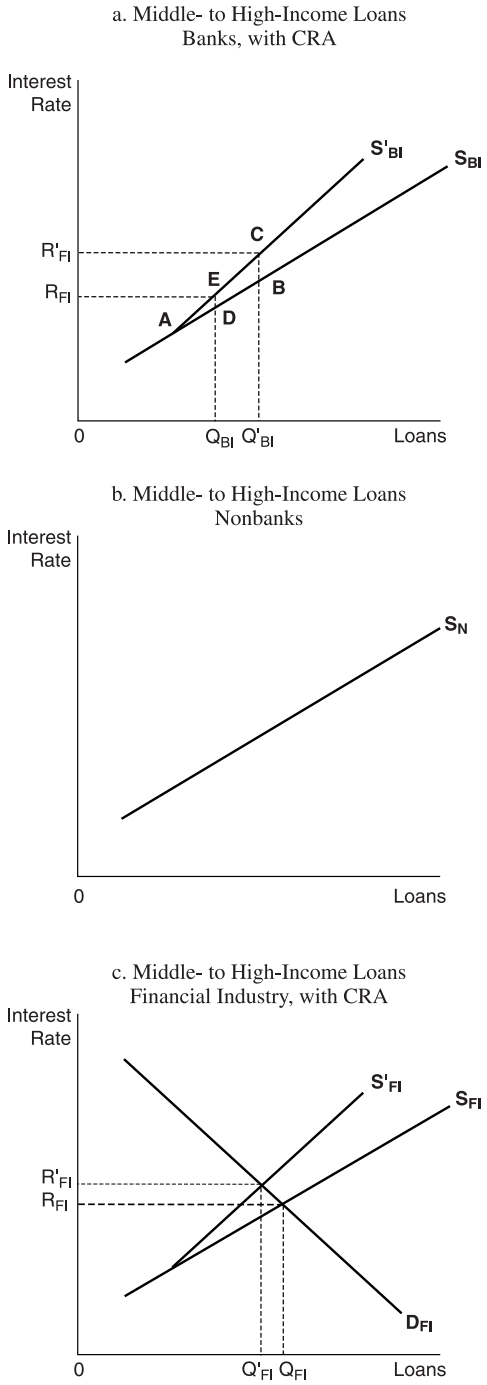
Nonbank competitors are introduced in our analysis in Figures 4a through 4c. Figure 4a depicts the banking industry's supply curves, which are like those in Figure 3. Figure 4b shows the supply curve, S_N, for nonbanks. We assume that nonbanks have the same cost structure as banks but are not subject to CRA, so the S_N curve is an exact replica of the S_{BI} curve in Figure 4a.[8] The financial industry supply curve S_{FI} in Figure 4c is the horizontal summation of the nonbanks' supply curve S_N and the banks' supply curve S'_{BI} reflecting CRA costs. The second supply curve, S'_{FI}, in Figure 4c, represents a summation of bank and nonbank supply curves in a theoretical case in which nonbanks are subject to CRA and have supply curves like S'_{BI} in Figure 4a. In this scenario, all firms incur higher costs because of CRA and thus S'_{FI} lies above S_{FI}. By comparing interest rates and loan quantities when the supply curve is S'_{FI} with rates and quantities when the supply curve is S_{FI}, we can illustrate changes that result when firms not subject to CRA are introduced. The industry price and quantity are found at the intersection of the D_{FI} and S_{FI} curves in Figure 4c—interest rate R_{FI} and quantity Q_{FI}. Because S_{FI} lies below and to the right of S'_{FI}, R_{FI} is lower than R'_{FI}.

At the market interest rate of R_{FI} in Figure 4a, banks will make Q_{BI} loans, as determined by the intersection of the S'_{BI} curve and the market interest rate of R_{FI} that was determined in Figure 4c. In contrast, if banks face only competitors that are subject to CRA, the market interest rate is R'_{FI} (as determined in Figure 4c), and the quantity of bank loans is higher at Q'_{BI} in Figure 4a. The lower market interest rate R_{FI} reflects the presence of nonbank competitors, not subject to CRA, and causes banks to make fewer loans. In addition, the size of the CRA subsidy is lower in the presence of nonbank competitors. Without these nonbanks, the subsidy is the ABC triangle in Figure 4a; with nonbank competitors, the subsidy is the smaller triangle ADE.

This CRA subsidy represents the transfer of costs from low-income to high-income borrowers. The subsidy is due to CRA's lending requirement, which results in low-income borrowers receiving loans at prices below banks'

[8] For expository purposes we have assumed that the cost structures of banks and nonbanks are the same. Such an assumption means, however, other things being equal, that banks would immediately convert to nonbank status if they were able to do so costlessly and could avoid CRA by converting. In practice, banks may not want to convert because they hold certain advantages as banks. Some observers maintain that transactions accounts offer banks a low-cost source of funds. While some nonbanks offer accounts with certain transaction features, accounts at nonbanks typically do not provide all of the payments features of a bank checking account. Further, banks may have cost advantages because of special lending skills, access to information not available to nonbanks, or access to what some argue is underpriced deposit insurance. A lower cost curve may allow banks to continue to compete successfully against nonbanks despite the costs that CRA imposes on high-income lending, so banks would not choose to convert to nonbank status. In Section 4 we provide further discussion of possible bank cost advantages.

Figure 4 Financial Services Industry



marginal cost of making the loans. While low-income borrowers benefit, high-income borrowers that receive a bank loan pay more than they would have paid in the absence of the CRA requirement.

Implications for Expansion of CRA

In addition to reducing a bank's MHI lending, the presence of nonbank competitors not subject to CRA is likely to diminish the effectiveness of CRA as a tool for expanding lending to low-income borrowers. Specifically, an increase in the stringency of CRA requirements in the presence of nonbank competitors will result in a smaller dollar amount of funds available for low-income lending.

The following example illustrates this constraint on regulators' ability to expand low-income lending. Assume first that banks are the only type of financial institution in the industry. With an S'_{BI} only slightly above S_{BI} in Figure 4a, meaning a CRA requirement causes banks to make only a small number of low-income loans above the quantity they would make without the requirement, the subsidy is small. The size of the subsidy will initially increase, but as the CRA requirement becomes more stringent, the subsidy will decline. In graphical terms, as the S'_{BI} curve swings counterclockwise and away from S_{BI}—that is, CRA requirements become more stringent—the area of the ABC triangle increases at first, then later decreases.[9] The area will shrink as the S'_{BI} curve becomes ever more vertical because the BC side of the triangle draws near point A of the triangle.

But with nonbanks competing with banks and attracting their customers, the market rate of interest will rise less than it would in the absence of nonbanks for any given increase in CRA stringency. The smaller increase occurs because the nonbanks' supply curve is unaffected by the increased stringency. As a result, S_{FI} in Figure 4c, derived by summing banks' S'_{BI} curve from Figure 4a and nonbanks' S_N curve from Figure 4b, does not swing as far as the S'_{BI} curve because its swing is damped by the stable nonbank curve. Thus, the interest rate changes little. Because the market interest rate rises less, the height of the BC side of the triangle in Figure 4a is also smaller in the presence of nonbanks. The ability of regulators to enlarge the size of the subsidy—i.e., the size of the triangle—is diminished by the existence of nonbanks.

Policymakers interested in assisting communities by expanding lending therefore face very real constraints. They could impose stricter requirements on banks in order to generate greater CRA lending benefits. For example, they might require that low-income loans account for 20 percent of total loans,

[9] In addition, as CRA requirements become more stringent, point A in Figure 4a shifts along the S_{BI} curve toward the origin. Point A represents the level of lending at which CRA requirements become binding. In this graph this point shifts toward the origin as the required minimum ratio of low-income loans to total loans is increased.

rather than the 15 percent in our earlier example. However, stricter requirements eventually produce a declining subsidy as the quantity of MHI loans made by banks declines and nonbank competitors acquire more of the banks' customers.

2. CRA IN PRACTICE

The graphical analysis in Section 1 captures the essential elements of CRA's low-income lending requirement.[10] However, there are two critical assumptions that need further elaboration. The analysis assumes that regulators expect banks to match or exceed a specific ratio of low-income lending to total lending. While this ratio is a simple statistic and only one of several indicators considered in a fairly complex CRA lending review, it provides regulators with an important benchmark of acceptable lending practice. The graphical analysis also assumes that absent CRA, banks would extend all the profitable low-income loans possible, and it thus implies that CRA causes banks to make unprofitable loans. Since the banking industry is currently quite competitive, and there is little evidence of discrimination based on geographical location, we believe our assumptions are reasonable in today's banking environment. These assumptions are explained in more detail below.

CRA Enforcement and Lending Ratios

The Community Reinvestment Act requires regulators to "assess the [bank's] record of meeting the credit needs of its entire community, including low- and moderate-income neighborhoods," and to "take such record into account" when evaluating a bank's merger application (12 U.S.C. 2901, sec. 804). Federal bank regulators make their assessment, assigning a bank one of four ratings: (1) substantial noncompliance, (2) needs to improve, (3) satisfactory, and (4) outstanding. Banks that receive low ratings are likely to have difficulty convincing regulators to approve merger applications. Thus, banks anticipating future mergers will tend to make at least enough low- and moderate-income loans to ensure that a merger will not be denied, which for some banks will likely involve an expansion in lower-income lending beyond the level they would choose in the absence of the Act. While the CRA regulations require examination of a bank's record of meeting the credit needs of its "entire community," the intent is clearly to promote more lending in low- to moderate-income communities. For example, the Federal Reserve's Regulation BB

[10] While our article focuses on CRA's low-income lending requirement, CRA enforcement also seeks to encourage expanded lending within banks' local markets as well as expanded small-business and small-farm lending. Measuring such lending by each bank and basing the bank's CRA rating, in part, on these types of lending provide the encouragement.

specifies "a very poor geographic distribution of loans, *particularly to low- or moderate-income geographies*" (emphasis added) as one of the factors that can cause a bank to receive a rating of substantial noncompliance. The Act also requires examiners to evaluate banks' efforts at meeting the "credit needs" of the community, which includes lending and investments, but the CRA rating is most dependent on lending.[11]

The discussion in Section 1 suggests that CRA examiners require banks to achieve a certain minimum ratio of low-income lending to total lending. And in practice, CRA examiners do calculate the ratio of a bank's low-income lending to total lending and compare that figure to the proportion of low-income population to total population in the bank's assessment area as a rough benchmark of appropriate lending procedures.[12] Examiners recognize, however, that there might be legitimate reasons why a bank's low-income lending would not be proportional to the low-income population. For example, examiners will take into consideration the prevalence of rental housing in an area when examining a bank's mortgage lending patterns, since the demand for mortgage lending would likely be lower in areas with a high proportion of rental properties. Likewise, adjustments are made for areas where income levels are extremely low or unemployment is high, since individuals with very low incomes or those that are unemployed are unlikely to be willing or able to borrow.

Still, in general, given these adjustments, if low-income households compose 20 percent of the population in the bank's assessment area, then examiners expect approximately 20 percent of the bank's total loans to be made to low-income households. Barring unusual circumstances, if the bank increases its total loans, it will be expected to increase its low-income loans proportionally. Our assumption of a fixed proportion of low-income loans to other loans in the graphical analysis reflects this regulatory approach to CRA enforcement.[13] Whenever the bank considers making a loan that does not qualify for CRA credit—say, a loan to an individual in a high-income community—it will take into account that it will also be required to add to its low-income lending.

---

[11] Large banks are rated based on three tests: a lending test, an investment test, and a service test. On each of the three tests, the bank receives a rating. The lending test, however, is predominant. A bank's numerical score on the lending test receives twice the weight of scores on either the investment or service tests before these three scores are summed to obtain the composite CRA score (FFIEC 1997, 15–16). Small bank CRA ratings generally depend only on lending, though investment and service activities can earn the small bank extra credit toward its final rating.

[12] The shorter term *low-income* will be used in place of the term *low- and moderate-income* throughout the remainder of the article.

[13] CRA performance evaluations for individual banks are available from websites maintained by the federal agencies that regulate banks. These evaluations provide a fairly detailed explanation of the factors examiners consider important in the determination of a CRA rating. An evaluation typically provides a calculation of the bank's proportion of low-income lending to total lending, which is compared to the proportion of low-income households in the assessment area. These evaluations often note when the two proportions are very different.

Does CRA Encourage Unprofitable Lending?

Without a doubt CRA has increased low-income lending (Litan et al. 2001). If in the absence of CRA, however, banks would have extended all profitable loans possible, then it is likely that the Act has resulted in at least some unprofitable loans.[14] Given the competitive nature of today's banking industry, it is hard to imagine that banks would overlook many opportunities to make additional profitable loans. When CRA was enacted in 1977, substantial regulatory restrictions on entry and pricing allowed some banks to exercise monopoly power in local banking markets. In such cases bankers would be expected to exploit monopoly power by restricting lending. If banks did restrict loans, it seems plausible that they would tend to prefer higher-income lending at the expense of low-income lending. Today, however, there are fewer entry and pricing restrictions in the banking industry, so monopoly power is likely to be quite limited. (See Table 1 for a review of competition in banking.)

Of course, one could argue that even in competitive markets, some banks might nevertheless prefer to avoid lending in certain low-income neighborhoods because of a bias against minorities predominant in a neighborhood or because of a lack of experience in lending to low-income borrowers, for example. Although not conclusive, a number of studies suggest there is little evidence of such failure to lend to individuals in low-income communities.[15] So, once again, in the absence of CRA we would expect banks to make all possible profitable loans, and if our expectation were met, CRA pressures to extend lending would produce unprofitable loans.

CRA may also be causing more low-income lending than is profitable because regulations generally do not make exceptions for differences in business strategies, market niches, or capabilities. Some banks are simply less adept at lending in low-income communities than others. For example, certain banks are especially skilled at making loans and gathering deposits from individuals in high-income communities or providing personal banking services to high-income individuals and make these services a large part of their deposits and loans. For such banks, CRA requires an expansion of low-income lending beyond the profit-maximizing equilibrium, acting as a tax on such banks' high-income lending. While one might imagine that there are few niche banks that specialize in personal banking, all banks will be on a continuum from those most capable at low-income lending to those most capable at high-income lending.[16]

---

[14] See, for example, Gramlich (2001).

[15] For reviews of the literature on geographic discrimination, see Lacker (1994, 6–9) and Evanoff and Segal (1996, 24–25).

[16] In some cases, a bank can improve its CRA rating not only by making low-income loans but also by purchasing such loans made by other lenders. If a niche bank, for example, is able to purchase low-income loans from a lender specializing in making such loans, then the cost imposed by CRA on the niche bank might be lowered.

Recently the Board of Governors of the Federal Reserve System surveyed banks in order to quantify the profitability of CRA lending. Profitability in the Board study was measured in terms of accounting profits, that is, return on equity. The Board asked respondents to measure profits based on revenues and costs associated with such items as overhead and the servicing, pricing, delinquency, and prepayment of CRA loans (Board of Governors 2000b). This definition of profitability differs from that used in Section 1, where profitability is defined as the difference between marginal revenue and marginal cost. The Board survey found that bank loans that qualify for CRA credit are profitable for most banks, with many institutions reporting that they are as profitable as comparable non-CRA loans. However, a high proportion of institutions reported that CRA loans are less profitable (44 percent of institutions in the case of home mortgage lending), and almost none reported them as more profitable.

Yet, to answer our question regarding unprofitable lending, we must move beyond an examination of the average profitability of all CRA loans and examine only those low-income loans that are made simply because of the presence of the Act. In other words, in the absence of CRA, banks would make a certain number of low-income loans, but only if these loans were expected to be profitable. The Act may require banks to make additional loans that are unprofitable. The average profitability of all low-income loans might be positive even if banks make unprofitable loans in response to CRA. So while the Board study does not answer our question, the fact that CRA loans are often less profitable suggests that unprofitable loans may be lowering the average. Given the competitive nature of the banking industry, we believe it is reasonable to assume that CRA encourages banks to make unprofitable loans.

Still, no industry is perfectly competitive. And the banking industry has some characteristics suggesting that it is no exception, despite the fact that its competition is quite strong. For example, observers have noted that switching costs may be significant in banking, meaning that it is costly for a consumer to change banks to take advantage of a superior interest rate or lower fee (Rhoades 2000). Switching costs might be significant for a consumer because shifting transaction accounts to a different bank would require the consumer to contact his or her employer, utilities, and mortgage company to modify direct deposit and direct withdrawal arrangements. As a result of switching costs, banks might exercise some pricing power over existing customers, providing a source of supracompetitive profits. If banks enjoy a measure of monopoly power in pricing, then some of the additional low-income loans CRA will induce are not necessarily unprofitable. Nevertheless, even if banks do not lose money on these additional loans, the requirement to make additional loans will lower profits below the bank's profit-maximizing level of lending, and the economic analysis found in Section 1 of this article will be unaffected.

Because profits are diminished for each additional CRA loan beyond the profit-maximizing quantity, banks will take account of the lost profits when deciding to make an MHI loan. Accordingly, a CRA requirement will mean an increase in the marginal cost of making MHI loans, or equivalently a leftward shift in the supply curves for bank MHI lending in the graphs.

3. EFFICIENCY AND EQUITY

While CRA has expanded lending in low-income neighborhoods, any benefit derived from the expansion may be offset by costs resulting from less efficient credit markets. Furthermore, some costs CRA imposes may be borne disproportionately by low-income individuals. The expansion of lending to low-income communities distorts credit decisions on loans made to individuals and businesses in both low- and high-income communities. At the same time, CRA can place banks at a cost disadvantage in relation to nonbanks, which are not subject to CRA, further distorting the market. Banks will attempt to shift the costs of making unprofitable loans to those customers who have the fewest alternatives to bank products.[17] Since low-income individuals frequently have few alternatives, they will tend to bear more than their share of such costs.

Efficiency Losses

Efficiency losses occur for several reasons. Imagine that one of the 50 borrowers discussed in Section 1 is a small-business owner in a low-income community who plans to undertake a project that will produce a rate of return equal to R_L in Figure 2a. Consequently, this borrower is willing and able to pay an interest rate as high as R_L to borrow to fund the project. But the bank, having already made 100 loans, finds that its cost of making this 101st loan is greater than R_L, as indicated by the S curve above R_L. The 101st loan is inefficient and results in a wasteful misallocation of resources inasmuch as costs exceed benefits as measured by the loan's rate of return. Because of CRA, the project is undertaken by the business owner, even though the project's economic benefits—as measured by the income the owner earns on the project—are smaller than the costs of the resources employed to fund it. This project is funded, but the resources could be employed elsewhere in a project that delivers benefits exceeding resource costs.

To illustrate the second source of inefficiency, imagine an owner of a business in a high-income community. This business owner, with a project capable of earning more than R_H, would have received a loan if the bank's

[17] For an example of cost shifting in banking see Fama (1985), who discusses banks shifting their reserve requirement costs to customers with the fewest alternatives.

cost curve had been S in Figure 2b. Instead, however, the loan is denied. The bank will make 900 loans if its cost curve is S, but only 850 loans if S'. As a result, the economic benefits from the project in the high-income community are unrealized.

A third type of inefficiency arises because CRA requirements result in less business for banks and more business for other, less efficient providers. While in Section 1 we assumed that costs of making unprofitable loans are borne by high-income borrowers, banks may shift some of these costs to depositors as well. The costs imposed by CRA, shifted to depositors in the form of lower interest rates paid by banks on deposits, means some individuals will elect to hold their transactions accounts in other places, such as money market mutual funds (MMMFs). Yet using a MMMF for payments purposes can be far less convenient than using a bank account and may result in an inefficient use of the individual's resources. Firms offering MMMFs typically do not have as many branches as banks, nor do they offer widespread ATM networks. Further, such firms often impose a minimum check size requirement on MMMFs, making such accounts more difficult to use for day-to-day purchases. While the consumer is willing to tolerate these inconveniences to earn the higher rate paid by the MMMF, the inconveniences would have been avoided if CRA had not led the bank to pay lower rates.

Furthermore, observers have often argued that banks can be especially effective lenders because they have access to information about borrowers' finances that allows them to better assess creditworthiness. For example, because bank borrowers often hold transactions accounts with the same bank, banks have unique access to information about the financial health of borrowers. But if such transaction deposit relationships are severed because banks lose customers by paying lower interest rates on deposits as a result of CRA, then this information will be lost. Such information is valuable because it lowers the cost of making loan decisions. If lost, the economy's resources are wasted, either because more resources are consumed by making lending decisions or because less creditworthy projects are funded.

Equity

While CRA is intended to benefit low-income communities by ensuring that banks do not overlook them, the law may at the same time impose additional costs on low-income individuals. Some of the costs of such lending will tend to be shifted to depositors in the form of lower interest rates or perhaps higher fees on transactions deposits. In effect, individuals holding transactions accounts are taxed so that more CRA loans can be made.

Such a tax may fall more heavily on low-income and minority individuals. High-income individuals hold a smaller percentage of their wealth in the form of checking account deposits than do low-income, and whites hold a smaller

percentage than do nonwhites (Davern and Fisher 2001, Tables 1 and H). While CRA may produce additional loans for such individuals, it also tends to tax them.

4. WHY HAS CRA SURVIVED?

Why has CRA survived despite the presence of nonbank competitors who can offer lower prices since they avoid cost shifts inherent in CRA's low-income lending requirements? After all, in a number of other regulated industries, most notably telecommunications and airlines, the entry of unregulated competitors made it difficult to pursue social objectives that require firms to shift costs from one customer group to another.

In the telecommunications industry, regulators were forced to slash telephone-rate subsidies as competition became more intense and deregulatory policies were implemented. Kahn (1990, 343) argues that during the 1980s the prices of long-distance service calling and basic residential service were brought closer to their respective costs. He provides as evidence an increase in the local telephone charges component of the Consumer Price Index (CPI) and a decline in the average price of long-distance calling from December 1983 to December 1989. Local telephone charges rose 19.3 percent in real terms, while average long-distance charges fell 44.5 percent interstate, and 24.1 percent intrastate, respectively, during the period. Temin (1990, 350) cites telephone price data from the CPI for 1977–1987 that show a sharp rise in the ratio of local to interstate telephone rates during the post-AT&T divestiture period of 1983–1987 and concludes that cross subsidies from long-distance to local calls were reduced but not eliminated.

In the airline industry, average airfares fell and fare subsidies diminished when the industry was deregulated in 1978 and prices began to be set in competitive markets. Prior to deregulation, fares on longer and more heavily traveled routes had been too high relative to costs, while fares on shorter and less heavily traveled routes had been too low; in effect, heavily traveled routes subsidized less-traveled routes (Joskow and Rose 1989, 1469). Airfares were declining prior to deregulation and the entry of new competitors, but the decline in real airfares was quicker and larger once the industry was deregulated (Winston 1998, 100). Morrison and Winston (1998, 484) estimate that average airline fares are approximately 33 percent lower in real terms since deregulation. But declines in airfares at airports in smaller communities—those designated as small hub or nonhub airports by the Federal Aviation Administration—were consistently smaller than declines in fares at airports in larger communities during the post-deregulation (1978–1996) period (Morrison and Winston 1997, 43). Despite continued interest in maintaining low rates

at smaller community airports, competition today will not allow the subsidies that could make this possible.

If banking followed the trend in the telecommunications and airline industries, one would expect CRA's influence on bank lending to have declined, but it has not. Two factors may help to explain CRA's resilience. First, banks maintain some competitive advantages compared to their unregulated competitors despite aggressive competition. These advantages allow banks to shift costs to certain bank customers and hold at bay nonbanks, which would otherwise lure bank customers with lower prices and cause CRA's low-income lending requirements to collapse. Second, as a practical matter, CRA low-income lending requirements may not impose large costs. Both factors are further discussed below.

CRA Costs Can Be Shifted If Banks Have Competitive Advantages

Among financial institutions, banks are unique in offering transaction deposits and widespread branch facilities to provide convenient deposit and withdrawal. While there are nonbank alternatives to transaction deposits—money market mutual funds for example—for most individuals the alternatives cannot completely substitute for a bank deposit account. According to the 1998 Board of Governors Survey of Consumer Finances, 90 percent of households have a bank transactions account, while only 16 percent have any kind of mutual fund (Kennickell, Starr-McCluer, and Surette 2000, 11). Because nonbank alternatives offer only imperfect substitutes for bank deposits, banks have greater leeway to charge higher prices for these accounts without incurring a loss of customers to nonbank competitors. As a result, some of CRA's cost of making unprofitable low-income loans can be shifted to holders of transaction accounts with little loss of these customers to nonbank competitors. As long as such shifts are possible, banks have less incentive to lobby for repeal of CRA.

CRA's costs might also be shifted to small-business borrowers. Observers argue that banks may hold a competitive edge in lending to small businesses because of long-standing relationships with these borrowers. The special lending skills that bank loan officers have developed over the years and the credit information that is obtained through long-standing relationships are difficult for nonbanks to acquire. While this competitive advantage may erode, it could be some time before nonbanks are on an equal footing with banks in the quality of lending services they offer.

Table 1 The Growth of Competition in Banking

Banking industry competition became more intense starting in the late 1970s for two reasons. First, the banking industry itself became more competitive as entry barriers were dropped, branching restrictions were removed, and interest rate restrictions fell. Second, banks faced mounting competition from nonbank competitors.

Barriers to Entry Fall

- After the massive bank failures of the Great Depression, fairly strict requirements were imposed on the granting of bank charters.
- The Comptroller of the Currency (Comptroller) denied applications for national bank charters when it determined that existing banks already adequately served markets. State banking authorities operated in a similar manner.
- In October 1980 the Comptroller ended its policy of assessing the competitiveness of markets when making charter decisions.

Branching Restrictions Fall

- Federal and state restrictions on banks' ability to branch were an important feature of the U.S. banking environment throughout the 20th century. Restrictions protected existing banks from competition.
- In addition to restrictions on bank branching, the ability of bank holding companies (BHCs) to operate across state lines was also restricted. The Bank Holding Company Act of 1956 largely prohibited bank holding companies from owning banks outside the BHC's headquarters state, but included a provision allowing ownership of banks across state lines if legislation in the non-headquarters state specifically provided for such rights. No states had such legislation.
- Interstate banking restrictions began to fall when, in 1978, Maine became the first state to pass legislation to allow BHCs headquartered in other states to purchase banks in its state. Other states followed suit, so that by 1990 all but four states allowed cross-border purchases—though interstate branching remained largely prohibited.
- In the early 1980s states began to remove restrictions on in-state bank branching.
- The Riegle-Neal Interstate Banking Act of 1994 largely eliminated restrictions on interstate branching.

Interest Rate Ceilings Fall

- The Banking Act of 1933 prohibited the payment of interest on checking accounts and authorized the Federal Reserve to regulate interest rates on time and savings deposits.
- Interest rate ceilings were initially set well above the interest rates banks were paying.
- Beginning in the mid-1960s the situation changed. Ceilings were set below market rates beginning in mid-1966 and generally remained below market rates until ceilings were eliminated in the mid-1980s. Ceilings were viewed at the time as a means of enhancing the flow of loans to mortgage borrowers. In effect, when market interest rates rose above the ceilings, depositors cross-subsidized mortgage borrowers.
- Bank depositors responded by moving their funds to money market mutual funds (MMMFs), and these funds grew rapidly in the 1970s.
- In March 1980 Congress responded by passing the Depository Institutions Deregulation and Monetary Control Act, which phased out all interest ceilings on savings and time deposits and authorized banks nationwide to pay interest on a new type of checking account, the Negotiable Order of Withdrawal (NOW) account. NOW accounts had previously only been available in certain states.

Table 1 The Growth of Competition in Banking (continued)

Increased Competition from Nonbanks

- Companies offering MMMFs compete with banks for consumer and business savings and checkable deposits. MMMFs gained prominence in the late 1970s. By the end of 1999, MMMF balances amounted to $1.46 trillion. This compares to $4.54 trillion in deposits at banks and savings institutions as of the end of 1999.
- Competition for loans increased as well. Large businesses can borrow by issuing commercial paper, and commercial paper as a source of funds has grown rapidly. Commercial paper outstanding in 1975 was $48 billion; as of the end of 1999 it totaled $1.4 trillion. In comparison, all business loans made by banks and savings institutions summed to $1.0 trillion.
- Nonbank lenders play an important and growing role in serving businesses that are too small to effectively borrow in the commercial paper market. As of 1993, such lending accounted for 35 percent of credit extended to small businesses.
- Nonbanks have made important inroads in consumer lending, accounting for 58 percent of such loans in 1998.

Sources: Board of Governors, *Federal Reserve Bulletin* (2000a), Tables 1.21 and 1.32; Cole and Wolken (1996); Federal Deposit Insurance Corporation (1999); Gilbert (1986); Hahn (1983); Kennickell, Starr-McCluer, and Surette (2000); Robertson (1995).

CRA May Impose Relatively Small Costs

Although CRA has undoubtedly resulted in an overall increase in low-income lending, at many banks the increase may have been relatively small. If CRA has caused only minor changes in bank behavior, banks are at little disadvantage to nonbank competitors because of the Act.

There is some empirical evidence to support the notion that the CRA-induced increase in low-income lending by banks has been fairly small. Evanoff and Segal (1996, 32) find that during the 1980s, growth of low-income mortgage lending by banks lagged growth of middle- and high-income lending. According to Litan et al. (2001, 26), from 1993 through 1999, low-income home purchase loans at institutions subject to CRA rose from 31.5 percent to 35.0 percent of total mortgage loans. Mortgage lending, the focus of CRA lending, accounts for only 15 percent of the average bank's assets, so as a percentage of bank assets the change in the 1990s was fairly small.[18] Moreover, a good bit of this increase seems to have been accounted for by factors outside of pressure brought by CRA regulations, for example by declining costs of lending to low-income borrowers. Such declines may be the result of improvements in the quality of information available to lenders on such

[18] Since banks often sell a large proportion of their mortgage loans in the secondary market, the proportion of mortgage loans to total assets tends to understate the importance of earnings from mortgage lending for bank profitability.

borrowers. Low-income lending by firms not subject to CRA grew even more rapidly than such lending by banks, suggesting that factors other than CRA have increased the attractiveness of low-income lending. For these nonbank institutions, low-income loans grew by 30 percent from 1993 to 1997 (Gunther 2000, 58).

So why haven't CRA regulations resulted in more low-income lending? In part because regulators must balance two conflicting goals when enforcing CRA: expanding low-income lending and assuring bank safety and soundness. The Act requires that additional loans be made only to the extent "consistent with safe and sound [bank] operation" (12 U.S.C. 2901). Thus, one would expect regulators to be reluctant to push banks too far in providing support for community development. Large increases in low-income lending imply a substantial reduction in profits or even losses for the bank. Because bank regulators are not only responsible for encouraging low-income lending but also for enforcing safety and soundness requirements, they are understandably loath to take steps that could undermine soundness.[19]

5. CONCLUSION

There is broad support for efforts to revitalize distressed low-income communities. A partnership of private and public interests as represented by CRA is considered by many an ideal way to accomplish this social goal. While a direct government transfer program might provide the needed funds, private organizations bring a business acumen honed from experience operating in a competitive marketplace. Furthermore, if private firms can contribute to community development, then government budgets are less burdened. But there can also be serious problems in relying on private efforts mandated by legislation. In the case of CRA, a requirement to expand lending in low-income communities may in some circumstances distort credit markets: projects for which costs exceed benefits are undertaken, and projects are rejected when benefits exceed costs. Further, CRA's costs may result in banks losing business to firms that are less efficient at providing deposit and lending services.

While some would like banks to play a greater role in community revitalization, a more aggressive low-income lending policy that further disadvantages banks relative to unregulated competitors would be hard to sustain. CRA will likely survive in a more competitive economy as a tool to fight discrimination against low-income neighborhoods, but those who expect CRA to play a growing role in community development funding may be disappointed.

[19] Gunther (1999) discusses the conflict between encouraging low-income lending and promoting safety and soundness. He provides evidence, at least for small banks, of the conflict between enforcement of safety and soundness standards and CRA compliance.

REFERENCES

Board of Governors. 2000a. "Money Stock and Debt Measures." *Federal Reserve Bulletin* 86 (June): Tables 1.21 and 1.32.

_____. 2000b. "Report on the Performance and Profitability of CRA-Related Lending." Report submitted to the Congress pursuant to Section 713 of the Gramm-Leach-Bliley Act of 1999. Washington: Board of Governors of the Federal Reserve System, July 17.

_____. Regulation BB Community Reinvestment. Code of Federal Regulations, Title 12, c. II, part 228.

California Environmental Protection Agency, Air Resources Board. 2001. "Amendments to the California Zero Emission Vehicle Program Regulations: Final Statement of Reasons." (December).

Canner, Glenn B., and Wayne Passmore. 1995. "Home Purchase Lending in Low-Income Neighborhoods and to Low-Income Borrowers." *Federal Reserve Bulletin* 81 (February).

Cole, Rebel A., and John D. Wolken. 1996. "Bank and Nonbank Competition for Small Business Credit: Evidence from the 1987 and 1993 National Surveys of Small Business Finances." *Federal Reserve Bulletin* 82 (November).

Community Reinvestment Act, U.S. Code, vol. 12, c. 30 (1977).

Davern, Michael E., and Patricia J. Fisher. 2001. "Household Net Worth and Asset Ownership: Household Economic Studies: 1995." U.S. Census Bureau, Current Population Reports, Household Economic Studies, Series P70-71. Washington: U.S. Government Printing Office (February).

Evanoff, Douglas D., and Lewis M. Segal. 1996. "CRA and Fair Lending Regulations: Resulting Trends in Mortgage Lending." Federal Reserve Bank of Chicago *Economic Perspectives* 20 (November): 19–46.

Fama, Eugene F. 1985. "What's Different About Banks?" *The Journal of Monetary Economics* 15 (January): 29–40.

Federal Deposit Insurance Corporation. 1999. *Quarterly Banking Profile.* Washington: Federal Deposit Insurance Corporation (Fourth Quarter).

Federal Financial Institutions Examination Council (FFIEC). 1997. "Community Reinvestment Act: Examination Procedures for Large Retail Institutions." (April).

Gilbert, Alton R. 1986. "Requiem for Regulation Q: What It Did and Why It Passed Away." Federal Reserve Bank of St. Louis *Review* (February): 22–36.

Gramlich, Edward M. 2001. "Preparing for CRA 2002." Speech delivered at the CRA and Fair Lending Colloquium, Boston, Mass. 23 October.

Gunther, Jeffrey W. 1999. "Between a Rock and a Hard Place: The CRA-Safety and Soundness Pinch." Federal Reserve Bank of Dallas *Economic and Financial Review* 2: 32–41.

_____. 2000. "Should CRA Stand for Community Redundancy Act?" *Regulation* 23 (3): 56–60.

Hahn, Thomas K. 1993. "Commercial Paper" In *Instruments of the Money Market,* ed. Timothy Q. Cook and Robert K. LaRoche. Richmond: Federal Reserve Bank of Richmond.

Joskow, Paul L., and Nancy L. Rose. 1989. "The Effects of Economic Regulation." In *Handbook of Industrial Organization*, vol. II., ed. Richard Schmalensee and Robert D. Willig. Elsevier Science/ North-Holland.

Kahn, Alfred E. 1990. "Deregulation: Looking Backward and Looking Forward." *Yale Journal on Regulation* 7 (Summer): 325–54.

Kennickell, Arthur B., Martha Starr-McCluer, and Brian J. Surette. 2000. "Recent Changes in U. S. Family Finances: Results from the 1998 Survey of Consumer Finances." *Federal Reserve Bulletin* 86 (January): 1–29.

Lacker, Jeffrey M. 1995. "Neighborhoods and Banking." Federal Reserve Bank of Richmond *Economic Quarterly* 81 (Spring): 13–38.

Litan, Robert E., et al. 2001. "The Community Reinvestment Act After Financial Modernization: A Final Report." Study prepared for the U.S. Department of Treasury (January).

Morrison, Steven A., and Clifford Winston. 1997. "The Fare Skies: Air Transportation and Middle America." *Brookings Review* 15 (Fall): 42–45.

_____. 1998. "Regulatory Reform of U.S. Intercity Transportation." In *Essays in Transportation Economics and Policy: A Handbook in Honor of John R. Meyer*, ed. Jose A. Gomez-Ibanez, William B. Tye, and Clifford Winston. Washington: Brookings Institute.

Rhoades, Stephen A. 2000. *Bank Mergers and Banking Structure in the United States*, 1980–98. Staff Studies 174. Washington: Board of Governors of the Federal Reserve System.

Robertson, Ross M. 1995. *The Comptroller and Bank Supervision: A Historical Appraisal.* Washington: The Office of the Comptroller of the Currency.

Temin, Peter. 1990. "Cross Subsidies in the Telephone Network after Divestiture." *Journal of Regulatory Economics* 2 (December): 349–62.

Winston, Clifford. 1998. "U.S. Industry Adjustment to Economic Deregulation." *Journal of Economic Perspectives* 12 (Summer): 89–110.

German Monetary History in the Second Half of the Twentieth Century: From the Deutsche Mark to the Euro

Robert L. Hetzel

Starting in January 2002, citizens of the European Monetary Union (EMU) replaced their national currencies with the Euro, issued by the European Central Bank (ECB). Europeans created a new pan-European central bank as a symbol of a future united Europe. However, what historical process explains the broad monetary policy of the ECB, that is, its objective of price stability and its strategy for achieving that objective? The short answer is that its founders designed the ECB to look like the Bundesbank. How then did the Bundesbank evolve? To answer that question, I survey German monetary policy in the second half of the twentieth century.

I divide this history into three main sections.[1] The first treats the Bretton Woods system of fixed exchange rates. The second treats the floating exchange rate period that began in 1973. It chronicles the Bundesbank's ultimate decision to accord primacy to reducing inflation rather than unemployment. The last explains how the Bundesbank dealt with the pressures created by movement toward a single European currency.

The evolution of the Bundesbank into an institution now identified as a modern central bank is fundamental to the article. A modern central bank

This article follows Hetzel (2002), which summarizes German monetary policy in the first half of the twentieth century. The author gratefully acknowledges helpful comments from Michael Dotsey, Martin M. Fase, Andreas Hornstein, Thomas Humphrey, Joachim Scheide, and Alex Wolman. The views expressed in this article are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] An examination of the monetary policies that central banks pursue requires a framework for understanding their choice of objectives and the way that they achieve those objectives. The

determines the behavior of the price level exclusively through (indirect) control over the liabilities on its balance sheet (the monetary base). Conversely, a modern central bank eschews control over the price level through interference with price setting in individual markets. Specifically, modern central banks avoid controls such as exchange and capital controls, tiered exchange rates, quantitative credit controls, moral suasion directed at the price setting of private parties, direct wage and price controls, and so on.

Another theme of the article is how free markets restrict a country's choice of monetary arrangements. The system of fixed exchange rates required disruptive changes in the price level. Political pressures to avoid those changes in turn produced pressures for exchange controls. Ultimately, Germany's commitment to a free market economy pushed it to reject fixed exchange rates and adopt floating exchange rates. The Bundesbank came to embody the modern conception of a central bank when, in the 1980s, it demonstrated how to achieve price stability in a free market economy. That outcome contrasted starkly with German monetary experience in the first half of the twentieth century (Hetzel 2002).

To create the Bundesbank that came to epitomize a modern central bank, Germany had to travel a long road. When West Germany came into existence in 1949, the intellectual consensus in the West held that the depression arose out of a grand failure of the free market model (Hetzel 2002). Similarly, very few economists understood the price level as a monetary phenomenon, that is, something under the control of the central bank. Instead, professional and private opinion held that powerful labor unions and large corporations determined prices through their monopoly power. In this intellectual environment and against the backdrop of staggeringly high unemployment during the depression, there was unanimous political opposition within postwar Germany to an independent central bank (Buchheim 2002).

At the same time, there was a demand for monetary stability. That demand arose out of the prewar experience of inflation, first hyperinflation and then the suppressed inflation of the Third Reich. The memory was still recent of the 1948 currency reform that extinguished most of the savings held by Germans in currency and bank deposits. In between inflations was the deflation of the

---

quantity theory provides such a framework. Modern central banks create only paper money, not wealth. For that reason, ultimately all they can control is the price level—the money price of goods. The quantity theory explains the behavior of the price level based on the way that central banks create and destroy money.

A core implication of the quantity theory is that a central bank must choose between two roles for the price level. With a system of fixed exchange rates, the central bank must allow the price level to vary to produce the real terms of trade that equilibrates the balance of payments. That is, the internal price level must vary to price the country's exports in a way that achieves balance of its international payments. Alternatively, with a system of floating exchange rates, the price level varies to endow the nominal money stock with the real purchasing power the public desires. The rate of inflation then depends upon the trend rate of growth of money the central bank chooses.

depression. Germany accepted the postwar policy consensus that monetary stability required pegged, infrequently adjusted exchange rates. Policymakers attributed the depression in part to the competitive devaluations that countries imposed to stimulate their economies and to the associated financial instability arising from the speculative capital flows those devaluations engendered (Yeager 1976, 375). In response, the Western world designed the Bretton Woods system.

1. THE BRETTON WOODS SYSTEM OF PEGGED EXCHANGE RATES

In July 1944 at Bretton Woods, the victorious West designed a monetary system intended to substitute stable exchange rates for the disruptive changes in exchange rates that characterized the depression. However, in time that system itself became a source of instability. The system became a dollar standard in which U.S. monetary policy determined the inflation rates of the other member countries. Also, the system of pegged but adjustable exchange rates allowed those rates to move far from equilibrium. The resulting one-way bets on exchange rate changes recreated the hot money flows that the designers of the Bretton Woods system had hoped to banish. This section explains how, over a two-decade period, Germany went from a pegged to a floating exchange rate regime.

How Did the System Work?

With the end of price controls in June 1948, Germany moved dramatically toward an internal free market. However, Germany continued to manage its foreign trade through trade deals arranged bilaterally with foreign governments. In an open economy, managed trade requires significant government control of economic activity. In order to move toward free trade, Germany needed to abandon achievement of balance-of-payments equilibrium through managed trade and restrictions on trade and capital controls.

Once Germany adopted the pegged exchange rates of Bretton Woods, free trade required it to adjust its internal price level to price its export goods at competitive international levels. After the war, like other European countries, Germany experienced chronic trade deficits with the United States. Commentators assumed an indefinite dollar shortage. External trade balance with unchanged exchange rates would have required deflation by Germany.

The postwar assumption that governments had a responsibility to manage the economy to prevent high unemployment made such deflation impossible. A major point of discussion below is how the Bretton Woods system worked in the postwar period without forcing deflation on countries with trade deficits.

Faced with a requirement to deflate in order to achieve external balance, countries would undoubtedly have resorted to trade restrictions and capital controls. Such recourse would have made impossible the postwar trade liberalization that actually occurred.

The Bretton Woods system worked because of the behavior of its two largest members, the United States and Germany. With its extraordinarily large gold reserves, the United States could maintain an overvalued exchange rate and lose gold for a long period without reacting.[2] At the same time, Germany maintained an undervalued exchange rate by recycling its trade surpluses to its neighbors through capital exports and foreign aid. Not until the early 1970s did the inherent instability of the Bretton Woods system and conflicting domestic policies of its member countries cause the pegged exchange rate system to collapse.

A New Central Bank

After the war, Germany had no central bank.[3] In its Western zones, the offices of the former Reichsbank assumed some of the responsibilities of a central bank. However, these banks could not issue currency and therefore constituted only the empty shell of a central bank. The political objective of preventing a return to a centrally organized banking system with interlocking links to large corporations motivated U.S. policy exclusively. The British had a more realistic attitude. They emphasized that economic integration required the existence of a note-issuing bank that could assure settlement of transactions on an economywide basis.

When the American and British occupation zones merged, British pragmatism prevailed. In March 1948, the Allies created the Bank deutscher Länder (BdL) to oversee the regional offices of the former Bundesbank. It possessed the power to issue currency. With the replacement of barter with monetary transactions after the June 1948 currency reform, the BdL became West Germany's central bank. The BdL possessed a governing structure modeled after the U.S. Federal Reserve System. The precedent of central bank independence set by this structure became impossible to reverse when Germany reorganized its central bank in 1957 to put it on a firm legal footing.

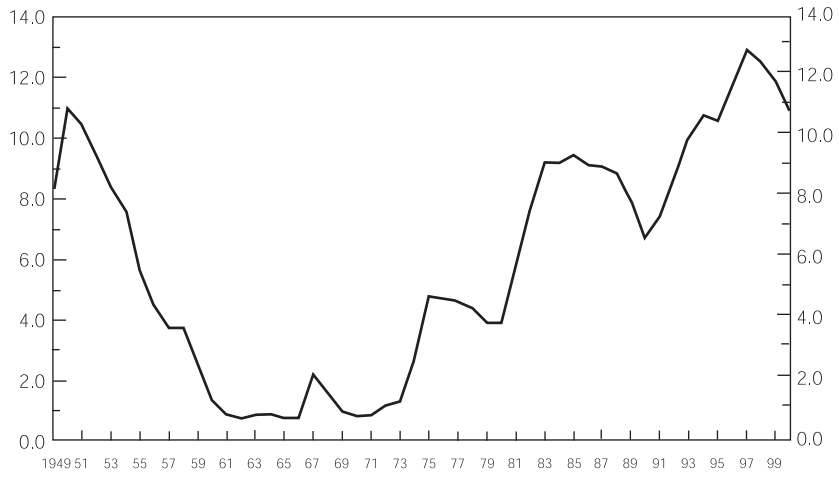
The Allied decree that established the BdL on 1 March 1948 required it "to stabilize the currency."[4] The Bank emphasized that priority despite the rise in

[2] Given domestic price levels, an overvalued exchange rate yields an excess demand for imports relative to exports. Similarly, an undervalued exchange rate yields an excess of exports over imports. In the first case, a country with a fixed exchange rate loses official reserve balances. In the second case, it gains reserves.

[3] The material in this and the following paragraph is from Buchheim (1999, Sections 3 and 4).

[4] Holtfrerich (1999, 318). This paragraph draws on Holtfrerich (1999, Section 4).

Figure 1 German Unemployment Rate



Notes: Data are from "Bevölkerung und Erwerbstätigkeit im Deutschen Reich und in der Bundesrepulik Deutschland" Bundesarbeitsblatt 7–8/1997, pages 110–11, Bundesanstalt für Arbeit, Ministerium für Arbeit und Sozialordnung und eigene Berechnungen.

the unemployment rate from 4.2 percent in 1948 (Holtfrerich 1999, 328) to 11 percent in 1950 (Figure 1).[5] From June 1948, the date of the currency reform, to October 1948, the cost of living rose by 14 percent.[6] The BdL responded strongly to the inflation.[7] The price level began to decline in 1949 and by June 1949 had reached its June 1948 level.

[5] Giersch et al. (1992, Chapter 3) argue that the rise resulted from an influx of refugees from former German territories and from increases in worker productivity.

[6] The inflation indicated the continued presence of a monetary overhang despite the drastic reduction in the units of circulating money. The inflation occurred despite the increase in demand for money implied by the more than 50 percent increase from June through December 1948 in the Allied index of bizonal industrial production (Buchheim 1999, 96).

[7] It did so in a way determined by its quantitative (as opposed to interest rate) procedures for the control of bank deposits and credit. The need for quantitative procedures came from the fact that the discount rate of 5 percent set in June 1948 did not bind, as it was 2 percentage points above market rates (Holtfrerich 1999, Figure 3). Commercial banks possessed large amounts of excess reserves and did not borrow significantly from the BdL.

In fall 1948, the BdL tightened restrictions on credit extension by banks. The Allies wrote off 80 percent of the deposits that remained frozen after the June reform but had been designated for ultimate one-to-one conversion to deutsche marks (DMs). (They amounted to 50 percent of the former bank deposits.) Government surpluses also reduced bank reserves.

By the end of 1949, reductions in banks' excess reserves and in the discount rate to 4 percent put banks into the BdL's discount window. The resulting change in procedures from quantitative credit controls to indirect control through the bank rate marked an important step toward market rather than administrative rationing of credit.

Integrating Germany into the World Economy

The Allies maintained the old dollar-reichsmark exchange rate of 0.30, which overvalued the deutsche mark (DM). Similar overvaluations of other currencies relative to the dollar resulted in an autarkic system of international trade after World War II.[8] Because their currencies were overvalued, the countries of Europe managed the trading of their residents so that transactions would balance bilaterally. By spring 1947, there were 200 bilateral agreements controlling trade in Europe alone. Importers had to obtain licenses, which limited total imports country by country. Governments made their imports conditional on another country's acceptance of their exports because they feared running short of the dollar reserves needed for essential food and fuel imports.

As a condition for aid, the United States insisted that European countries replace bilateral trade deals with free trade and multilateral clearing arrangements. Backed by $350 million of Marshall Plan money, Western European countries agreed in September 1950 to create the European Payments Union (EPU).[9] However, the EPU created no mechanism for eliminating overall payments imbalances.

Very quickly, the arrangements for the EPU came close to collapse because of German balance-of-payments deficits. With the outbreak of the Korean War in 1950, Germans, like Americans, tried to buy goods for fear inflation would resurge (Hetzel 2001; Holtfrerich 1999, 334; Yeager 1976, 413). Germany came under both foreign and domestic political pressure to control its imports of raw materials through a system of central administration and to reimpose price controls. "This was the last time that the essence of the liberal economic system which West Germany had adopted in mid 1948 was actually put in jeopardy" (Giersch et al. 1992, 101).

Germany's Minister of Economics, Ludwig Erhard, refused to abandon his free market reforms and predicted that his country's trade deficit would disappear. Despite opposition from Chancellor Adenauer, the Bundesbank raised interest rates (Marsh 1992, 152). Erhard's prediction came true. By spring 1951, Germany began to run trade surpluses with the EPU. Germany's

[8] This paragraph and the next two summarize material in Yeager (1976, Chapter 21).

[9] The multilateral clearing of EPU achieved a great simplification by removing the detailed government interventions necessary for bilateral clearing. By making the members jointly responsible for the credit of each member, the EPU allowed consolidation of a member's transactions into a single overall net claim on the EPU.

Whether overall the Marshall Plan encouraged free trade is unclear. Its dollar payments, most of which went to Britain, allowed countries to maintain overvalued exchange rates. Governments possessed an incentive to maintain overvalued exchange rates because they lowered the cost of food imports. In the post-depression intellectual environment of the time, "elasticity pessimism" implied that devaluation would work only very slowly to reduce a trade imbalance. The trade deficits created by overvalued currencies also provided European governments with the economically perverse but politically attractive incentive to retain restrictions on trade (see Giersch et al. [1992, 94–95]).

free market reforms endured. German economic growth in the 1950s earned the name of Wirtschaftswunder (economic miracle).

Maintaining Balance-of-Payments Equilibrium

The West succeeded in establishing a peaceful Europe after World War II, where it had failed after World War I. Its success depended in part on the establishment of a monetary system that did not entail the deflation of the late 1920s and early 1930s. German and U.S. behavior allowed movement toward free trade with a system of pegged exchange rates while avoiding deflation.[10]

Immediately after the war, an undervalued dollar created a dollar shortage, and the United States ran a huge balance of trade surplus. In part, the United States recycled the resulting reserve inflows through unilateral transfers to the rest of the world.[11] Equally important, in 1949, the United States encouraged its trading partners to devalue their currencies (revalue the dollar).[12] By ceasing to overvalue their currencies, these countries eliminated pressures to either deflate or resort to protectionism.

After a small deficit in 1951 and surplus in 1952, Germany began to run persistent current account surpluses (Giersch et al. 1992, Table 28). Left alone, the surpluses would have widened because of the increasing competitiveness of German industry. After the war, Western European countries purchased their capital goods exclusively from the United States. However, in the 1950s, Germany replaced the United States as the major exporter of capital goods to European countries (Giersch et al. 1992, 88–89; Yeager 1976, 486). Germany

---

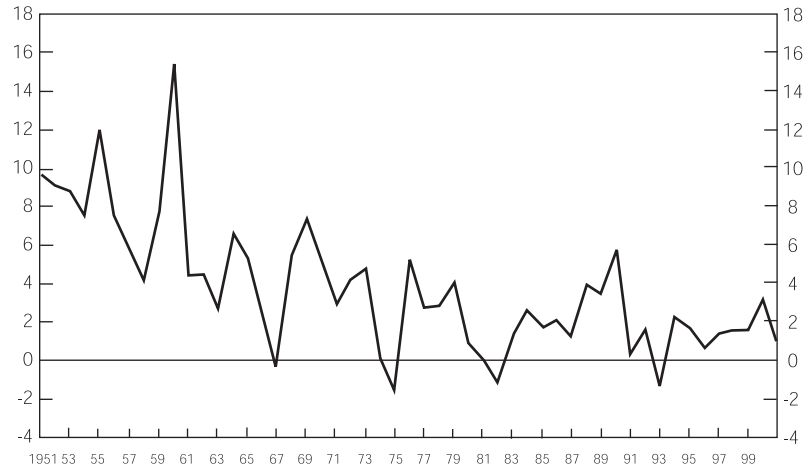
[10] The rejection of isolationism by the United States was also of central importance. U.S. aid replaced punitive reparations. The resulting cooperation between European countries in the form of the Organization for European Economic Cooperation (OEEC) facilitated the re-entry of Germany into the European community. With the establishment of GATT, the United States encouraged an open, multilateral trading system.

[11] The sum of government and private unilateral transfers and capital outflows was $6.8 billion in 1949. The figure fell to around $4 billion in the mid-1950s, but rose again to almost $7 billion in 1957 (U.S. Historical Statistics, Part 2, Series U 1-25, "Balance of International Payments: 1790 to 1970").

[12] On 18 September 1949, after a sterling crisis and with U.S. encouragement, Britain devalued the pound by 30.5 percent. Thirty other countries, accounting for approximately two-thirds of all world trade, followed in devaluing relative to the dollar (Yeager 1976, 444–45). Those devaluations left the dollar somewhat overvalued for most of the 1950s. Giersch et al. (1992, 93) present a graph of the difference between the free market DM exchange rate on the Zürich market and the official exchange rate. According to this measure, the DM was about 20 percent overvalued in early 1950. By the end of 1953, the overvaluation disappeared.

Starting in 1950 and for the remainder of that decade, the United States incurred a deficit on current account (with the exception of 1956 and 1957, when the United States benefited from special factors related to the Suez crisis). From 1950 through the end of the Bretton Woods system in 1973, the United States persistently lost gold. (In 1949, the U.S. gold stock was $24.6 billion. It declined steadily and reached $10.5 billion in 1972, the last full year of the Bretton Woods system.) The willingness of the United States to allow reductions in its gold stock allowed other countries to rebuild their reserves.

Figure 2 German Real GDP Growth



Notes: Data are from Ritschel and Spoerer (1997), Tables A.1, A.2, and A.3.

responded to balance-of-payments surpluses by liberalizing its trade faster than other countries (Holtfrerich 1999, 331).[13]

After 1955, liberalization of its capital controls allowed Germany to become a significant exporter of capital (Yeager 1976, 490–96; Giersch et al. 1992, Table 28). When the EPU disbanded in December 1958, Germany was granting it significant amounts of credit to cover the deficits of other countries (Yeager 1976, 412). In the 1960s, Germany pursued a large-scale foreign aid program and capital exports (Holtfrerich 1999, 377, 393).

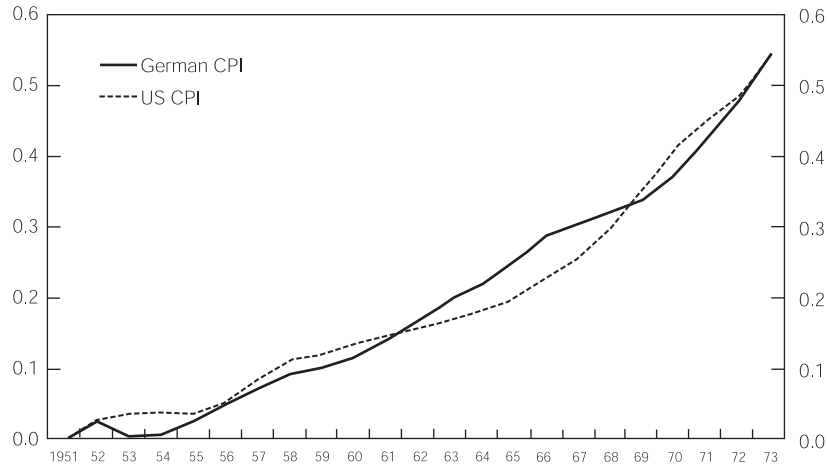
An Independent Bundesbank

Although the system of pegged exchange rates left the BdL without goal independence, it still had to establish credibility for instrument independence.[14]

[13] Germany became a "pioneer of European [trade] liberalization" in an effort to "become a respected member of the Western world" (Giersch et al. 1992, 108–09). If Germany enjoyed monopoly power in its exports of capital goods, reducing tariffs would reduce balance-of-payments surpluses by weakening its terms of trade.

[14] For a bank that needs to establish credibility, an exchange rate peg serves as a clear objective, the accomplishment of which is evident to everyone. Because of the random variability that relative price changes impart to short-term movements in the price level and the long lags with which money affects the price level, central bank control of a price level objective becomes evident only over a long period of time. There are then reasons for a new central bank to peg its currency to that of a stable currency.

Figure 3 Price Levels



Notes: Annual observations of the natural logarithm with 1951 as the base. The source of the data is Deutsche Bundesbank, ed. *Geld und Bankwesen 1876–1975* (1976).

It had to do so despite the fact that "the Allies had established a totally independent central bank contrary to German wishes" (Buchheim 2002, 10). The success of the BdL in maintaining the pegged exchange rate in an environment of falling unemployment, strong real growth, and price stability created public support for independence (Figures 1, 2, and 3). When in 1956 the Bundesbank increased its discount rate by 1 percentage point, Chancellor Adenauer challenged the BdL publicly. He stated that "it is the little ones who will suffer most. . . . [T]he guillotine falls on the man in the street, and that is what grieves me so much" (Neumann 1999, 291). Public reaction, however, supported the BdL.

When in 1957 Germany replaced the Allied promulgation establishing the BdL with the law establishing the Bundesbank, it had to respect public support for an independent central bank (Buchheim 2002, 11ff).[15] The 1957 Bundesbank Act created the Bundesbank and instructed it to "regulate the amount of money in circulation and of credit supplied to the economy with

[15] The Bundesbank also demonstrated its (instrument) independence later, in 1966 when output stopped growing and the unemployment rate rose (Figures 2 and 3). Despite a public attack by the Erhard government, the Bundesbank refused to lower its discount rate. Chancellor Erhard's government fell in November 1966. Because of the unsatisfactory behavior of wage growth and the government budget, the Bundesbank then refused the demands of the new government of Chancellor Kiesinger to cut interest rates (Holtfrerich 1999, 380–81).

the aim of safeguarding the currency" (Holtfrerich 1999, 318). The written report of the Chairman of the Committee for Money and Credit of the German parliament stated:[16]

> The security of the currency... is the highest precondition for the retention of a market economy, and hence in the final analysis that of a free constitution for society and the state.... [T]he note-issuing bank must be independent of these [political bodies] and subject only to the law.

However, the Act's mandate for "safeguarding the currency" left unstated whether the Bundesbank should stabilize the internal or external value of the DM.

U.S. Inflation Destroys Bretton Woods

Germany had strong reasons for maintaining the peg of its currency with the dollar at a fixed rate. The United States' technological and manufacturing supremacy coming out of World War II had made the dollar a symbol of strength. A stable exchange rate with the dollar gave prestige to the mark. Later, Germans associated exchange rate stability with the export boom that powered postwar economic recovery. The export industries that benefited did not want a revalued mark that would erode their profits.

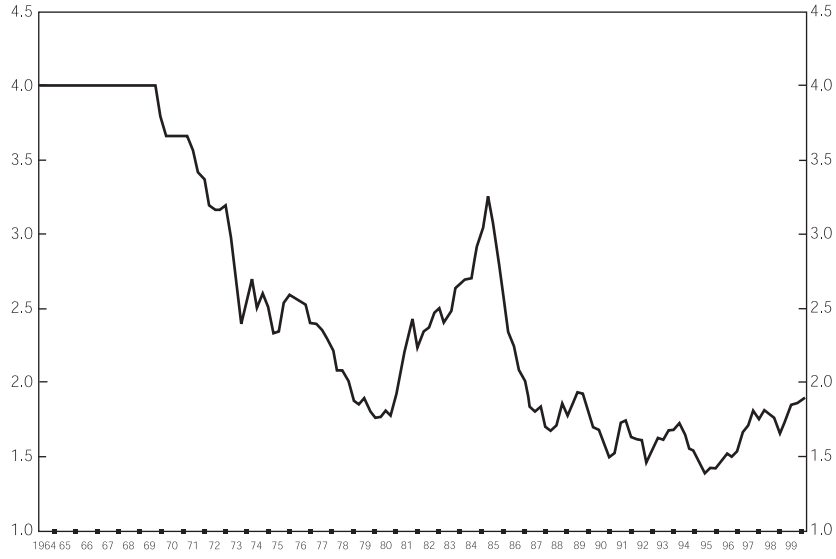
However, the inflationary monetary policy the United States began in the mid-1960s forced fundamental change on Germany both in its monetary arrangements and in its intellectual environment. Maintenance of a fixed exchange rate required Germany to match U.S. inflation. Initially, Germany resorted to capital controls in a futile attempt to retain fixed exchange rates and to avoid imported inflation. Eventually, Germany chose floating exchange rates. Ten years of intellectual ferment then passed before the Bundesbank used its newfound freedom to make price stability its overriding objective.

Like other countries in the 1970s, Germany experimented with aggregate demand policies aimed at controlling inflation. In Germany, the unemployment rate rose far above the levels of the 1960s (Figure 1). Despite this fact, inflation reached peaks of 7 percent in 1974 and 1982. Two results became apparent. First, to control inflation, the central bank had to control money growth. Second, high rates of money growth produced inflation, not low unemployment.

The Bretton Woods system of pegged exchange rates required Germany to allow its price level to rise along with that of the United States. Figure 3 shows that over the period of the Bretton Woods system, U.S. and German price levels

[16] Quoted in Stern (1999, 149).

Figure 4 Deutsche Mark–Dollar Exchange Rate



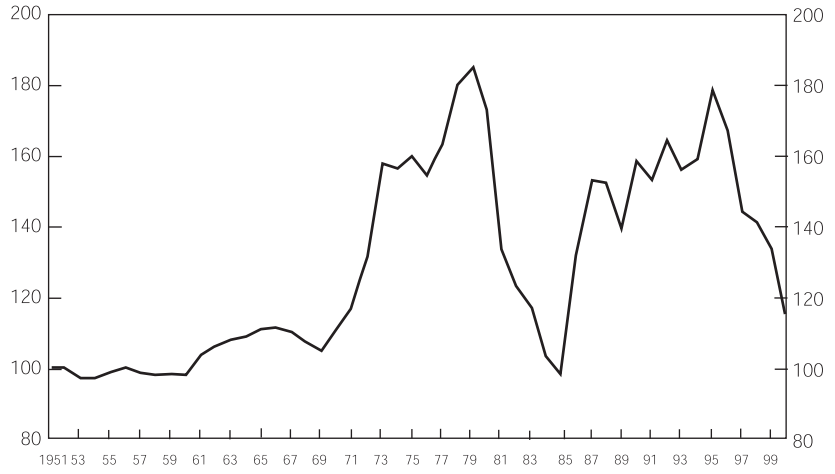
Notes: The source of the data is DRI-WEFA.

rose by the same amount.[17] The increase in the rate at which prices rose in the United States, beginning in 1965, ultimately destroyed the Bretton Woods system. Germany was willing to accept 3 percent "adjustment inflation" as a cost of the system (Holtfrerich 1999, 383). Moderate inflation was less costly to the government than a confrontation with the export industries over a revaluation of the DM. However, in the early 1970s, U.S. inflation pushed the Bundesbank beyond its limit. Even so, abandonment of an exchange rate peg and the move to a floating exchange rate came only after bitter debate and a wave of inflation.

Figures 6 and 7 display graphically the dilemma Germany faced. The high rates of money growth created by the Bundesbank's defense of the mark-dollar

[17] Given the similar rise in price levels, Germany's terms of trade relative to the United States rose by the amount of its revaluations. Germany revalued its currency relative to the dollar by 5 percent in March 1961. Germany let the mark float upward in September 1969 and then set a peg at a rate that revalued the mark by 9.3 percent (Figure 4). It again let the mark float upward in May 1971. In December 1971 as part of the Smithsonian accords, Germany pegged to the dollar at a rate that revalued the mark by 13.6 percent relative to its prior Bretton Woods parity. Despite these revaluations, Germany's terms of trade rose with the breakdown of Bretton Woods in 1973 (Figure 5). That fact suggests that the capital controls imposed by the United States in the 1960s kept the mark from becoming more undervalued (the dollar from becoming more overvalued).

Figure 5 Real Exchange Rate



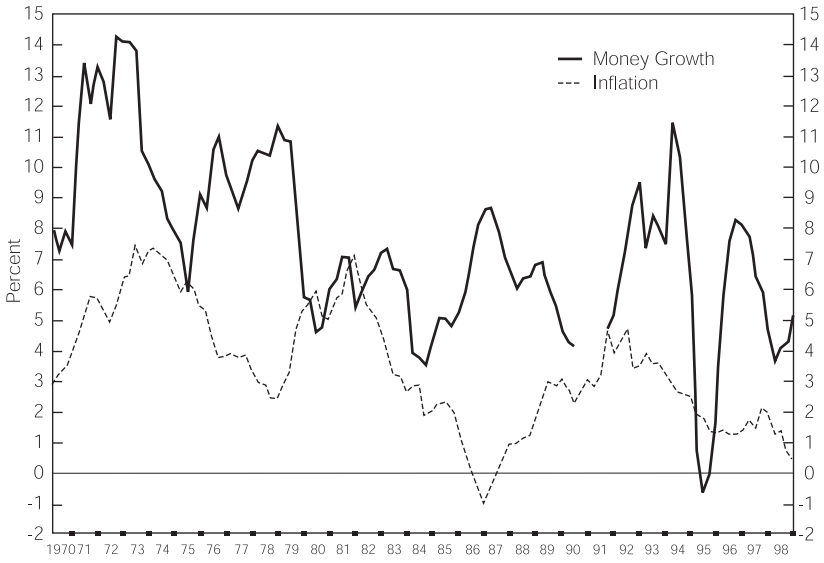
Notes: The real exchange rate is the ratio of the DM to dollar exchange rate multiplied by the ratio of the U.S. CPI to the German CPI. 1951 equals 100.

exchange rate in the early 1970s produced high rates of inflation. Figure 6 plots quarterly observations of four-quarter growth rates of money and prices. Figure 7 fits a step function to annualized quarterly growth rates of money and prices. It highlights the lagged relationship between changes in money growth and inflation through arrows that connect steps in money growth to subsequent steps in inflation (see also Table 1).

In response to recession, in 1970 the Federal Reserve pushed down U.S. short-term interest rates. Germany, with strong real output growth, maintained a high level of short-term rates (Figure 2; von Hagen 1999, Figure 4). As a result, large amounts of short-term capital flowed into Germany, while its significant export of long-term capital ceased (Giersch et al. 1992, Table 28). Net capital inflows forced the Bundesbank in May 1971 to buy large amounts of dollars, including $1 billion in the last 40 minutes of trading on 5 May. Germany let the mark float and then revalued it as part of the December 1971 Smithsonian agreement.

Opposing camps within the Bundesbank and the government debated whether to stabilize the external or the internal value of the DM. Johnson (1998, 70) quotes a Bundesbank official who characterized this debate as "a Glaubenskrieg (religious war) of Wagnerian proportions." Karl Klasen became Bundesbank president in January 1970. He favored capital controls to preserve the foreign exchange value of the mark and credit controls to limit

Figure 6 Money Growth and Inflation in Germany



Notes: Quarterly observations of four-quarter growth rates of the CPI and M3. Data are from DRI-WEFA. Heavy tick marks indicate fourth quarter. The gap in the money growth series arises from deletion of the observations distorted by the discrete jump in money that occurred with German unification in 1990.

the credit extension of German banks (Johnson 1998, 70–84; Solomon 1982, 178–80).[18]

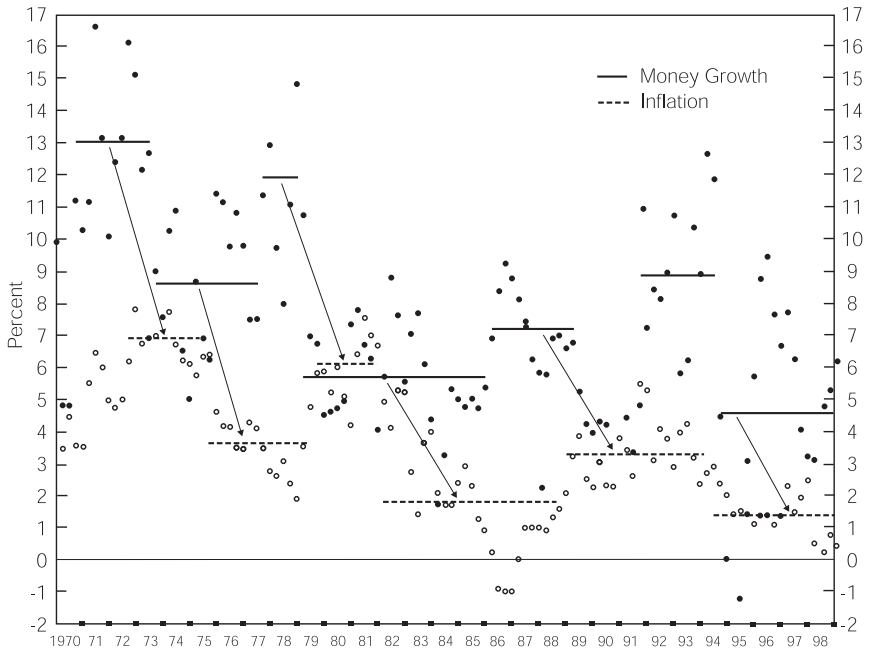
In April 1972, in an arrangement called the Snake, most European countries agreed to limit fluctuations in their exchange rates more strictly than provided for by the Bretton Woods system. When Britain abandoned the Snake in June, speculators attacked the mark. Germany responded by adopting capital controls (Yeager 1976, 514).[19] Again, as in 1950, Germany had to

[18] In 1961, Bundesbank President Blessing had strongly opposed revaluation. It was Ludwig Erhard, Minister of Economics and author of Germany's free market reforms, who understood that revaluation would allow Germany to maintain internal stability without external controls (Holtfrerich 1999, 375).

The essence of such controls is the implicit taxation of international investment flows, that is, fiscal policy. Fiscal policy is properly the province of the government not the central bank. Capital and credit controls are ultimately incompatible with central bank independence and with free markets.

[19] The Federal Reserve encouraged the Bundesbank ("Memorandum for the President," Herbert Stein, 11 July 1972, Stein Box 49, "Nixon Presidential Materials Project" National Archives

Figure 7 Money Growth and Inflation in Germany



Notes: The dots are two-quarter moving averages of the current and preceding quarters' annualized M3 growth. The solid line is a step function fitted to the quarterly growth rates. Observations are omitted for M3 from 1990Q3 through 1991Q2. 1990Q3 includes the discrete jump in M3 of 15 percent. The circles are two-quarter moving averages of the current and preceding quarters' annualized quarterly growth of the CPI. The dashed line is a step function fitted to the quarterly growth rates. The arrows connect the money step to the subsequent inflation step. Heavy tick marks indicate fourth quarter. See Table 1.

decide whether to adopt monetary arrangements that would require a retreat from free markets.

Overwhelmed by an inflationary American monetary policy, the Bretton Woods system collapsed definitively in early March 1973. In February 1973, the Bundesbank monetized foreign exchange inflows equal to 15 percent of

---

and Records Administration in College Park, MD): "In his letter of July 8 to you, Arthur Burns describes the circumstances surrounding the German decision to impose exchange controls rather than allow the mark to rise relative to the dollar. The German cabinet was apparently motivated by the desire to minimize political difficulties—including difficulties for us in November. Dr. Burns congratulated Dr. Klasen, head of the German Central Bank, on his victory in behalf of international monetary stability."

Table 1 Money and Inflation Steps

Time Period	Averages of Annualized M3 Growth	Time Period	Averages of Annualized CPI Inflation
70Q3–73Q2	13.0	72Q3–75Q2	6.9
73Q3–77Q2	8.6	75Q3–79Q2	3.6
77Q3–78Q4	11.9	78Q3–81Q4	6.1
79Q1–85Q4	5.7	82Q1–88Q3	1.8
86Q1–89Q1	7.2	88Q4–94Q1	3.3
91Q3–94Q2	8.9		
94Q3–98Q4	4.6	94Q2–98Q4	1.4

its monetary base (von Hagen 1999, 686). The Bundesbank's purchase of 2.7 billion dollars on a single day, 1 March (3 percent of its monetary base), forced a reluctant German government to float (Marsh 1992, 165).[20] Henceforth, the mark floated against the dollar (Figure 4). Although Germany finally chose free markets and a floating exchange rate, its stubborn defense of pegged exchange rates allowed the Bretton Woods system to continue long enough to turn U.S. inflation into a worldwide economic boom and inflation.[21]

This experience made clear to the Bundesbank that a system of pegged exchange rates had stolen its independence. An exchange rate peg meant that the Federal Reserve determined German monetary policy. Loss of independence came from loss of control over money creation. Members of the Bundesbank Council in their meetings described the Bundesbank as a "self-service store for central bank money" (von Hagen 1999, 686).

Pegged Exchange Rates and Price Fixing

The Bretton Woods system contained an inherent contradiction: the attempt to combine stable and adjustable exchange rates. Its architects wanted stable

[20] For a fuller account, see Yeager (1976, 515–16) and Solomon (1982, Chapter XIII). Johnson (1998, 82–83) contains an account of the drama. Otmar Emminger, the Bank's vice-president, led the free market camp that favored a floating exchange rate and the indirect control of inflation through monetary control. Helmut Schlesinger, a newly appointed Bundesbank Council member, supported Emminger. By law, the German government set the external value of the mark. Backed by an implicit threat of resignation by the Bundesbank's Directorate, Emminger issued a challenge to the government's policy of exchange rate pegging. In the climactic meeting with Willy Brandt, both Bundesbank President Klasen and Finance Minister Schmidt, who opposed floating, were in the hospital and Emminger prevailed.

[21] The economic boom and strong demand for energy made possible the sharp rise in oil prices engineered by OPEC in 1973 and 1974. Hetzel (1999, Section 3) argues that excessive money growth, not oil price shocks, caused the inflation of the 1970s and early 1980s.

exchange rates to eliminate what they saw as the destructive, competitive devaluations of the 1930s. The supposed failure of markets in the depression led policymakers to believe that floating exchange rates would be destabilizing. At the same time, its architects wanted adjustable exchange rates so that individual countries could equilibrate their balance of payments without recourse to deflation or inflation. The resulting system delayed changes to maladjusted exchange rates until a speculative crisis forced a large change.

Independent central banks that could sterilize international reserve flows robbed the Bretton Woods system of an automatic market mechanism for equilibrating the balance of payments. Governments faced an incentive to use capital controls to reconcile the conflicting demands of external and internal stability of their currency. Central bank discretion in a system of pegged exchange rates created all the problems associated with government price fixing. The pressure on government to impose capital controls politicized monetary policy. Consider the 1969 revaluation of the DM.

In May 1968, the attempt to defuse riots in France led the French government to grant large wage increases to workers.[22] De Gaulle's refusal to devalue the franc then set up a one-way bet for speculators: Although Germany would not devalue the DM, it could well revalue it relative to the franc. In November 1968, large inflows of foreign currency into the Bundesbank prompted Germany to enact a "pseudo-revaluation" with a special tax on exports and tax allowance on imports. When a leaked Bundesbank memorandum showed that the Bundesbank had in fact not supported the measure, the head of the SPD party accused it of having "stabbed its own government in the back."

The Bundesbank imposed a non-interest-bearing reserve requirement of 100 percent on the growth of foreign deposits. In early May 1969, despite heavy inflows of foreign exchange, the government, which had authority over the foreign exchange parity of the DM, again decided against revaluation, "finally, unequivocally and eternally" as a government spokesman put it (Holtfrerich 1999, 388; Yeager 1976, 509). The issue of whether to revalue dominated the September 1969 elections. The new government of Willy Brandt revalued the DM. Speculative capital flows then reversed as more than DM 20 billion in hot money flowed out of Germany.

2. CHOOSING THE GOAL OF MONETARY POLICY

The March 1973 float of the mark gave the Bundesbank goal independence. However, in the 1970s, the Bundesbank remained divided over how to exercise that independence. The 1957 law that created the Bundesbank required it to "safeguard the value of the currency." A history of inflation leading to

[22] This paragraph and the next draw on Holtfrerich (1999, 384–89).

social discontent encouraged Germany to interpret that phrase to mean that the Bundesbank had to safeguard the internal value of the mark. However, Germany also had a history of unemployment leading to social discontent. That history encouraged a monetary policy directed at holding down unemployment. In the 1970s, the Bundesbank had to contend with pressure from the governments of Willy Brandt and Helmut Schmidt, which made unemployment the top priority.[23]

The Indecision of the 1970s

After the 1973 float of the mark, the Bundesbank concentrated on lowering inflation and money growth fell (Figure 6). However, by the end of 1974, the unemployment rate had risen significantly (Figure 1). Some Bundesbank council members supported the Brandt government's policy goal of full employment. Others, including Helmut Schlesinger and Otmar Emminger, opposed an activist stabilization policy.[24]

This division appeared in a disagreement over how to use the newly invented money target. In 1973, the Bundesbank had adopted a target for "central bank money."[25] In December 1974, it began annual public announcements of the target. The hawks wanted to use money targets to ratchet down money growth and inflation over time. The doves wanted to use them to reassure the public that temporarily expansionary monetary policy would not become inflationary (von Hagen 1999, 425).

In a victory for the doves, toward the end of 1974, the Bundesbank began to lower interest rates. It also announced a generous target of 8 percent for money growth. Beginning in 1977, the mark began to appreciate strongly against the dollar. That appreciation made the Bundesbank unwilling to raise interest rates (von Hagen 1999, 423; Baltensperger 1992, 442). Monetary policy remained inflationary throughout the remainder of the 1970s (Figures 6 and 7). The Bundesbank regularly overshot its already expansionary money targets. In 1978, growth of central bank money exceeded its average target of 8 percent by 3 percentage points.

The decision made in March 1973 to float the mark gave the Bundesbank goal independence and made possible the path that would lead to the Bundesbank of the 1980s: the independent, monetarist pillar of West German society.

---

[23] Finance minister Karl Schiller resigned in June 1972 over the refusal of the government to allow a float of the mark against the dollar. Willy Brandt replaced him with Helmut Schmidt, who later replaced Brandt as Chancellor. Schmidt used the slogan, "Five percent inflation is better than five percent unemployment" (Johnson 1998, 78).

[24] This paragraph and the following one draw on von Hagen (1999, 686) and Johnson (1998, 90–95).

[25] This monetary aggregate was akin to the monetary base adjusted for changes in reserve requirements.

However, in the 1970s, other paths beckoned. Germany and other Western countries undertook the Keynesian experiment of aggregate demand management. The failure of such policies to "buy" low unemployment through high inflation changed the intellectual and political environment.

In that new environment, countries were willing to assign the control of inflation to their central banks. Japan had steadily pursued a policy of a gradual return to price stability since 1974. In 1979, with the election of Margaret Thatcher as prime minister, Britain began to disinflate. In that same year, Federal Reserve Chairman Paul Volcker led the United States down the path of disinflation. The stagflation of the 1970s also made Germany willing to forswear activist macroeconomic policies and to assign to the Bundesbank the goal of price stability (Dyson and Featherstone 1999, 747, 752).

The EMS Threatens Bundesbank Independence

In the postwar period, European economic integration became a strategy for anchoring Germany within a democratic Europe (Arestis, McCauley, and Sawyer 1999, 1–4). Wim Duisenberg (1999, 4), first head of the European Central Bank, expressed the goal using Thomas Mann's phrase: "A European Germany, rather than a German Europe."

In December 1969 at a conference at The Hague, governments of the European Economic Community (EEC) agreed on European monetary union as a goal. After the breakdown of Bretton Woods in 1973, Germany wanted to make the Snake into a joint float of European currencies against the dollar. One reason was that when capital flowed out of the United States in response to inflationary worries, it primarily went to Germany. The mark then appreciated not only relative to the dollar but also relative to Germany's major European trading partners, and German exports suffered. However, European countries were not ready to sacrifice independent national monetary policies. In 1977 and 1978, the dollar again depreciated strongly (Figure 4). The corresponding appreciation of the mark gave Germany an incentive to revive the Snake once again.

In summer 1978, Germany's Chancellor Helmut Schmidt and France's President Valery Giscard d'Estaing agreed to link their currencies within a European Monetary System (EMS).[26] Their motivations foreshadowed those that would later lead to the EMU. Germany needed its foreign policy interests to be identified with an aspiration to build a united Europe. Other countries would then be more receptive to German diplomatic initiatives, especially

[26] The EMS's "exchange rate mechanism" of fixed parities is referred to as ERM.

toward Eastern Europe.[27] Schmidt said later that the EMS was part of a "grand strategy for integrating Europe" (Marsh 1992, 202).

France wanted to build pan-European institutions to ensure that its influence within Europe would remain on par with that of an economically dominant Germany. France was confident that its civil servants would assure attention to French interests within such institutions. However, because the EMS did not create a single central bank for all of Europe, it left unanswered the roles of the Bundesbank and the Banque de France in maintaining the exchange rate peg. The Bundesbank, however, did not wait for Bonn and Paris to define its role.

The Bundesbank Defines Itself

The launch of the EMS in March 1979 initiated the Bundesbank policy of setting money targets to achieve price stability (von Hagen 1999, 433–36). The EMS had the potential to recreate the experience of Bretton Woods, with the Bundesbank forced to create money and inflation by supporting a weak franc instead of a weak dollar. If France wanted the EMS, the Banque de France would have to subordinate its monetary policy to pegging the franc to the mark. Karl Otto Pöhl, Bundesbank president from 1979 to 1991, said, "The Bundesbank turned the original concept [of the EMS] on its head by making the strongest currency the yardstick for the system" (Marsh 1992, 203).

Starting in 1979, the Bundesbank began to pursue the goal of price stability. Not only the EMS, but also inflation motivated the change in policy. From a low of 2.7 percent in 1978, CPI inflation rose to 6.3 percent in 1981 (Figure 6). The Bundesbank lowered its monetary target range from 1979 (6 to 9 percent) through 1985 (3 to 5 percent). Only in one year of this period, 1983, did money growth slightly exceed the target range. Money (M3) growth fell from 10 percent in the 1970s to 6 percent in the 1980s. By retaining price stability as its primary objective despite the high unemployment rates of the 1980s, the Bundesbank gained credibility for its policy of price stability.

The Bundesbank as Guarantor of Stability

From modern Germany's postwar inception in 1949, its polity has been corporatist: the major organized groups—political parties, corporations, and trade unions—determine the economic and political consensus. By the 1980s, this

[27] The most complete account of the history of European monetary union is in Dyson and Featherstone (1999) and Connolly (1995).

framework for achieving consensus had lost its balance. The Bundesbank restored balance by joining that framework as the representative of stability.

The German corporate consensus exercised a dramatic effect on the labor market in the 1950s and 1960s (Giersch et al. 1992, Chapter 4, Section A). The labor unions kept real wages below their market clearing value to help Germany regain its prominence as one of the world's great exporters of manufactured goods. An influx of foreign workers met the resulting labor shortage. In the early 1950s, the unemployment rate fell slowly because of the problems in absorbing German-speaking immigrants. In the 1960s, however, it generally remained well below 1.5 percent (Figure 1).

As described above, in the latter part of the 1960s under the Bretton Woods system, the rise in inflation in the United States rendered the dollar overvalued (the mark undervalued).[28] The profits of German corporations soared because of the export boom set off by the undervalued mark. At the same time, imported inflation eroded the real value of nominal wage contracts. CPI inflation, which had averaged 1.6 percent in 1967, 1968, and 1969, rose steadily until it reached 7 percent in the early 1970s. In response, labor unions broke the corporatist social contract and launched a wave of wildcat strikes in autumn 1969.[29] Pushed by "shop floor radicals," major unions like chemicals and autos went on strike in the early 1970s (Johnson 1998, 72–73, 90–95).

The postwar German consensus that had produced the Wirtschaftswunder continued to erode in the 1970s. In 1978, the printers' union went on strike to prevent the introduction of labor-saving technology. Also in the late 1970s, the Social Democratic Party began to identify with the program of labor. This identification "endangered the consensual pillars of corporatism" (Giersch et al. 1992, 214–16).

With the full-employment pledges by the Brandt and Schmidt governments, with labor unions desirous of large wage increases, and with corporations reluctant to allow appreciation of the mark, the Bundesbank became the member of the corporatist framework defending "stability"—price stability and balanced budgets. Moreover, when the Bundesbank agreed on the primacy of maintaining the DM's internal value, it could represent the general public desire for economic stability. As Otmar Emminger said in his inaugural speech upon becoming Bundesbank president in 1977, "Monetary stability is linked up with general social stability—and with political stability" (Marsh 1992, 37).

---

[28] Giersch et al. (1992, 164–66) document the sharp rise in the German trade surplus.

[29] Giersch et al. (1992, 154–58) discuss reasons for the "sudden switch of union behavior from moderation to aggressiveness in the late 1960s and early 1970s." Holtfrerich (1999, 387) quotes Schiller (Minister of Economics) as arguing for a DM revaluation in 1969 to eliminate the threat to "social symmetry" produced by the surge in corporate profits. Rising inflation that eroded the value of collective wage agreements was a social "bomb."

Marsh (1992, 145) writes, "If it [the Bundesbank] feels inflationary pressures are getting out of hand, the central bank reserves the right to confront the politicians, industrialists, and trade unionists who exert the main influence on corporate Germany." However, the Bundesbank could not confront the unions directly. To do so would come "dangerously close to compromising their constitutionally guaranteed right to autonomous wage negotiations" (Marsh 1992, 145). The Bundesbank could not on its own conduct a "disguised incomes policy," that is, tell the unions what wage increases to negotiate (Johnson 1992, 92, 94). The Bundesbank could, however, use its money targets to make its objective of price stability credible and thus exercise indirect influence over wage negotiations. Those procedures became "an integral component of German 'stability culture.'"(Schmid 1996, 42).[30]

The Bundesbank's stability policy succeeded because of widespread public support.[31] Capie and Woods (2001), in their review of *Fifty Years of the Deutsche Mark,* cite Richter (1999, 562):

---

[30] The Bundesbank used a quantity theory framework to derive a target for money growth. There are many descriptions of these procedures. See, for example, Schlesinger (1984) and Schmid (1996). The Bundesbank began by setting a target for growth in nominal aggregate demand (nominal output). The target for growth in nominal demand had two components. One was an estimate of the trend rate of growth of real output. (By using trend growth, the Bundesbank sent the message that monetary policy was not an appropriate instrument for countercyclical aggregate demand management.) The other component was "unavoidable" inflation.

In 1986, the Bundesbank restored price stability and discarded the idea of "unavoidable inflation" for an inflation target of 2 percent or less. The Bundesbank worked hard to achieve a consensus for its inflation target from government, labor, and business (Schlesinger 1984). The Bundesbank then set a money growth target equal to the target for nominal output growth minus estimated growth of monetary velocity. Clarida, Gali, and Gertler (1998) estimate a monetary policy reaction function for Germany, which they interpret as evidence of inflation targeting.

[31] Baltensperger (1999, 462–63) reviews "the often highly controversial and turbulent debate" over monetary policy in the early 1980s, but concludes, "all in all, Bundesbank policy enjoyed broad public support." Why this support despite the sustained rise in the unemployment rate over the period from 1973 through 1983 (Figure 1)? Connolly (1995, 33, 301) contends that Germans liked making the EMS into an "undeclared DM-zone" and having other currencies devalue relative to the DM. More important, the willingness of most major countries to assign to their central banks primacy for a price stability objective attests to the importance of a fundamental change in the political and intellectual environment. Kitterer (1999, 192) points to the "failure of demand management despite a sharp rise in the public sector deficits" and to "the simultaneous sluggish growth and high inflation of the period." Countries became disillusioned with the ultimate ineffectiveness and the inflationary consequences of the aggregate demand policies of the 1970s. Within Germany, the more conservative environment appeared with the program of fiscal consolidation of the CDU/CSU and FDP coalition of Helmut Kohl that replaced the SPD/FDP coalition of Helmut Schmidt in October 1982 (Giersch et al. 1992, 192–96).

In the postwar period, the German government had become increasingly leftist and interventionist (Giersch et al. 1992, 125). Although the government had always regulated the economy, "public subsidization and heavy legal regulation of economic activity…took on a new qualitative dimension in the 1970s and 1980s" (Giersch et al. 1992, 216). For a while in the 1970s, Keynesian aggregate-demand policies had offset the rise in unemployment produced by the resulting increased inflexibility in labor markets. However, inflation vitiated the effectiveness of those policies. By the 1980s, "public opinion and the vast majority of academic economists…supported the Bundesbank's line, if only because there seemed to be no realistic alternative" (Giersch et al. 1992, 193).

> The German public... after having lost their savings twice within 25 years [1923 and 1948], definitely wanted a stable currency.... No Bonn government in its right mind would have... put the Bundesbank under pressure.

The Bundesbank as European Central Bank

By 1980, almost all the major industrial countries had rejected the activist policies of the 1970s that had led to inflation. France alone retained expansionary fiscal and monetary policies. In May 1981, François Mitterand became president of France. He pursued a program of government intervention in the economy and expansion of aggregate demand. Capital flowed out of France and the franc weakened. When the Bundesbank refused to lower interest rates and inflate to support a weakened franc, France had to devalue. In a series of devaluations ending March 1983, the value of the French franc fell by 30 percent against the DM.

With inflation and a weak currency, France could not exercise the same leadership within Europe as Germany.[32] After March 1983, France followed conservative fiscal policies, and the Banque de France gave priority to preserving the parity of the franc with the DM. The combination of Bundesbank commitment to internal price stability and the peg of the franc to the DM determined the character of the EMS. Continental Europe gained a central bank—the Bundesbank—when other countries chose to peg to the DM even at the expense of their own domestic monetary policies (Baltensperger 1999, 440). A stability-oriented Bundesbank policy of monetary targeting designed to achieve price stability provided a nominal anchor for the EMS. At the same time, de facto establishment of the Bundesbank as the European central bank gave France an incentive to regain influence over monetary policy by creating a de jure European central bank.

Backsliding with the Louvre Accord

In 1987, the Louvre Accord initiated expansionary and ultimately inflationary monetary policies among the world's major industrial countries.[33] The Louvre Accord and simultaneous EMS problems due to weakness in the franc pushed the Bundesbank away from its stability-oriented policy (Baltensperger 1999, 466–75). In January 1987, labor unrest in France weakened the franc, and the Bundesbank intervened in the foreign exchange market to prop up the franc.

---

[32] Comments on French politics reflect the themes developed in de Boissieu and Pisani-Ferry (1999).

[33] The political problem was the rise in U.S. protectionism produced by a large U.S. current account deficit. The Plaza Accord of September 1985 had attempted to use coordinated intervention

Already by 1986 the Bundesbank had significantly overshot its target range for money. Nevertheless, in early 1987, the Bundesbank reduced the repurchase rate. France still had to devalue the franc.

The German money stock continued to overshoot its target range through 1987. In response, early in October, the Bundesbank nudged its repurchase rate up slightly. On Friday 16 October, U.S. Treasury Secretary James Baker criticized Germany for backing off its pledge to stimulate its economy (Connolly 1995, 40). On Monday 19 October, the U.S. stock market crashed.[34]

Concerned that the fall in stock prices would depress economic activity, central banks lowered interest rates. In 1988 in Germany, money again overshot the top of its target range. The CPI, which had remained basically unchanged from 1985 through 1988, began to rise sharply in 1989. The Bundesbank's attempt to resist an appreciation of the mark led to excessive money growth and inflation just as it had in the early and the late 1970s. The Bundesbank kept its repurchase rate basically steady in 1986 and lowered it in 1987 despite significant overshoots in its money targets in both years. German CPI inflation climbed steadily from -0.1 percent in 1986 to 4 percent in 1992 (Figure 6).

3. EUROPEAN MONETARY UNION

German Chancellor Kohl and French Prime Minister Mitterand put the EMU on the agenda for discussion in 1988 and 1989. "But, once German unification was set under way, the EMU became endowed with a new significance: as a test of the political resolve of a unified Germany to bind itself into Europe" (Dyson and Featherstone 1999, 369). The Bundesbank accepted the decision

---

in the foreign exchange market to depress the value of the dollar. Policymakers hoped that a depreciated dollar would lower the U.S. deficit by stimulating U.S exports. However, a steady depreciation of the dollar (Figure 4), which had already begun in early 1985, failed to lower the U.S. current account deficit.

The U.S. Treasury then lobbied other countries to expand domestic demand as a way to reduce their trade surpluses and lower the U.S. trade deficit. The agreement reached at Paris in February 1987 required Japan and Germany to stimulate their economies to increase imports while the United States reduced its fiscal deficit. Other countries saw the U.S. fiscal deficit as the cause of the U.S. current account deficit. The Gramm-Rudman-Hollings agreement to balance the U.S. budget gave credibility to the U.S. side of the Louvre agreement.

At the time, worldwide stimulus appeared acceptable. The disinflationary monetary policies of the first half of the 1980s had tamed inflation everywhere. Together with the transitory effect of the fall in oil prices, virtual price stability had emerged in 1986. Also, the desire of Japan and Germany to prevent further appreciation of their currencies against the dollar encouraged them to reduce interest rates. For a discussion of this period, see James (1996, 433–53), Solomon (1999, 21–29), and Volcker and Gyohten (1992, 248–58).

[34] Consensus is difficult to establish over the causes of a one-time event like the crash, but the Louvre Accord must have been one of them. It promised an end to dollar depreciation through coordinated government policies. The public rift between the U.S. Treasury and the Bundesbank cast doubt on the viability of that coordination and, therefore, on the value of the dollar. With the value of the dollar suddenly open to question, foreign investment in the United States became less attractive. The fall in the U.S. stock market quickly spread to foreign markets.

to replace it with a European central bank. At the same time, the Bundesbank worked to ensure that its successor would continue its policy of price stability. That continuity required an explicit mandate for price stability with the force of a treaty among countries. To bequeath the Bundesbank's credibility to the new ECB, the Bundesbank also lobbied for the replication of its own institutional structure. Finally, the Bundesbank pursued a monetary policy that would enable the new central bank to begin operation in an environment of price stability. To do so, it had to undo the post-Louvre inflation of about 5 percent (Figure 6). That task, which took place in an extremely difficult political environment, constituted one of the great successes of central banking.

German Reunification

The Berlin Wall fell in November 1989. On 6 February 1990, Chancellor Kohl decreed that West Germany would exchange the DM for the ostmark at a ratio of one to one, which compared with a free market exchange rate of 7 to 1 (Marsh 1992, 178). Monetary union on 1 July 1990 came before reunification on 3 October 1990.

The decision to reunify required immediate monetary reform. East Germany knew that West Germany would exchange ostmarks for DMs not at their black market rate, but rather at a politically determined rate. Chancellor Kohl, mindful of the symbolism of the one-for-one exchange in 1948 between the old reichsmarks and the new DMs, decided to exchange East for West German currency at a one-for-one rate.[35] East Germany could therefore obtain resources just by printing ostmarks.[36]

---

[35] However, while the 1948 reforms had allowed the market to determine relative prices, the 1990 reform superseded the market. To avoid an inflow of East German workers into West Germany, Chancellor Kohl decided to raise real wages in East Germany to move them toward parity with those of West Germany (Marsh 1992, 183, 187). The West German government converted East German wage rates one for one into DM wage rates, despite the argument of the Bundesbank for a two-to-one conversion (Streit 1999, 660–62). Moreover, Germany granted East German workers large pay raises after unification. "Following monetary union with West Germany in 1990, the real wage of East German workers rose 83%" (Hunt 2001, 190).

The attempt to move toward real wage parity with West German workers conflicted with economic reality. East German workers were less productive than West German workers. In 1991, productivity in East Germany (output per worker) was only 30 percent of the West German level (Marsh 1992, 171). The economics of that political decision meant that Germany had to raise the capital-labor ratio in East Germany. However, the effort to make East Germany into a capital-intensive economy like West Germany, when the low productivity of its workers demanded a labor-intensive economy, created large-scale unemployment in East Germany.

[36] In March 1991, Kohl faulted Bundesbank president Pöhl publicly for the latter's criticism of the terms of monetary union with East Germany. Marsh (1992, 147) commented on the relationship between the Chancellor and the Bundesbank president: "Bound by the common desire to maintain confidence in the conduct of state affairs, the two are condemned to harmonious coexistence." Lacking a harmonious working relationship with Kohl, Pöhl resigned on 15 May 1991.

Maastricht and the Birth of the Euro

The same political forces that had brought about the EMS now led to the EMU. In the early 1980s, Germany wanted to undertake its diplomatic initiatives within the context of building a united, democratic Europe, thereby lessening other countries' fears of a Europe dominated by Germany.[37] France wanted to constrain future German influence within Europe through pan-European organizations (Marsh 1992, 198).

After the EMS crisis of March 1983, France abandoned its policy of aggregate demand expansion to lower unemployment in an attempt to remain within EMS with no further devaluations. France thus fulfilled a condition necessary for movement toward monetary union with Germany. These actions required a prior fundamental decision by President Mitterrand to reshape his presidency by abandoning the agenda to substitute socialism for a free market economy and replacing that agenda with *construction européenne* (*Le Monde,* 19 May 2001, 1; Dyson and Featherstone 1999, 199).[38]

In 1986, Europe moved closer to economic integration when it passed the Single European Act, which required the abolition of all remaining trade impediments within the European Union. Supporters of the European Union wanted Europe to possess an economic and political stature comparable to that of the United States. They talked of Eurosclerosis and believed that European integration would make European firms competitive with American firms. France especially wanted to create a European economic bloc that would rival the United States in economic influence (Boissieu and Pisani-Ferry 1999, 68).

In 1988, Hans-Dietrich Genscher, the German foreign minister; Edouard Balladur, the French Finance Minister; and Jacques Delors, the president of the European Commission (EC), began the process that would lead to the creation of a European central bank. Behind them stood Chancellor Kohl and President Mitterand. The European Council met in Hanover in June 1988 and set up the Delors Committee to devise a plan for a single currency. The Committee delivered its report at the Madrid European Council Meeting in June 1989.[39]

---

[37] For example, *Le Monde* (1 May 2001, 8) wrote: "It is for Germany to regain the power that it lost in 1945. However, because of its past, in order not to scare its partners, it can only do so through Europe."

[38] "When President François Mitterrand opted in 1983 for what came to be called the franc fort policy, he was effectively discarding most of his election commitments to old-fashioned socialism in France" (*Financial Times* 31 July–1 August 1993, 2). (Note that in French the pronunciation of the term "franc fort" is the same as "Frankfurt," the home of the Bundesbank.)

The extraordinary depth of this commitment appeared in France's willingness to alter fundamentally the dirigiste character of its economy. At the beginning of Mitterrand's presidency, nationalized firms made up a quarter of the French economy, with the remainder heavily regulated or subsidized. France was protectionist and retained capital controls. European integration required economic and financial liberalization. Within a decade, France moved dramatically away from its traditional interventionist model of government control (de Boissieu and Pisani-Ferry 1999, 56–57).

[39] Connolly (1995) and Vanthoor (1999) provide this chronology. The most detailed account is in the early chapters of Dyson and Featherstone (1999).

The conviction that the EMU would advance the political unification of Europe united the participants (Dyson and Featherstone 1999, 273).

The unification of Germany provided the impetus for governments to make the hard political decisions necessary for the realization of the EMU. A reunified Germany would not only be stronger economically, but also would lie in the center of a Europe rejuvenated by the fall of communism. Possessed of a sense of history, Chancellor Kohl wanted to shape two great historical movements: German reunification and European federation. With the fall of the Berlin Wall in November 1989, those movements came together.

Kohl wanted a reunified Germany, but not a Germany that Europe would fear as a bully. For Kohl, European monetary union was the instrument that would bring reunification and European federation together.[40] He said, "Political union and economic and monetary union are inseparably linked. The one is the unconditional complement of the other" (Marsh 1992, 211).

At Maastricht in December 1991, members of the European Union signed the Treaty on European Union, which laid down conditions for membership in the EMU. Monetary union required that Germany forsake the DM—the symbol of everything that it had done right after World War II. German public opinion could accept such a sacrifice only if EMU members adopted the German model of stability symbolized by the Bundesbank (*New York Times* 15 March 2000, C1). France acceded for the sake of its foreign policy objectives of countering U.S. worldwide hegemonic influence and German European hegemonic influence (Dyson and Featherstone 1999, 252; Marsh 1992, 204).

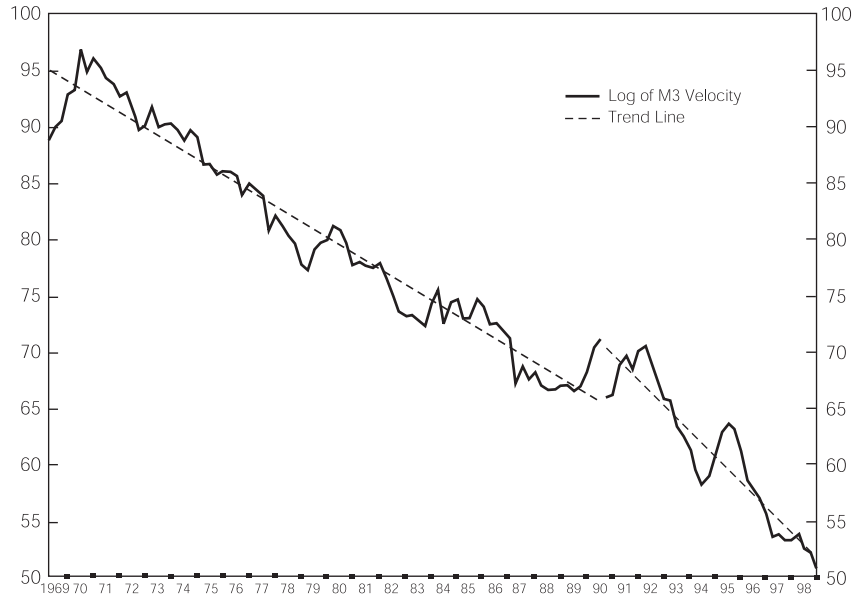
In 1990, Bundesbank president Pöhl chaired the Committee of EC Central Bank Governors that drafted the ECB statute, and the Bundesbank prepared the single draft for negotiations. The Bundesbank worked to preserve its Stabilitatspolitik.[41] It replaced the ambiguous reference in the Bundesbank Law of 1957 to "safeguarding the currency" with the explicit language, "the primary objective of the ESCB shall be to maintain price stability."[42] Furthermore, it included the statement that "the ESCB shall act in accordance

[40] Dyson and Featherstone (1999, 307–08) write that German unification provided Kohl with "an opportunity to make an historic contribution to unifying Europe and to reinvent himself as an historic Chancellor.... His European vision was a generational as well as personal matter. It was bound up with a notion of a special historical responsibility to create a Europe that would never again experience the horrors of 1914–18 and 1933–45."

[41] Realization of the EMU required convincing Germans that its members would accept the German stability culture of fiscal responsibility. To be eligible to join the EMU, countries had to meet convergence criteria that included guidelines on inflation and budget deficits (Connolly 1995, 79; Vanthoor 1999, 131; Dyson and Featherstone 1999, 3). By making the EMU potentially into a club of outsiders and insiders, Maastricht created an enormous incentive to meet these criteria. The Bundesbank's role was then to make certain that governments did not relax the convergence criteria.

[42] The ESCB is the European System of Central Banks. All members of the EU belong to it. The ECB is the European Central Bank, and only a subset of EU countries are members.

Figure 8 M3 Velocity



Notes: Observations are quarterly values of the logarithm of German velocity. Velocity is the ratio of nominal GDP to M3. Until 1990Q2, figures are for West Germany. Thereafter, they are for unified Germany. The slope of the trend line fitted from 1969Q1 through 1990Q2 is −1.4. From 1990Q3 through 1998Q4, the slope is −2.3. Heavy tick marks indicate fourth quarter.

with the principle of an open market economy with free competition, favoring efficient allocation of resources" (Dyson and Featherstone 1999, 387–89).[43]

The Breakdown of the EMS

German reunification created economic as well as political shock waves. The increased demand for capital investment in a united Germany required Germany to change from a capital exporter into a capital importer. To provide the additional resources needed in Germany, Germans would have to buy more from foreigners, who would in turn have to buy less from Germans. In 1989,

[43] This statement required monetary arrangements compatible with free markets. Such arrangements imply control of the price level through monetary control rather than through government intervention in the marketplace in the form of incomes policies. Monetary control cannot come from selected credit control and government credit rationing. Such arrangements require freely floating exchange rates rather than pegged exchange rates maintained by capital controls.

Germany's current account surplus was almost 5 percent of GDP. In 1991, Germany moved to a current account deficit equal to 1 percent of GDP (Whitt 1994, 23). This reversal required that prices in Germany rise more than the prices of its trading partners. The required relative change in international price levels could have happened automatically through a revaluation of the DM, as desired by the Bundesbank; however, France refused the required mirror devaluation of the franc. Why?

The enhanced role of a reunified Germany in Europe made the French objective of intertwining the destinies of Germany and a united Europe all the more pressing. Accordingly, the goal of European monetary union took on greater urgency. France wanted the same prestige as Germany in negotiations over the design of the ECB, not to be treated as a weak currency country (de Boissieu and Pisani-Ferry 1999, 63–64). Monetary union would also be less palatable in Germany if the most important currency after the DM, the franc, was seen as weak. Moreover, at considerable political cost, France had pursued restrictive economic policies since 1983 to prevent franc devaluations, and it did not want its sacrifice to have been in vain.

The alternative to revaluing the mark was either inflation in Germany or disinflation by Germany's EMS partners. In the event, both occurred to some extent, along with devaluations by some EMS members. The initial unwillingness of other countries to devalue required them to disinflate. Immediately following reunification, this bitter medicine appeared to work and the currency parities of the EMS held. However, the EMS came apart in 1992 and 1993.

In 1992, German money growth exceeded the 5 1/2 percent top of the Bundesbank target range, and in July the Bundesbank raised its discount rate. The discount rate did not affect the money market rate, which was determined by the repurchase rate. However, the rise was enough to make the financial markets doubt the ability of other EMS countries to maintain the punishing level of interest rates necessary to prevent the depreciation of their currencies. In September 1992, the markets forced Britain and Italy out of the EMS, and Spain, Portugal, and Ireland devalued.[44]

One can understand the unyielding actions of the Bundesbank in this period in the context of the decision taken at Maastricht the previous December to proceed with monetary union. In 1992, CPI inflation in Germany was 4 percent (Figure 6). Although the Bundesbank might have to relinquish control over monetary policy to the new ECB, it was going to do so in a way that preserved its stability culture. The convergence criteria for joining the EMU, which enshrined fiscal rectitude and price stability, would mean little

[44] Estimates put the Bank of England's losses from intervention in the foreign exchange market at approximately six billion dollars. George Soros gained one billion dollars. The figures are, respectively, from Pringle (1996, 123) and Hanke and Walters (1994, 141).

The name given to the day Britain left the ERM, Black Wednesday, reveals the drama of the events. *The Financial Times* (9–10 March 1996) wrote later: "[John] Major [British Prime

if Germany itself did not itself enter into monetary union with price stability (Marsh 1992, 217). In 1992, the Bundesbank could not ignore the significant overshoot in its money target, regardless of the consequences for the EMS.

During the fall 1992 exchange rate crisis, the franc held. Michel Sapin, the French finance minister, reminded the markets that speculators had been beheaded during the French Revolution. More important, the Banque de France followed a highly restrictive monetary policy. (The degree of restriction appears in Figure 9 in the sharp increase in the interest rate difference between France and Germany.) Restrictive monetary policy in France led to weakness in the French economy in 1993.[45]

In July 1993, financial markets forced France to allow the franc to depreciate by 6 percent relative to the central rate within the EMS exchange rate system. Speculators assumed that France and the other countries forced to devalue would lower interest rates to stimulate their economies. However, France, Denmark, and Belgium maintained the existing high level of interest rates. These countries feared that a "competitive devaluation" would split Europe into weak and strong currency blocks and create protectionist pressures (*Financial Times* 3 August 1993, 2; Boissieu and Pisani-Ferry 1999, 78). By maintaining a restrictive monetary policy, France returned the franc in 1996 close to the central EMS rate of 3.35 francs to the mark. The small difference between French and German interest rates showed that the mark/franc peg had become credible (Figure 9).

France restored the external value of the franc by maintaining a lower inflation rate than Germany. Through 1990, France had higher inflation. In

---

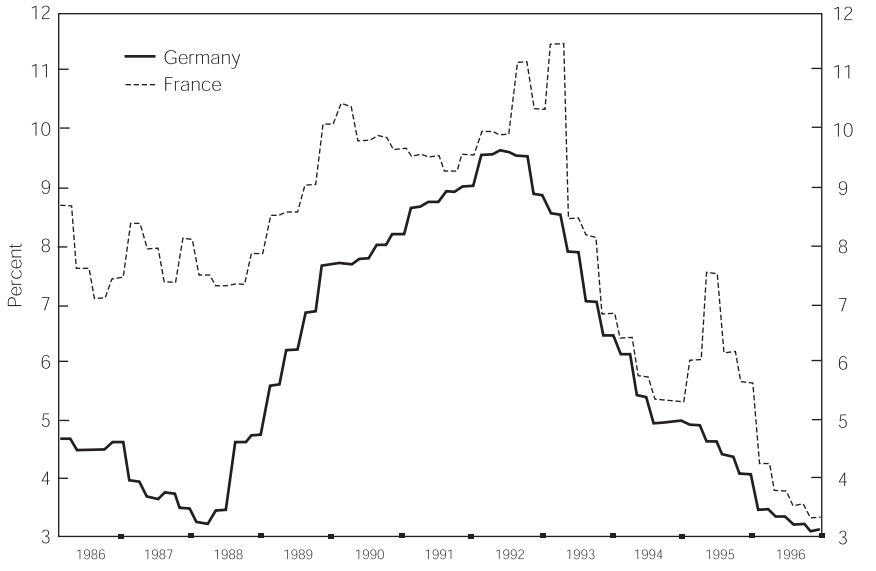
Minister] now fell into a perilous trap. Like Winston Churchill and Harold Wilson before him, he treated sterling's exchange rate as a badge of national pride.... Interest rates had been raised to 15 percent. More than $30 billion had been thrown in vain at the markets, all but exhausting the reserves. The government's economic policy lay in ruins. Major's political reputation was in shreds, his party torn asunder. The failure robbed the Prime Minister of the authority of his office."

Norman Lamont, Chancellor of the Exchequer during the ERM crisis, blamed the Bundesbank for Britain's forced departure. He believed that the Bundesbank should have lowered German interest rates to help defend the pound (Lamont 1999–2000). England's Prime Minister Major and Chancellor Lamont were especially unhappy over an interview in the German newspaper *Handelsblatt* reporting a statement by Helmut Schlesinger, Bundesbank president, that "further devaluations are not excluded." That newspaper article precipitated the final crisis (Frowen 1999–2000).

[45] In response to this weakness, the French lowered interest rates. Initially, the franc held against the mark. However, in July, the French economic institute, INSEE, predicted that in 1993 French GDP would fall 0.7 percent (Stanley 1993, 3). Speculators bet that a weak French economy would force the Banque de France to lower rates further and abandon the franc-mark parity. At the same time, the Bundesbank, concerned about a pickup in money growth, refused to cut its discount rate. The unwillingness of the Bundesbank to lower rates to defend the EMS weakened the credibility of the EMS and speculators attacked the franc.

The *New York Times* (31 July 1993, 47) wrote: "The attack on the franc came after a decision Thursday by the Bundesbank to fight German inflation by refraining from a cut in its discount rate.... For France and Mr. Balladur [French Prime Minister], the options look bleak: either devalue, which would amount to a political humiliation after vows of resistance, or try to hang on, at potentially devastating further cost to the economy.... Already last week, the Bank of France was obliged to raise overnight rates to 10 percent from 7.75 percent."

Figure 9 German and French Money Market Rates



Notes: Observations are monthly averages of money market rates on day-to-day money for Germany and France. Heavy tick marks indicate December observations.

the subsequent four years, it maintained an inflation rate somewhat more than 1 percentage point lower than Germany. By 1995, France had come close to achieving price stability with an inflation rate of about 1.5 percent. In contrast, from 1991 to 1994, Germany had an inflation rate of about 3.5 percent.

Was Bundesbank Policy Inflationary or Disinflationary?

A characterization of German monetary policy in the early 1990s turns on whether the Bundesbank pursued an expansionary policy to relieve the exchange rate stresses in the EMS. Alternatively, did it attempt to reestablish price stability in an effort to help the new ECB begin operation in an environment conducive to the establishment of credibility? The answer depends upon how one assesses the high rates of money growth in the years 1992, 1993, and 1994 (Figures 6 and 7).

To begin, the inflation of the early 1990s derived from the expansionary character of earlier monetary policy. On 1 July 1990, monetary union increased the monetary aggregate M3 by 15 percent. However, unified German

GDP increased only 8 percent. At the time, the Bundesbank put the resulting monetary overhang at 5 percent (Baltensperger 1999, 479). This increase in M3 combined with the earlier rapid increases begun in 1987 explains the inflation rates of the early 1990s.

The Bundesbank did not pursue a stimulative monetary policy as it would have had to do in order to prevent the depreciation of the franc. In August 1992, in response to strong economic growth, the Bundesbank had raised the repurchase rate to almost 10 percent. Only when signs of recession began to appear did it lower rates significantly. When in July 1993 the franc fell to the level requiring central bank intervention, the Bundesbank did not lower its rates. Newly appointed Bundesbank president Hans Tietmeyer rejected the demand "that Germany must immediately abandon its monetary sovereignty" (Connolly 1995, 324).[46]

Low real growth accompanied restrictive monetary policy. Annualized real GDP growth in Germany averaged only 0.8 percent over the years 1992 through 1997. In return, by 1998, German CPI inflation averaged less than 1 percent. French CPI inflation was also less than 1 percent in 1998. The social cost was high. In France, the unemployment rate was 9 percent in 1990 and 12 percent in 1993. In Germany, the unemployment rate reached 12 percent in 1998 (Figure 1).

Nevertheless, together, the Bundesbank and the Banque de France bequeathed a priceless gift to the new ECB. They created virtual price stability. If in 1999 the new ECB had put Europe through a recession in order to lower inflation, it might not have survived as an institution.

4. CAN THE ECB BECOME THE BUNDESBANK AND THE EURO THE MARK?

Can the ECB summon wide support for a policy of price stability? An affirmative answer will require that Europeans accept the ECB and its objective

[46] Figures 6 and 7 reveal high money growth in the early 1990s. However, that money growth corresponded to an unusually high demand for money associated with turbulence in currency markets and the view that the mark was immune from devaluation. It seems likely that the probable division of Europe into strong and weak currency blocks after the breakup of the ERM led to a general substitution into the DM. The DM also became a "parallel currency" in some East European countries (Baltensperger 1999, 482). Figure 8 shows the behavior of M3 velocity. Adjusted for trend, the demand for M3 (measured by the inverse of velocity) was reasonably stable through mid-1989. Velocity rose after unification. However, as indicated by the more steeply sloped trend line, money demand grew unusually fast beginning in 1992.

The high money growth step from 1991Q3 through 1994Q2 does not lead to a subsequent high inflation step—the only exception in the figure (Figure 7). Instead, inflation fell steadily beginning in 1993. Although the Bundesbank allowed M3 growth to exceed its target range in 1992 and 1993, the overshoot failed to compensate for the increased money demand. Starting in 1994Q3, the Bundesbank maintained trend M3 growth below 5 percent—a historically low rate.

of price stability as part of a constitutional framework. Such a framework limits government discretion. The clearest example of constitutional limitations on the sphere of government action is the protection of fundamental human rights. Freedom of speech is not subject to majority vote as part of the normal political process.

The alternative to making monetary policy part of the constitutional framework is to make the purchasing power of money subject to ongoing democratic debate. Just as they do for the farm subsidies of the European Union, constituencies would organize and lobby on behalf of the objective of "low" unemployment rather than price stability. If Europeans come to believe monetary policy should be part of the democratic process rather than a constitutional framework, they will see the ECB as undemocratic and elitist. Political attacks will then erode its legitimacy. To maintain its independence and support for the objective of price stability, the ECB must explain why a constitutional framework should constrain monetary policy.[47]

The ECB can point out that in the 1970s, Europe pursued policies meant to manage real aggregate demand and achieve socially desirable low unemployment rates. Political pressures to reconcile conflicting objectives for low unemployment and price stability created a demand for incomes policies that would control the price setting of private markets. Such pressures threatened both central bank independence and free markets.[48]

The Maastricht Treaty appropriately specified price stability as the objective for the ECB. However, the EMU's geographical composition derives from a political, not an economic, consideration. Europeans desired an EU-wide symbol that would promote an EU-wide identity. The EMU is a mechanism to promote the political unification of Europe. But imposition of a common monetary policy on an economically diverse area entails a cost.

EMU member countries experiencing an adverse change in their terms of trade with other member countries will experience deflation.[49] Increasing the overall EMU-wide inflation rate to deal with periodic regional deflationary stresses will not avoid the need for real economic adjustment. Regional deflations will be the price incurred for a common symbol. They will not imply that the EMU or its objective of price stability is inappropriate.

---

[47] Otto Pfleiderer (quoted in Buchheim 2002, 3), later a member of the BdL, expressed the alternative view in 1946. "Only the government itself could in a democracy bear responsibility for the principal measures of monetary policy... which is an important part of general economic policy.... There is the danger that the central bank would develop as a kind of second government. As such it would be able to counteract the economic policy aims of the responsible government."

[48] A commitment to EU-wide free markets is the necessary precondition for the economic integration of Europe. Modern central banks like the Bundesbank control the price level through management of their balance sheet. They forswear the direct government interference in markets entailed by incomes policies.

[49] A specific example can clarify the issue. After the creation of the EMU in January 1999, the Euro depreciated due to outward capital flows, especially to the United States. However, for

Europeans now have the opportunity to make the Euro the kind of symbol that the DM was for Germany in the postwar period.[50] Europeans can now create the right future for the Euro to represent. The Euro can then become the symbol of a prosperous, democratic, and unified Europe.

REFERENCES

Arestis, Philip, Kevin McCauley, and Malcolm Sawyer. 1999. "From Common Market to EMU: A Historical Perspective of European Economic and Monetary Integration." Working Paper 263, the Jerome Levy Economics Institute.

Baltensperger, Ernst. 1999. "Monetary Policy under Conditions of Increasing Integration (1979–96)." In *Fifty Years of the Deutsche Mark,* ed. Deutsche Bundesbank. Oxford: Oxford University Press.

Bernholz, Peter. 1999. 'The Bundesbank and the Process of European Monetary Integration." In *Fifty Years of the Deutsche Mark*, ed. Deutsche Bundesbank. Oxford: Oxford University Press.

Board of Governors of the Federal Reserve System. 1976. *Banking and Monetary Statistics: 1941–1970.*

Buchheim, Christoph. 1999. "The Establishment of the Bank Deutscher Länder and the West German Currency Reform." In *Fifty Years of the Deutsche Mark,* ed. Deutsche Bundesbank. Oxford: Oxford University Press.

_____. 2002. "How Did the Bundesbank Become Independent?" Manuscript, University of Mannheim.

countries enjoying capital inflows within the EMU, that depreciation was not appropriate. For example, Ireland and the Netherlands experienced above-average EMU inflation rates, which kept their terms of trade with the rest of the world from deteriorating. The analogue to these countries for the United States was Argentina, which had a monetary union with the United States through its currency board. Due to capital inflows, the United States experienced an appreciation in its terms of trade. Because that appreciation was not appropriate for Argentina, which experienced capital outflows, it had to deflate.

[50] Germany endowed its 1948 currency reform with vitality by combining it with a vast deregulation, namely, an end to price controls. For Europe today, the analogous measure would be to deregulate its labor markets. Milton Friedman (1953, 187) identifies the major problem that governments must confront in a world of ongoing change: They must ignore "the urge for security that is so outstanding a feature of the modern world and that is itself a major source of insecurity by promoting measures that reduce the adaptability of our economic systems to change without eliminating the changes themselves."

Bureau of the Census, U. S. Department of Commerce. 1975. *Historical Statistics of the United States, Colonial Times to 1970, Part 2*.

Clarida, Richard, Jordi Gali, and Mark Gertler. 1998. "Monetary Policy Rules in Practice: Some International Evidence." European Economic Review 42 (June): 1033–67.

Connolly, Bernard. 1995. *The Rotten Heart of Europe.* London: Faber and Faber.

de Boissieu, Christian, and Jean Pisani-Ferry. 1999. "The Political Economy of French Economic Policy in the Perspective of EMU." In *Forging an Integrated Europe,* ed. Barry Eichengreen and Jeffry Frieden. Ann Arbor: The University of Michigan Press.

Deutsche Bundesbank. 1992. "Commentaries." *Deutsche Bundesbank Monthly Report* 44 (August).

Duisenberg, Willem F. 1999. "The Past and Future of European Integration—A Central Banker's View." The Per Jacobsson Lecture, Washington.

Dyson, Kenneth H. F., and Kevin Featherstone. 1999. *The Road to Maastricht: Negotiating Economic and Monetary Union.* Oxford: Oxford University Press.

Financial Times. 1931. "The Nightmare Facing French Policymakers: Pursuing Rational Economics Means Political Humiliation." 31 July–1 August.

_____. 1993. "Welcome Flexibility, Goodbye Simplicity." 3 August.

_____. 1996. "The Countdown to Meltdown." 9–10 March.

Friedman, Milton. 1953. "The Case for Flexible Exchange Rates." In *Essays in Positive Economics,* ed. Milton Friedman Chicago: The University of Chicago Press.

Frowen, Stephen. 1999–2000. "Unjustified British Critique of the Bundesbank." *Central Banking* 10: 40–45.

Giersch, Herbert, Karl-Heinz Paqu, and Holger Schmieding. 1992. *The Fading Miracle: Four Decades of Market Economy in Germany.* Cambridge: Cambridge University Press.

Hanke, Steve H., and Sir Alan Walters. 1994. "Easy Money." *Forbes*, 31 January.

Hetzel, Robert L. 1999. "Japanese Monetary Policy: A Quantity Theory Perspective." Federal Reserve Bank of Richmond *Economic Quarterly* 85 (Winter): 1–25.

———. 2002. "German Monetary History in the First Half of the Twentieth Century." Federal Reserve Bank of Richmond *Economic Quarterly* 88 (Winter): 1–35.

Holtfrerich, Carl-Ludwig. 1999. "Monetary Policy under Fixed Exchange Rates (1948–70)." In *Fifty Years of the Deutsche Mark,* ed. Deutsche Bundesbank. Oxford: Oxford University Press.

Hunt, Jennifer. 2001. "Post-Unification Wage Growth in East Germany." *The Review of Economics and Statistics* 83 (February): 190–95.

James, Harold. 1996. *International Monetary Cooperation Since Bretton Woods.* Washington: International Monetary Fund.

Johnson, Peter A. 1998. *The Government of Money: Monetarism in Germany and the United States.* Ithaca: Cornell University Press.

Keynes, John M. [1923] 1971. "A Tract on Monetary Reform." In *The Collected Writings of John Maynard Keynes,* vol. 4. London: Macmillan.

Kitterer, Wolfgang. 1999. "Public Finance and the Central Bank." In *Fifty Years of the Deutsche Mark,* ed. Deutsche Bundesbank. Oxford: Oxford University Press.

Lamont, Norman. 1999–2000. "Black Wednesday—The Controversy Continues." *Central Banking* 10: 65–69.

Le Monde. 2001. "Le 10 Mai 1981, La Gauche et 'Le Monde.'" 19 May.

Marsh, David. 1992. *The Most Powerful Bank: Inside Germany's Bundesbank.* New York: Random House.

Neumann, Manfred J. M. "Monetary Stability: Threat and Proven Response." In *Fifty Years of the Deutsche Mark,* ed. Deutsche Bundesbank. Oxford: Oxford University Press.

New York Times. 1993. "Attack on Franc Threatens System." 31 July.

———. 2000. "In the Midst of Upheaval, Yet out of Sight: Horst Koehler." 15 March.

Pringle, Robert. 1996. "Black Wednesday Rides Again." *Central Banking* 6 (Spring): 121–23.

Richter, Rudolf. 1999. "German Monetary Policy as Reflected in the Academic Debate." In *Fifty Years of the Deutsche Mark,* ed. Deutsche Bundesbank. Oxford: Oxford University Press.

Schlesinger, Helmut. 1984. Luncheon address at symposium, Federal Reserve Bank of Kansas City, 1–3 August, in Jackson Hole, Wyoming. Extract from press articles, Deutsche Bundesbank, 10 August 1984.

———. 1992. Speech at the Fiftieth Anniversary of the Bank of Thailand, 11 December. Reproduced in press articles, Deutsche Bundesbank, 15 December.

Schmid, Peter. 1996. "Monetary Targeting." In *Monetary Policy in Transition in East and West: Strategies, Instruments and Transmission Mechanisms,* ed. Austrian National Bank.

Solomon, Robert. 1982. *The International Monetary System, 1945–1981.* New York: Harper & Row.

———. 1999. *Money on the Move: The Revolution in International Finance since 1980.* Princeton, N.J.: Princeton University Press.

Stanley, Stephen. 1993. "ERM Collapse." Federal Reserve Bank of Richmond Pre-FOMC Memo, 10 August.

Streit, Manfred E. 1999. "German Monetary Union." In *Fifty Years of the Deutsche Mark,* ed. Deutsche Bundesbank. Oxford: Oxford University Press.

Stern, Klaus. 1999. "The Note-Issuing Bank within the State Structure." In *Fifty Years of the Deutsche Mark,* ed. Deutsche Bundesbank. Oxford: Oxford University Press.

Vanthoor, Wim F. V. 1999. *A Chronological History of the European Union 1946-1998.* Cheltenham, UK: Edward Elgar.

Volcker, Paul A. and Toyoo Gyohten. 1992. *Changing Fortunes: The World's Money and the Threat to American Leadership.* New York: Random House.

von Hagen, Jörgen. 1999. "A New Approach to Monetary Policy (1971-8)." In *Fifty Years of the Deutsche Mark,* ed. Deutsche Bundesbank. Oxford: Oxford University Press.

———. 1999. "Money Growth Targeting by the Bundesbank." *Journal of Monetary Economics* 43: 681–701.

Wall Street Journal. 1993. "German Stance on Rates Sends ERM to Brink." 4 August.

Whitt, Joseph A. 1994. "Monetary Union in Europe." Federal Reserve Bank of Atlanta *Economic Review* 1 (January/February): 11–27.

Yeager, Leland B. 1976. *International Monetary Relations: Theory, History and Policy.* New York: Harper & Row.

Towards a Theory of Capacity Utilization: Shiftwork and the Workweek of Capital

Andreas Hornstein

Among the large number of economic indicators that provide information on the current or future state of the economy, the index of capacity utilization (CU) published by the Federal Reserve Board is one of the more prominent. Low levels of the CU index number tend to be associated with below-average aggregate activity, and high levels are supposed to indicate inflationary pressures (Corrado and Mattey 1997).[1] For the most part, CU is an empirical concept that is only loosely related to economic theory, and until recently it has not played an important role in models of the business cycle. There are various interpretations of what CU means, but in this article I address one particular aspect of CU from the point of view of standard production theory. I first review some empirical evidence on the workweek of capital, a measure that makes the concept of CU operational. I then extend standard production theory to incorporate the workweek of capital into the neoclassical growth model. Finally, I argue that recent attempts to use variations in the workweek of labor in order to get CU-adjusted measures of short-term productivity growth are potentially misguided, since the workweek of labor is not an unbiased measure of the workweek of capital.

The Board's CU index is defined as the ratio of actual output to potential output. Potential output reflects "sustainable practical capacity, defined as the

The author would like to thank Mike Dotsey, Margarida Duarte, Tom Humphrey, and Yash Mehra for helpful comments. The views expressed in this paper are those of the author and do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Finn (1995) argues that the CU index is not particularly useful in forecasting future inflation rates.

greatest level of output each plant in a given industry can maintain within the framework of a realistic work schedule, taking account of normal downtime and assuming sufficient availability of inputs to operate machinery and equipment in place" (Corrado and Mattey 1997, 152). The capacity measures are limited to the manufacturing industries, mining, and electric and gas utilities, and they are based on the *Survey on Plant Capacity,* which is produced by the U.S. Census Bureau in the fourth quarter of each year. Capacity measures for quarters are obtained by smooth interpolation of the fourth-quarter numbers. Given the smooth interpolation of capacity and the volatility of output, movements in the CU index are mainly due to movements in actual output.

The standard theory of production views the quantity of output produced as a function of the quantities of inputs used. How does the concept of CU fit into this theory? I focus on the "utilization" aspect of CU and disregard the "capacity" aspect.[2] When we measure inputs to production we usually consider the capital stock, total hours worked by production workers/employees, and the quantities of intermediate inputs (materials and energy) used. We implicitly assume that the input services provided by capital are proportional to its accumulated stock. Looking at the production of a plant, we can see that its output obviously depends on the extent to which it uses its existing capital stock: how many machines are running and for how long? That is, the plant can vary the service flow per unit of capital, and this input variation is not covered in the usual input measures. This opportunity to vary the flow of services from capital creates a problem for productivity measurement when we want to attribute movements in output to movements in inputs and changes in productivity.

The workweek of capital is supposed to capture the service flow of the capital stock, which is proportional to the average duration of time for which a unit of capital is operated. The workweek of capital is different from the workweek of labor, which is the average duration of time a unit of labor (worker) is employed. To the extent that labor and capital are complementary inputs (for example, a certain number of workers are needed to operate a machine), the workweek of capital and the workweek of labor are related, but they need not be the same. For instance, if a plant is operating multiple shifts, then the workweek of capital will be a multiple of the workweek of labor. Furthermore, if the extent to which a plant uses shiftwork changes over the cycle, the cyclical behavior of the workweeks of capital and labor will be different.

[2] The capacity concept implies that there is a maximal output level in the short run, the "capacity constraint," and that this level cannot be exceeded no matter how many variable inputs are hired. Economic variables will respond differently to changes in the environment, depending on whether the capacity constraint is binding or not. This behavior can introduce a nonlinearity into observed economic relations. For a recent model with occasionally binding capacity constraints, see Hansen and Prescott (2000).

In the remainder of the article I first review evidence on the workweek of capital from micro- and macrostudies. I then describe a simple model with variable employment in late shifts. Finally, I discuss the implications of variations in shiftwork for the measurement of productivity changes.

1. OBSERVATIONS ON THE WORKWEEK OF CAPITAL

The workweek of capital varies widely across industries and across plants. The average length of the workweek of capital in an industry depends on common elements of the production processes used by different plants in an industry. Within an industry, plants deviate from the industry average in response to variations in demand across plants because production cannot be reallocated between plants. Depending on the structure of the production process, there are limits on the extent to which firms can vary the workweek of capital. Microevidence indicates that plants can adjust the workweek of capital along a number of margins. Aggregating plant level data to get industry data shows that the workweek of capital varies substantially over time and indeed is more volatile than the workweek of labor. It also appears that at the aggregate level a substantial fraction of the capital workweek's volatility is due to movements in the share of the labor force that works on late shifts.

How long do plants operate in a quarter and how do they change the duration for which they operate? Mattey and Strongin (1997) answer this question based on plant level data on actual and capacity hours worked per quarter from the *Survey on Plant Capacity*. They break down total operating hours in a quarter as follows:

$$\frac{Weeks}{Quarter} \cdot \frac{Days}{Week} \cdot \frac{Shifts}{Day} \cdot \frac{Hours}{Shift}.$$

Note that as long as plants do not operate 24 hours a day for every day of the quarter, there is scope for variation in the workweek of capital. Mattey and Strongin (1997) find that in their sample 35 percent of all plants do not operate every week, 62 percent do not operate every day of the week, 20 percent have only one shift, and 13 percent do not work overtime. Plants design their operating margins in order to vary the workweek of capital. For example, in some industries almost all plants seem to operate essentially every hour of the quarter. Within other industries the workweek of capital varies substantially across plants. Mattey and Strongin (1997) classify industries according to operating margins, and they distinguish between "continuous process" industries and "variable workweek" industries. The two classifications do not exhaust their sample.

An industry is classified as continuous process if its plants report that at capacity they essentially operate every hour of the quarter. Continuous process industries contain one-fourth of all plants in the sample. Even within this class, not all plants actually operate at full capacity: 20 percent shut down for five

weeks in a quarter, 11 percent work only five days a week, and 9 percent work only two shifts. For continuous process plants, output variations take place mainly through the variation in material inputs use and not through variations in the workweek of capital. Petroleum is an example of a continuous process industry.

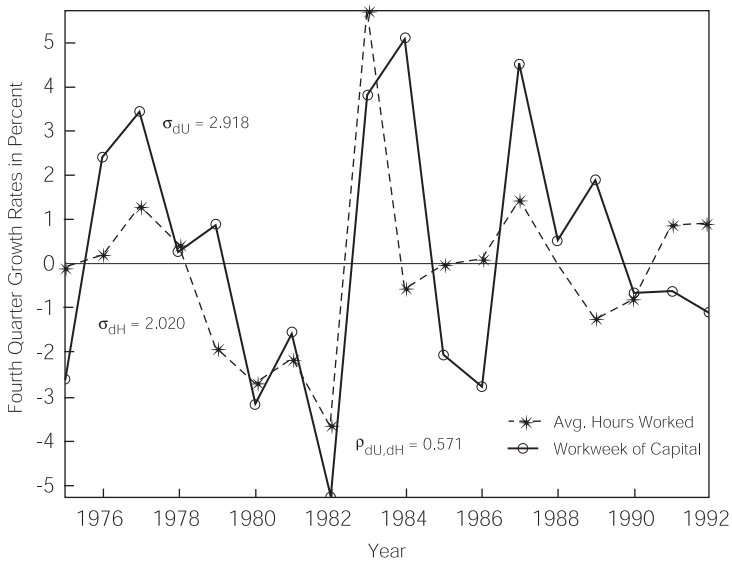
An industry is classified as variable workweek if individual plants in that industry show substantial variation in their actual workweek over time. Most plants in these industries (85 percent) operate at least 12 weeks per quarter and five to six days a week (93 percent), so that differences in the workweek of capital across plants are evident mainly in the number of shifts operated: 27 percent operate one shift, 40 percent operate two shifts, and 32 percent operate three shifts. A substantial share (16 percent) of these plants also use overtime as an option.

Obviously the boundaries between continuous process and variable workweek industries are not clear cut. For example, the steel industry appears to be close to the continuous process ideal: an individual blast furnace is a continuous process technology. Yet Bertin, Bresnahan, and Raff (1996) argue that for iron production, the relevant unit of observation is a plant that may operate several blast furnaces, and therefore the average time its furnaces are operated (workweek) is the more relevant measure. Another example is an apparel workshop that can vary the average time its sewing machines are used along two margins: (1) the length of time of time an individual machine is used, and (2) the number of machines actually operated.

The prototypical example of a variable workweek industry is automobile production, particularly its assembly plants. Bresnahan and Ramey (1994) study 50 U.S. automobile plants from 1972–1983. They find that variations in the workweek of capital are mainly due to weekly shutdowns related to model changeovers, inventory adjustment, and holidays. Individual plants only infrequently change the number of shifts they operate. Nevertheless, since shift changes have a huge impact on output, variations in the number of shifts operated make a substantial contribution to output volatility (about 25 percent). Hall (2000) in a study of 14 Chrysler plants from 1990–1994 also finds that most of the variation in an individual plant's workweek is due to short-term shutdowns rather than to variations in the number of shifts operated. Even though at the individual plant level the shift margin does not appear to be important, it can still be important at the industry level if the relative shares of plants operating at different shifts change systematically over the cycle.

What are the implications of plant-level patterns for industrywide movements in the workweek of capital? Beaulieu and Mattey (1998) construct capital workweek series for two-digit Standard Industrial Classification (SIC) industries in the manufacturing sector for the period 1974–1992. They find that the capital workweek is both longer and more volatile than the workweek of labor. For the overall manufacturing sector, the mean workweek of capital

Figure 1 Workweek of Capital and Workweek of Labor in Manufacturing



Notes: The average workweek of capital for manufacturing is from Beaulieu and Mattey (1998). The average hours worked are total hours worked divided by the number of employees for manufacturing only. For both series, the growth rate is calculated from the fourth quarter in the previous year to the fourth quarter of the current year.

is 97 hours. In Figure 1, I plot the fourth-quarter to fourth-quarter percentage growth rates of the workweek of capital and average hours worked in manufacturing.[3] We can see that the workweek of capital is substantially more volatile than average hours worked, and although the two series tend to move together, the fit is not very tight: the correlation coefficient is about 0.6. Beaulieu and Mattey (1998) also find significant differences in the statistical properties of the workweek of capital across industries. The mean workweek of capital ranges from 44 hours in apparel to as high as 156 hours for the continuous-process-type petroleum refining industry. The volatility of the workweek of capital is also quite different across industries; for example, the standard deviation of percentage changes in the workweek of capital ranges from a high of 10.0 in primary metals to a low of 3.0 for chemicals and petroleum.

[3] The SPC is undertaken in the fourth quarter of each year, that is, the workweek of capital refers only to that quarter. For consistency I have used the same procedure for average hours worked.

Unfortunately, the work by Beaulieu and Mattey (1998) is limited to the manufacturing sector. Shapiro (1996) constructs an alternative measure of the workweek of capital based on the employment pattern of shiftwork in the *Current Population Survey*, and his work covers manufacturing and nonmanufacturing industries of the economy. With respect to the manufacturing sector, Shapiro (1996) suggests that the capital workweek is shorter (52 hours) and only half as volatile as Beaulieu and Mattey (1998) argue. Like Beaulieu and Mattey (1998), Shapiro (1996) also observes substantial variation of the mean and volatility of the capital workweek across industries in manufacturing. With respect to nonmanufacturing industries, Shapiro (1996) finds that the capital workweek is only 44 hours, which is close to the workweek of labor, and that the capital workweek is substantially less volatile than in manufacturing.

Why does the capital workweek tend to be more volatile than the workweek of labor, especially in the manufacturing sector? Shapiro (1996) points to variations in the extent to which shiftwork is used in production. He finds that in overall manufacturing about 25 percent of all production workers are working late shifts.[4] Furthermore, in each industry the late-shift share of employment is quite volatile and tends to increase with overall employment. In particular, Shapiro estimates that a 1 percent increase of employment increases late-shift employment by 1.5 percent. For nonmanufacturing industries, where the capital workweek is less volatile, late-shift work is not as prevalent, and for most service industries the late-shift employment share tends to decline with overall employment.

2. A MODEL OF SHIFTWORK AND THE WORKWEEK OF CAPITAL

I now construct a simple model where capacity utilization is reflected in the workweek of capital and the workweek of capital is closely related to the share of late-shift work in total employment. This model builds on work by Kydland and Prescott (1991), Bils and Cho (1994), and Hall (1996). I argue that there are systematic differences between the workweek of capital and the workweek of labor. The model serves as an illustration only, and I do not provide a complete analysis of all of its properties in this article.

Shiftwork in Production

In the standard model of production, we view output y as a function of the capital stock k and total hours worked nh:

$$y = zk^{\alpha}(hn)^{1-\alpha}, 0 < \alpha < 1, \tag{1}$$

[4] The employment share of late-shift work ranges from 4 percent in apparel to 40 percent in tobacco.

where n is employment and h is average hours worked per worker (that is, the workweek of labor), and z denotes productivity.[5] We usually assume that output increases as inputs increase, the marginal product of each input is positive, and the marginal product of each input is declining. Furthermore, production is constant-returns-to-scale (CRS) in the capital stock and total hours worked: if we double the capital stock and total hours worked, then output doubles. This structure assumes that the contribution of the input capital is proportional to the stock of capital.

We can allow for variations in the utilization of capital through changes in the workweek of capital, assuming that labor works a single shift:

$$y = zk^{\alpha}n^{1-\alpha}h = z\,(hk)^{\alpha}\,(hn)^{1-\alpha}\,. \tag{2}$$

We now assume that production per unit of time is CRS in the capital stock and workers employed. Total output is then proportional to the hours capital and workers are employed. The relevant inputs for this production structure are the services capital and labor provide, and these services are proportional to the hours worked. Furthermore, the production structure continues to be CRS with respect to the capital and labor services employed. This production structure has been used by Kydland and Prescott (1991), Bils and Cho (1994), and Hall (1996). Note that for this production structure, the workweeks of capital and labor are the same.

What happens if we allow for more than one shift and if we can vary the relative employment levels on the two shifts? I consider the case where the economy can operate the capital stock with two employment shifts. That is, in any given period the existing capital stock can be used twice in production. For this case I assume that production takes place with machines and that machines and workers are complementary, meaning that if a worker is matched with a machine containing $\tilde{k}$ units of capital, then output per worker per unit of time is $z\tilde{k}^{\alpha}$. Assuming that all machines are the same, the number of machines m is limited by the available total capital stock, $m\tilde{k} \leq k$. Given the available machines, the economy can employ $n_1 \leq m$ workers, each working for h_1 hours on the first shift, and output from the first shift is $\tilde{k}^{\alpha}h_1n_1$. The same machines can be used on the second shift with employment $n_2 \leq m$ and shift length h_2, with corresponding output $\tilde{k}^{\alpha}h_2n_2$. The sum of shift lengths is bounded by the duration of a period, $h_1 + h_2 \leq \bar{h}$. Total production in a period is then

$$y = z\tilde{k}^{\alpha}\,(n_1h_1 + n_2h_2)\,.$$

For an efficient allocation of capital, all capital is used in machines and for at least one shift all machines are employed. An efficient allocation means that

[5] For concreteness I have assumed that the production function is Cobb-Douglas. All the arguments apply for a general concave constant-returns-to-scale production function.

for a given employment decision, the vector $(\tilde{k}, m)$ maximizes output. Since the marginal product of capital is positive, and there is no additional cost of building a machine besides the amount of capital it contains, it is always efficient to use all available capital, that is, $\tilde{k}m = k$. For the same reason, the efficient number of machines is equal to the maximum of the two employment levels. Without loss of generality let $n_1 \geq n_2$, in which case we need at least $m = n_1$ machines. It is not efficient to create more machines, because that would only reduce the capital-labor ratio and therefore reduce output for a given employment decision. This argument simplifies the representation of the production function to

$$y = z\,(k/n_1)^{\alpha}\,(n_1 h_1 + n_2 h_2)\,, \text{ with } h_1 + h_2 \leq \bar{h} \tag{3}$$

$$= z\,(uk)^{\alpha}\,(hn)^{1-\alpha}\,. \tag{4}$$

In this economy, the workweek of capital (that is, the average duration a unit of capital is operated) is $u = (n_1 h_1 + n_2 h_2)\,/n_1$. The workweek of labor (that is, the average duration a worker is employed) is $h = (n_1 h_1 + n_2 h_2)/n$, where n is total employment $n = n_1 + n_2$. Variations in the employment share of the first shift $\omega = n_1/n$ drive a wedge between the workweek of capital and the workweek of labor:

$$u = h/\omega. \tag{5}$$

I make one more assumption regarding the dynamic structure of production. Specifically, I assume that it is costly to change the capital-labor ratio in machines. For simplicity, I choose an extreme case where the capital-labor ratio is fixed at the beginning of the period. This means that the capital-labor ratio cannot be adjusted in response to new information on the state of the economy. The capital-labor ratio, however, can be adjusted at no cost at the end of a period. This assumption essentially makes employment on the first shift a predetermined variable. On the other hand, variations of employment on the second shift allow the economy to respond more flexibly to contemporaneous shocks. This feature of the economy will give rise to variations in the employment ratio of the second shift, and it will increase the volatility of the workweek of capital relative to the workweek of labor.

The remainder of the production structure is standard:

$$\begin{aligned} y &= c + x + g \\ k' &= (1-\delta)\,k + x \\ \ln\left(g'/\bar{g}\right) &= \rho_g \ln\left(g/\bar{g}\right) + \varepsilon_g \\ \ln z' &= \rho_z \ln z + \varepsilon_z. \end{aligned}$$

Output can be used for private consumption c, government spending g, and investment x. Investment augments the capital stock, which depreciates at a constant rate δ. Primes denote the next period's values. Government spending is financed by lump-sum taxation, and log-deviations from its mean $\bar{g}$ follow an

AR(1) process with autocorrelation coefficient ρ_g. Productivity also follows an AR(1) process with autocorrelation coefficient ρ_z. The disturbance terms ε_g and ε_z in the government spending and productivity equations are i.i.d. and uncorrelated with each other.[6]

Capital is not always utilized to the fullest extent. In our setup full capital utilization means that both shifts use all available machines, $n_1 = n_2$, and machines are continuously operated through the period, $h_1 + h_2 = \bar{h}$. Capital will not be fully utilized if there are increasing marginal costs to the utilization of capital. One way to model these costs of capital utilization is to assume that the rate at which capital depreciates increases as capital utilization increases (see Greenwood, Hercowitz, and Huffman [1988], Burnside and Eichenbaum [1996], and Basu and Kimball [1997]). In my setup there is another reason why capital would not always be operated at full capacity. Since capital utilization is tied to the use of labor, higher capital utilization requires more employment at less desirable times and longer work hours. If there is a wage premium for work at extended hours and that wage premium is sufficiently high, then capital may never be used at full capacity. I now describe preferences that give rise to a wage premium for shiftwork and overtime work.

Preferences and the Wage Premium

There is an infinitely-lived representative household with a large number of members. In any time period the household can send n_1 of its members to the first shift, where they will work h_1 hours, and it can send n_2 of its members to the second shift, where they will work h_2 hours. The number of employed household members cannot exceed the total number of household members, $n_1 + n_2 \leq \bar{n}$. The household's expected utility from a random consumption and labor supply process is

$$E_0 \sum_{t=0}^{\infty} \beta^t \left\{ \log c_t - \sum_{i=1,2} \sigma_i \left[\frac{n_{it}^{1+\gamma}}{1+\gamma} + \psi_i n_{it} \frac{h_{it}^{1+\phi}}{1+\phi} \right] \right\}, \tag{6}$$

with discount rate $\beta \in (0, 1)$, and $\sigma_i, \psi_i, \gamma, \phi \geq 0$.[7]

The household is assumed to maximize the expected present value of utility subject to budget constraints. The household purchases consumption goods, saves, and supplies its labor. The market wage rates for employment on the two shifts are given by the wage functions $w_{1t}(h_{1t})$ and $w_{2t}(h_{2t})$. The

[6] My treatment of government spending follows Hall (1996). In the last section I will discuss some issues in the measurement of productivity within the context of the capacity utilization model. The methods I discuss there use Instrumental Variable (IV) techniques, that is, they require the use of a variable which is exogenous to productivity but affects production decisions in the economy. In my model economy, government spending is such an instrumental variable.

[7] These preferences are based on those described by Bils and Cho (1994).

household's period budget constraint is then

$$c_t + a_{t+1} \leq R_t a_t + w_{1t}\left(h_{1t}\right) n_{1t} + w_{2t}\left(h_{2t}\right) n_{2t},$$

where R_t is the return on asset holdings a_t. Note that the household can choose only which shifts to work, but not how long to work on each shift. On the other hand, when firms make their employment decisions, I assume that they choose employment in each shift and the length of each shift, given the wage functions they see in the labor market. The cost minimization problem of a firm is

$$\begin{aligned} &\min_{h_i, n_i} w_1\left(h_1\right) n_1 + w_2\left(h_2\right) n_2 \\ &\text{s.t.} \quad y = z\left(k/n_1\right)^{\alpha}\left(n_1 h_1 + n_2 h_2\right), \\ &\qquad 0 \leq n_2 \leq n_1. \end{aligned}$$

We can define a competitive equilibrium for this economy, and it turns out to be the solution to the planning problem where we choose an allocation that maximizes the household's utility subject to the constraint that the allocation is feasible.[8]

The optimal employment decision by a household implies that the marginal disutility of employment at a given shift length is equal to the wage for a shift of that length:

$$w_{it}\left(h_i\right) = \frac{\sigma_i}{\lambda_t}\left\{n_{it}^{\gamma} + \psi_i h_{it}^{1+\phi}/\left(1+\phi\right)\right\}, \tag{7}$$

where λ_t is the Lagrange multiplier on the period budget constraint. In an equilibrium we can take this condition as the definition of the wage function. We can see that for the same employment levels and shift lengths, work on the second shift will require a higher wage than work on the first shift if $\sigma_1 \leq \sigma_2$ or $\psi_1 \leq \psi_2$, which means that work on the first shift creates less disutility than work on the second shift.

A Quantitative Evaluation of the Model

What does this model say about the behavior of the workweeks of capital and labor? In particular, does it predict that the workweek of capital is more volatile than the workweek of labor and that the late-shift employment share is procyclical? The model is sufficiently complicated that analytical characterizations are not feasible. I therefore parameterize the model, obtain a numerical solution, and calculate the response of the workweeks of capital and labor to a productivity shock.

[8] In a more general formulation we have shifts of different lengths with wages depending on the length of the shift. Nevertheless, in an equilibrium, the household and firms would choose to operate each shift at one particular length. See Hornstein and Prescott (1993).

The main difference between the model described above and the standard growth model relates to the description of employment, in particular how labor market variables enter preferences and production. With respect to preferences, hours worked and employment are separate arguments in the utility function; furthermore, there are two types of employment (shifts). For the specification of the employment and hours elasticities, I follow Bils and Cho (1994) and select $\phi = 2$ and $\gamma = 1.6$.[9] I calibrate the scale parameters on the disutility of work based on assumptions on the relative steady state values of employment and hours worked for the two shifts.

Total employment is normalized at one, and I assume that 20 percent of total employment is in the second shift, $n_1 = 0.8$ and $n_2 = 0.2$. As stated above, Shapiro (1996) reports a mean of 25 percent for late-shift employment in manufacturing, but manufacturing represents only a subset of the economy, and late-shift work is less prevalent outside manufacturing. Since I do not have any information on the relative length of shifts, I simply assume that in the steady state both shifts are of equal length, which I normalize to one, $h_1 = h_2 = 1$. The two assumptions on relative employment and shift length imply that the workweek of capital is 25 percent longer than the workweek of labor. The calibrated capital workweek is substantially shorter than what Shapiro (1996) reports for the capital workweek in manufacturing based on *Survey on Plant Capacity* data, but it is comparable to his capital workweek estimates based on the *Consumer Population Survey*. The assumptions on steady state employment and hours worked and the assumptions on the elasticity parameters together determine the scale coefficients σ_i and ψ_i in the utility function.

We can evaluate the parameterization of labor supply based on the implied shift premium and labor supply elasticities. First, the implied steady state shift premium of the second shift is quite high, about 70 percent. This premium is substantially higher than the 10 percent shift premium Bils (1995) argues for or the 20 percent night-shift premium Shapiro (1996) suggests. Second, from the equilibrium wage function the implied elasticity of shift employment and hours worked to changes in the wage rate are[10]

$$1/\eta_n \equiv \frac{\partial w(h)}{\partial n}\frac{n}{w(h)} = \gamma \frac{n^\gamma}{n^\gamma + \psi h^{1+\phi}/(1+\phi)},$$
$$1/\eta_h \equiv \frac{\partial w(h)}{\partial h}\frac{h}{n} = (1+\phi)\frac{\psi h^{1+\phi}/(1+\phi)}{n^\gamma + \psi h^{1+\phi}/(1+\phi)}.$$

[9] I should note that the argument which Bils and Cho (1994) make for these particular parameter values is not strictly applicable to my model since their interpretation of the employment and hours worked variables is different from mine.

[10] These are wealth compensated supply elasticities since the Lagrange multiplier (marginal utility of consumption) is taken as constant.

For the given parameterization of preferences these labor supply elasticities are[11]

$$\eta_{n1} = 1.46;\ \eta_{n2} = 0.94;\ \eta_{h1} = 0.58;\ \text{and}\ \eta_{h2} = 1.00.$$

The labor supply elasticities are relatively low compared to other specifications used in dynamic general equilibrium models, where supply elasticities around 2 are more common. The labor supply elasticities in this model and in a standard growth model are, however, not directly comparable for two reasons. First, standard dynamic general equilibrium models usually do not distinguish between the labor supply elasticity of employment and the labor supply elasticity of hours worked per worker as I have done. Second, supply elasticities are usually stated in terms of percentage response of hours to a percentage change in the hourly wage rate.[12]

The remaining parameter values are comparable to those used in other studies. I choose the time discount factor $\beta = 0.99$ such that the annual steady state interest rate is 4 percent, the depreciation $\delta = 0.02$ such that the annual depreciation rate is 8 percent, and the capital coefficient $\alpha = 1/3$ such that the capital income share is one-third. The parameterization of productivity and government spending is based on Hall (1996). For the productivity process I assume that $\rho_z = 0.97$ and the standard deviation of innovations to the productivity process is 0.006. For the government spending process I assume that $\rho_g = 0.97$ and the standard deviation of innovations to the productivity process is 0.009. The steady state share of government spending in output is $\bar{g}/y = 0.18$.

Figure 2 displays the responses of the workweeks of capital and labor to a percentage point deviation of productivity from the steady state value. We can see that on impact, the workweek of capital increases more than does the workweek of labor. Since employment on the first shift is predetermined, the economy cannot use this margin when it responds to the contemporaneous productivity shock. The economy is, however, free to increase employment

[11] Some algebra shows that the steady state values of these labor supply elasticities are

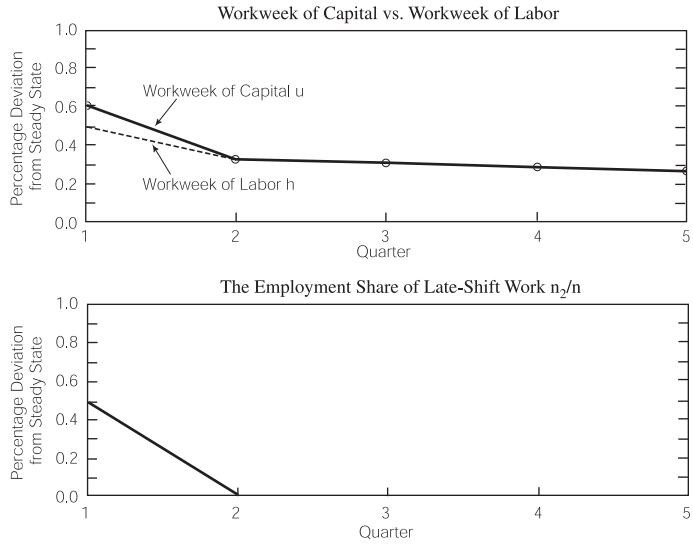
$$1/\eta_{n_1} = \gamma\left[\rho/(1-\rho) - \alpha u/h_1\right]/\left[1 - \alpha u/h_1\right],$$
$$1/\eta_{n_2} = \gamma\rho/(1-\rho),$$
$$1/\eta_{h_1} = 1/(1 - \alpha u/h_1),$$
$$1/\eta_{h_2} = 1.$$

[12] A more appropriate procedure would define the labor supply elasticity with respect to average wage changes. This does not affect the measure of employment supply elasticity, but it does change the measure of hours elasticity:

$$1/\hat{\eta}_h \equiv \frac{\partial\left[w(h)/h\right]}{\partial h}\frac{h}{w(h)/h} = 1/\eta_h - 1.$$

According to this alternative measure, the hours supply elasticity is actually quite high.

Figure 2 Workweek of Capital and Workweek of Labor in Model



on the second shift, which increases the employment share of late-shift work. This increase in the employment share of the second shift in turn amplifies the response of the workweek of capital.

The model does not generate persistent differences between the workweek of capital and the workweek of labor. That is to say, only unanticipated shocks create a divergence between the two workweek definitions. After the first period, when employment on the first shift can be adjusted again, the economy attains its target late-shift employment ratio. This feature of the model is due to the particular assumption on adjustment costs for the capital-labor ratio, namely infinite adjustment costs at the beginning of the period and zero adjustment costs at the end of the period. Alternatively, we could assume that any changes in the production structure, specifically the capital content of machines, involve some resource costs (see, for example, Bils and Cho [1994]). With this alternative assumption, there will be persistent deviations between the workweek of capital and the workweek of labor in response to a productivity shock.

Another way to evaluate the role of shiftwork in the model is to compare its business cycle properties to those of the U.S. economy and other models without shiftwork. The business cycle properties are defined with respect to Hodrick-Prescott filtered time series of output, consumption, investment, capital stock, total hours worked, employment, and the workweek of labor (see columns 1 and 5 of Table 1). For our purposes it is of interest to note

Table 1 Business Cycle Properties of U.S. Data and Model Economies

	Percentage Standard Deviations				Correlation with Output			
	U.S.	**I**	**II**	**III**	**U.S.**	**I**	**II**	**III**
Output	1.74	1.08	1.00	1.02	1.00	1.00	1.00	1.00
Consumption	1.29	0.50	0.48	0.48	0.85	0.89	0.89	0.89
Investment	8.45	3.58	3.30	3.37	0.91	0.99	0.99	0.99
Capital Stock	0.63	0.25	0.24	0.24	0.05	0.03	0.02	0.02
Total Hours Worked	1.74	0.40	0.34	0.35	0.77	0.94	0.90	0.92
Employment	1.50	0.26	0.23	0.22	0.81	0.94	0.74	0.83
Workweek of Labor	0.46	0.14	0.15	0.15	0.76	0.94	0.88	0.90
Workweek of Capital		0.14	0.15	0.19		0.94	0.88	0.83

Notes: U.S. Data are from Bils and Cho (1994) and they cover the time period 1955:III–1984:I. Model I is the one-shift model with employment not predetermined; Model II is the one-shift model with employment predetermined; Model III is the two-shift model with employment of the first shift predetermined. All time series are detrended with the Hodrick-Prescott filter. The model statistics are based on 100 simulations where each simulation consists of 30 years of quarterly observations.

that employment is more volatile than the workweek of labor (average hours worked), and employment is slightly more correlated with output than is the workweek of labor. I do not have quarterly observations on the workweek of capital, but in the previous section I note that for annual growth rates, the workweek of capital is more volatile than the workweek of labor and that the workweeks of capital and labor are only weakly correlated.

To evaluate the contribution of shiftwork as modeled in this article, I consider three models. The first and second models assume that there is a distinction between employment and the workweek of labor, but that all employment is in one shift only. Preferences and production are as previously described, with the restriction that $n_2 = h_2 = 0$. For Model I, employment and the workweek of labor are determined at the beginning of the period, after the productivity and government spending shock have been observed. For Model II, employment is determined before observations on the current productivity and government spending shock are available. Model I is similar to Kydland and Prescott (1991) and Bils and Cho (1994), whereas Model II is similar to Burnside, Eichenbaum, and Rebelo (1993) and Hall (1996) in that part of the employment decision is predetermined. Finally, Model III is the economy with shiftwork as described above.

Table 2 The Workweeks of Capital and Labor, Annual Growth Rates

	Std. Dev. Δu	**Std. Dev.** Δh	**Corr.** $(\Delta u, \Delta h)$
U.S.	2.92	2.02	0.57
Model III	1.14	0.99	0.98

Notes: For U.S. data, see Figure 1.

For each model I generate 100 random samples, each with 30 years of quarterly observations. The artificial time series are detrended with the Hodrick-Prescott filter, and I calculate the average standard deviations and correlations with output of the detrended series. The results are listed in Table 1. We can see that given the exogenous disturbances, the model economies are not as volatile as the U.S. economy. The model economies capture the fact that investment is more volatile than consumption, but consumption tends to be too smooth. The model economies also capture the fact that employment is more volatile than the workweek of labor, but overall labor is too smooth relative to the U.S. economy. Concerning the comovement with output, we see that in the models employment is more closely correlated with output than it is in the U.S. economy.

In Model III, the workweek of capital is indeed more volatile than the workweek of labor for detrended quarterly data. Since quarterly U.S. data on the workweek of capital are not available, I have calculated the standard deviations and correlations for the annual percentage growth rates of the workweeks of capital and labor (see Table 2). We can see that in the U.S. manufacturing sector, the workweek of capital is relatively more volatile than the workweek of labor. Furthermore, the relationship between the workweek of capital and the workweek of labor is much tighter in the model than in the data. While the model captures the qualitative features of the workweek of capital, it does not come close yet to replicating its quantitative properties.

3. IMPLICATIONS FOR PRODUCTIVITY MEASUREMENT

I now study the implications of variable shiftwork for the measurement of productivity changes. Since productivity change is defined as output changes that cannot be attributed to input changes, unobserved input movements obscure our measures of productivity change. Changes in capital services, such as changes in the workweek of capital, represent important input movements that are not reflected in our standard measures of inputs. Recently, Basu and Kimball (1997) have argued that unobserved variation in the utilization of

inputs is related to the observed variation in the workweek of labor.[13] They use this relationship to obtain a utilization-corrected measure of productivity change. There are two potential problems with the approach of Basu and Kimball (1997). First, their procedure requires that the workweek of capital be strictly proportional to the workweek of labor. If this is not true, as suggested by the available evidence and theory on the workweek of capital, then their procedure does not necessarily generate unbiased estimates of the volatility of productivity change. Second, even if the estimates of the volatility of productivity are unbiased, they may not be precise because the estimates rely on instrumental variables that may be quite poor.

Consider the standard production function (1), which takes as inputs the capital stock and total hours worked. We can measure productivity growth through the Solow residual, which defines productivity growth as output growth less input growth weighted by the output elasticities of inputs:

$$\hat{z}^m \equiv \hat{y} - \alpha\hat{k} - (1-\alpha)\left(\hat{h} + \hat{n}\right) = z.$$

The Solow residual is an operational concept because in a competitive equilibrium with CRS production, we can identify the elasticities with the factor income shares of inputs. Consider now the production function (2) with a variable workweek of capital but only one shift. If we continue to assume that the relevant inputs are the stock of capital and total hours worked, then measured productivity growth no longer reflects true productivity growth:

$$\hat{z}^m = \hat{z} + \alpha\hat{h}.$$

However, a simple correction of the Solow residual, made by subtracting the growth rate of average hours worked weighted by the factor income share of capital, retrieves the true productivity change:

$$\hat{z}^m - \alpha\hat{h} = \hat{z}.$$

Empirically, this simple correction does not deliver measures of true productivity change. Suppose you have variables—call them instrumental variables—that on a priori grounds are considered to be independent of true productivity change. On these grounds, the instrumental variables should be uncorrelated with the Solow residual corrected for average hours worked. In empirical applications, however, it turns out that the Solow residual corrected for average hours worked remains correlated with these instrumental variables. However, in an instrumental variables regression of the measured Solow residual on average hours growth, Basu and Kimball (1997) find that the coefficient on average hours growth is around one, larger than the factor income share of capital, which is substantially less than one. They argue that the relatively large coefficient on average hours worked reflects other unobserved input utilization, which is strictly proportional to average hours worked. Furthermore,

[13] See also Basu, Fernald, and Kimball (2000).

they argue that once they correct the Solow residual for movements related to movements in average hours worked, they can recover exogenous movements in productivity. By contrast, I argue here that this contention is not true when the workweek of capital and the workweek of labor are not perfectly correlated.

Consider the production function (3) of the two-shift model described above. With this production structure, measured productivity growth based on changes in the capital stock and total hours worked is

$$\hat{z}^m = \hat{z} + \alpha\hat{u}. \tag{8}$$

Again, the true productivity disturbance can be recovered by correcting for the workweek of capital. Shapiro (1996) argues that industry Solow residuals that are corrected for the workweek of capital are essentially uncorrelated with instrumental variables. A problem with this approach is that only a limited number of observations on the workweek of capital are available. Shapiro (1996) uses the workweek of capital numbers constructed by Beaulieu and Mattey (1998), and this sample is limited to the years 1974–1992. Given the limited availability of direct observations on the workweek of capital, the argument of Basu and Kimball (1997) for the use of average hours worked as a proxy for different forms of capacity variation is attractive. Especially so since, as they argue, average hours worked not only covers variations in the workweek of capital, but also variations in capital utilization that are not related to corresponding changes in the worktime of labor.

Basu and Kimball (1997) suggest estimating the regression equation

$$\hat{z}^m = b\hat{h} + e \tag{9}$$

using instrumental variable techniques.[14] Let q denote an instrumental variable that is uncorrelated with the true productivity shock, then the two-stage instrumental variable estimator of b is

$$\bar{b} = \frac{E\left[\hat{z}^m\hat{q}\right]}{E\left[\hat{h}\hat{q}\right]} = \frac{E\left[\hat{q}\left(\alpha\hat{u} + \hat{z}\right)\right]}{E\left[\hat{q}\hat{h}\right]}$$

given the true relationship (8). If we suppose for the moment that changes in the workweek of capital are proportional to changes in the workweek of labor, $\hat{u} = \mu\hat{h}$, then the estimator simplifies to

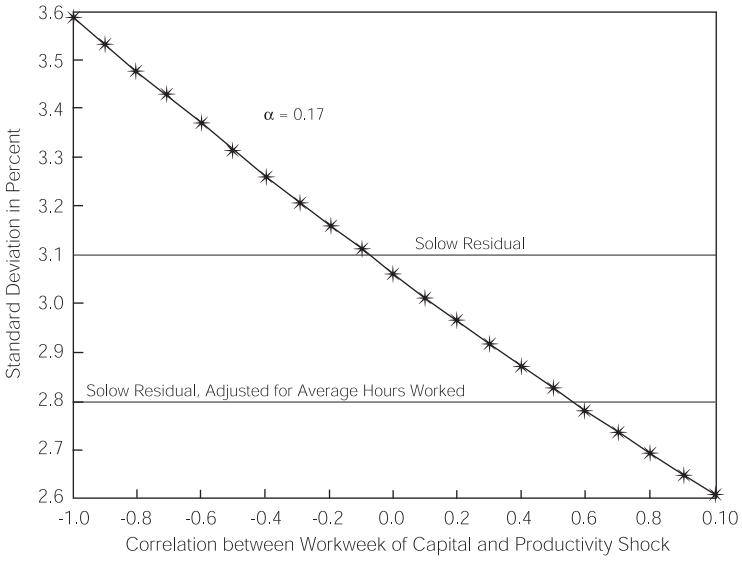
$$\bar{b} = \alpha\mu.$$

[14] Basu and Kimball's (1997) approach is actually somewhat more complicated in that they consider the possibility of noncompetitive behavior. They estimate the equation

$$\hat{y} = \gamma\hat{m} + b\hat{h} + e,$$

where γ is the average markup of price over marginal cost. I assume competitive behavior, that is, $\gamma \equiv 1$, and only the coefficient b must be estimated.

Figure 3 Productivity Volatility Implied by Workweek Volatility



Notice that the estimator does not have a structural interpretation, since μ reflects a relation between two endogenous variables that depends on elements of a fully specified equilibrium model. Nevertheless, correcting the Solow residual recovers the true changes in productivity

$$\hat{z}^m - \bar{b}\hat{h} = \left(\alpha\mu\hat{h} + \hat{z}\right) - \bar{b}\hat{h} = \hat{z}.$$

The only problem with this approach is that the above model of the workweek of capital predicts that for a reasonable specification of the production structure, the changes in the workweek of capital are not strictly proportional to changes in the workweek of labor. Furthermore, in my review of the empirical evidence on the workweek of capital, I have shown that the relation between the workweek of capital and the workweek of labor is not very tight; the correlation coefficient between their respective percentage changes is only 0.6.

If the workweek of capital is not tightly related to the workweek of labor, then the average hours-corrected estimates of the volatility of productivity disturbances obtained by Basu and Kimball (1997) are not unbiased. Suppose that the relation between the workweek of capital and the workweek of labor is $\hat{u} = \mu\hat{h} + \hat{v}$, where $\hat{v}$ is an endogenous movement in the workweek of capital that is orthogonal to the workweek of labor. Since $\hat{v}$ is endogenous, we would expect it to be correlated with the instrumental variable $\hat{q}$, $E\left[\hat{v}\hat{q}\right] \neq 0$. The

estimated parameter and the hours-corrected Solow residual are then

$$\bar{b} = \alpha\mu + \frac{E\left[\hat{v}\hat{q}\right]}{E\left[\hat{h}\hat{q}\right]} \text{ and } \hat{z}^m - \bar{b}\hat{h} = \hat{z} + \left(\hat{v} - \bar{b}\hat{h}\right).$$

Notice that the hours-corrected Solow residual no longer provides an estimate of true productivity movements.

Can we say anything about the volatility of the true productivity shocks conditional on what we know about the workweek of capital? From equation (8) we can write the variance of the Solow residual $\sigma^2_{z^m}$ as a function of the variance of the true productivity shock σ^2_z, the variance of the workweek of capital σ^2_u, and the correlation of true productivity shocks and the workweek of capital ρ_{zu}. This expression defines an implicit equation in the volatility of the true productivity shock conditional on the volatility of the measured Solow residual, the volatility of the workweek of capital, and an assumption on the correlation coefficient between the true productivity shock and the workweek of capital:

$$0 = \sigma^2_z + \left(2\alpha\rho_{zu}\sigma_u\right)\sigma_z + \left(\alpha^2\sigma^2_u - \sigma^2_{z^m}\right).$$

Consider now the manufacturing sector. From Beaulieu and Mattey (1998), the volatility of the workweek of capital for the time period 1974–1992 is 2.9 percent, and from Basu, Fernald, and Kimball (1999), the volatility of the Solow residual is $\sigma_{z^m} = 3.1$ percent. Because Basu, Fernald, and Kimball (1999) estimate the Solow residual from gross output data, and intermediate inputs make up a substantial share of total payments to inputs, I choose a capital coefficient $\alpha = (1/3)\,(1/2)$. In Figure 3, I plot the implied volatility of the true productivity shock for values of the correlation coefficient between negative one and positive one. The two horizontal lines indicate the volatility of the unadjusted Solow residual and the hours-worked-adjusted Solow residual from Basu, Fernald, and Kimball (1999) for the manufacturing sector. We can see that the workweek-of-labor-corrected Solow residual underestimates (overestimates) the true volatility of the productivity shocks when the workweek of capital and the true productivity disturbance are weakly (strongly) correlated. The critical value for the correlation coefficient is 0.6. Since we expect a relatively strong positive correlation between the capital workweek and productivity disturbances, the actual bias might not be very large.

Suppose that Basu and Kimball's (1997) estimates of the volatility of production are unbiased. Can we say anything about how precise these estimates are? This question is relevant since the estimate of the coefficient b in equation (9) and the estimated productivity change $\hat{z}^m - \hat{b}\hat{h}$ are based on instrumental variables that are quite poor and the sample size is quite small (only 40 years). We can evaluate the uncertainty surrounding the estimates using the workweek of capital model described above. This model captures the qualitative features that the workweek of capital is more volatile than the workweek of

Table 3 Ratio of Estimated to True Volatility of Productivity Growth

Sample Size	Mean	Std. Dev.
40 years	0.88	0.35
200 years	1.02	0.18
400 years	1.03	0.12

labor and that the two workweeks are not perfectly correlated. In order to evaluate the uncertainty about the estimates, I generate 1,000 samples of 40 years of quarterly observations for the model. For each sample I construct annual data from the quarterly data and then estimate equation (9) with the annual data. In the model, government spending is exogenous and affects other endogenous variables, that is, it can be used as an instrumental variable. For the estimation of equation (9), I use contemporaneous and lagged growth of annual government spending as instruments. I then use equation (9) to calculate the volatility of estimated productivity growth rates. The model tells me the volatility of the two productivity growth rates. In Table 3, I display the means and standard deviations of the ratio of estimated to true productivity volatility across the samples. We can see that for a small sample of 40 years, on average we underestimate the true productivity volatility. More important, the standard deviation on the estimates is very large: the two-standard deviation error band for the ratio reaches from 0.18 to 1.58.

Given the small sample of available data, the estimate of true productivity volatility is very imprecise. Basu and Kimball (1997) increase the sample size by using industry data rather than aggregate data. They assume that in equation (9), the coefficient b is the same across industries, and then they pool industry data. For manufacturing they pool durable-goods-producing industries (ten industries) and nondurable-goods-producing industries (seven industries). I replicate the industry pooling approach by assuming that each industry represents another 40 years of observations. With more observations the estimate of productivity volatility appears to be unbiased (see Table 3). This result should not be too surprising since the differences between the workweek of capital and the workweek of labor are not quantitatively important in the model, as opposed to the data. We might therefore expect that the use of the workweek of labor rather than the workweek of capital in equation (9) would not generate a large bias in the estimate of the volatility of productivity. Even with a larger data set, however, the estimate of productivity volatility is not very precise: the two-standard deviation error band for the ratio still ranges from 0.79 to 1.27.

4. CONCLUSION

Variations in capital utilization as measured by the workweek of capital are large; indeed, the workweek of capital is substantially more volatile than the workweek of labor. This observation suggests that for output fluctuations, short-term variations in the utilization of the capital input are at least as important as short-term variations in the utilization of the labor input. Yet, official statistics are collected only for variations in the workweek of labor, not for the workweek of capital. Improved measurement of the workweek of capital is clearly called for (Shapiro 1996). Improved data would allow for a better assessment of the role of productivity disturbances.

REFERENCES

Basu, Susanto, and Miles S. Kimball. 1997. "Cyclical Productivity with Unobserved Input Variation." NBER Working Paper 5915.

_____, John Fernald, and Miles S. Kimball. 1999. "Are Technology Improvements Contractionary?" Manuscript, University of Michigan.

Beaulieu, J. Joseph, and Joe Mattey. 1998. "The Workweek of Capital and Capital Utilization in Manufacturing." *Journal of Productivity Analysis* 10 (October): 199–223.

Bertin, Amy L., Timothy F. Bresnahan, and Daniel M. G. Raff. 1996. "Localized Competition and the Aggregation of Plant-Level Increasing Returns: Blast Furnaces, 1929–1935." *Journal of Political Economy* 104 (April): 241–66.

Bils, Mark. 1995. "Measuring Returns to Scale from Shift Practices in Manufacturing." Manuscript, University of Rochester.

_____, and Jang-Ok Cho. 1994. "Cyclical Factor Utilization." *Journal of Monetary Economics* 33 (April): 319–54.

Bresnahan, Timothy F., and Valerie A. Ramey. 1994. "Output Fluctuations at the Plant Level." *Quarterly Journal of Economics* 109 (August): 593–624.

Burnside, Craig, and Martin Eichenbaum. 1996. "Factor-Hoarding and the Propagation of Business-Cycle Shocks." *American Economic Review* 86 (December): 1154–74.

_____, and Sergio Rebelo. 1993. "Labor Hoarding and the Business Cycle." *Journal of Political Economy* 101 (April): 245–73.

Corrado, Carrol, and Joe Mattey. 1997. "Capacity Utilization." *Journal of Economic Perspectives* 11 (Winter): 151–67.

Finn, Mary G. 1995. "Is 'High' Capacity Utilization Inflationary?" Federal Reserve Bank of Richmond *Economic Quarterly* 81 (Winter): 1–16.

Greenwood, Jeremy, Zvi Hercowitz, and Gregory W. Huffman. 1988. "Investment, Capacity Utilization, and the Real Business Cycle." *American Economic Review* 78 (June): 402–17.

Hall, George J. 1996. "Overtime, Effort, and the Propagation of Business Cycle Shocks." *Journal of Monetary Economics* 38 (August): 139–60.

_____. 2000. "Non-convex Costs and Capital Utilization: A Study of Production Scheduling at Automobile Assembly Plants." *Journal of Monetary Economics* 45 (June): 681–716.

Hornstein, Andreas, and Edward C. Prescott. 1993 [1990]. "The Firm and the Plant in General Equilibrium." In *General Equilibrium, Growth, and Trade. Volume 2. The Legacy of Lionel McKenzie,* ed. Robert Becker et al. San Diego, CA: Academic Press: 393–410.

Kydland, Finn E., and Edward C. Prescott. 1991. "Hours and Employment Variation in Business Cycle Theory." *Economic Theory* 1 (January): 63–81.

Mattey, Joe, and Steve Strongin. 1997. "Factor Utilization and Margins for Adjusting Output: Evidence from Manufacturing Plants." Federal Reserve Bank of San Francisco *Economic Review* 2: 3–17.

Shapiro, Matthew D. 1996. "Macroeconomic Implications of Variation in the Workweek of Capital." *Brookings Papers on Economic Activity* 2: 79–119.

Can Risk-Based Deposit Insurance Premiums Control Moral Hazard?

Edward Simpson Prescott

Calls for deposit insurance reform regularly sound the refrain to make deposit insurance premiums more risk based.[1] Those who support such a change believe that risk-based premiums will discourage insured banks from taking excessive risk because a bank facing higher premiums will think twice before undertaking a risky activity.

This logic seems impeccable: Let banks face the true cost of risk and they will appropriately balance the tradeoff between risk and return. While seemingly correct from the standard perspective of price theory, this argument requires the deposit insurer to be able to observe the risk characteristics of a bank's investment portfolio. There are good reasons to think that this is not the case; it is hard for outsiders to evaluate a bank loan or a complicated portfolio of financial derivatives. Under these conditions, risk-based deposit insurance premiums are not enough to control moral hazard. Instead, other devices such as performance-based insurance payments and supervisory monitoring are needed as well.

When one party to a transaction has information that the other party does not have, economists describe the transaction as one with *private information*. Various types of information may be private, but I am concerned with a payoff-relevant action. This model is sometimes referred to as the moral-hazard or hidden-action model. In this article, the action that may be hidden from others is the risk characteristics of a bank's investment decisions. The economic literature on moral hazard emphasizes the importance of state-contingent payments

The author would like to thank Huberto Ennis, Tom Humphrey, John Walter, Roy Webb, and John Weinberg for helpful comments. The views expressed in this article do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] For recent examples, see FDIC (2000) or Blinder and Wescott (2001).

for giving people the right incentives.[2] A simple example of state contingencies is a salary plus a commission. Sales representatives are frequently paid this way to give them an incentive to work hard. In contrast, risk-based deposit insurance premiums are not state contingent. They are entirely ex ante. As we will see, this limits their usefulness as a tool to control moral hazard.

I have three goals in this article. The first goal is to show what risk-based deposit insurance premiums can and cannot do. Risk-based premiums are useful for preventing transfers between different risk classes of banks, but they cannot control moral hazard. This idea is not new. It appears to be widely known among banking economists, but it rarely seems to have been formally expressed.[3] The second goal is to illustrate how state contingencies in deposit insurance payments can be used to control moral hazard. As indicated above, this illustration will use a model with private information. The final goal is to formally develop a role for supervisory activities like safety and soundness exams. These exams are modeled as a costly means for reducing the amount of private information between the deposit insurer and the bank.[4] Most of the literature on bank regulation takes the amount of private information as given. But as long as these supervisory activities reduce private information, they play a crucial role in any well-designed deposit insurance system.

The ideas in this article can be expressed with an analogy to an insurance contract. In dealing with different risks, insurance companies do more than adjust premiums. They also alter deductible amounts, copayment rates, and the probability of inspections. These contractual features are designed to prevent the insured from altering the risks it faces in a way that is detrimental to the insurance company, while still providing a degree of insurance. Of course, these characteristics of the insurance contract change with the risks, so in that sense well-designed deposit insurance contracts are risk based. Nevertheless, the premium level is not the only thing that changes. The analogy carries through to deposit insurance, which is why a well-designed deposit insurance system needs to do more than make premiums risk based.

1. THE MODEL

There is a deposit insurer who insures the depositors of one bank. The insurer is risk-neutral and has access to outside funds, so it has enough resources to cover its exposure. For simplicity, I assume that the bank is fully funded by

[2] For a survey of moral-hazard models, see Hart and Holmstrom (1987) or Prescott (1999).

[3] One exception is John, John, and Senbet (1991), and there are probably others as well.

[4] There is a literature on costly monitoring and auditing. Examples include Townsend (1979) and Dye (1986).

deposits.[5] I also ignore any liquidity or payment services provided by deposits. For my purposes, it is sufficient to treat deposits as just another form of debt. These deposits are fully insured and pay a gross rate of return of one.

The bank has access to several investment strategies. Each strategy requires one unit of capital to be invested. I assume that because of investment indivisibilities, the bank can engage in only one strategy at a time. The return r of each investment strategy i is uncertain. The probability distribution of returns for a given investment strategy is written $f(r|i)$. For simplicity, I assume that only a finite number of returns are possible. The bank is risk neutral but has limited liability. If the investment's return is less than one, the depositors receive everything produced by the bank plus enough of a payment from the deposit insurer that they receive the guaranteed gross return of one. If the return is greater than one, depositors receive a payment of one, any charges imposed by the deposit insurer are paid by the bank, and the bank keeps the remainder (if any) of its return.

The objective in this economy is to design the deposit insurance scheme so that the bank chooses the highest net present value investment project. Because of deposit insurance, however, meeting this objective is not straightforward. In the following sections, I work through the following three variations on the environment.

1. In the first variation, I assume that the deposit insurer observes the bank's investment strategy. Risk-based premiums are sufficient to control risk in this case.

2. In the second variation, I assume that the deposit insurer no longer observes the bank's investment strategy. This is the hidden-action or moral-hazard model. Risk-based premiums do not control moral hazard in this case and state-contingent payments are needed.

3. In the final variation, I develop a role for safety and soundness exams. The deposit insurer may spend resources that reduce (but do not eliminate) private information. In the example, the optimal deposit insurance system requires an exam in addition to state-contingent payments.

Full Information

In this section, I assume that the bank's investment decision is observed by the deposit insurer. In this case, economists say there is *full information*. It is under full-information conditions that risk-based deposit insurance premiums can succeed.

[5] For a related analysis of capital regulations, see Marshall and Prescott (2001) and Prescott (2001).

Table 1 Probability Distribution of Returns

	Return			
Investment	0.9	1.05	1.20	$E(r\|i)$
i_s	0.1	0.6	0.3	1.08
i_r	0.3	0.3	0.4	1.065

Notes: Probabilities and expected return of each investment strategy. The row labeled i_s corresponds to the high-mean, low-risk strategy, while the row labeled i_r corresponds to the low-mean, high-risk strategy. The last column lists the expected return or mean.

I illustrate this point with a simple example. Assume that the bank can choose between two investment choices. One of these choices is a low-risk, high-mean strategy, i_s, while the other is a high-risk, low-mean strategy, i_r.[6] There are three possible returns: a low one of 0.9, a medium one of 1.05, and a high one of 1.2. Table 1 lists the probability distribution of returns $f(r|i)$ as well as the expected return.

The socially desirable investment strategy is i_s. Its expected output is higher than that of the risky investment strategy i_r. The distribution of returns also differs between the two strategies. The safe strategy usually produces the medium return of 1.05, while the risky strategy is much more likely to produce either low or high returns.

Without Deposit Insurance

Without deposit insurance, the market prices deposits to reflect risk. If the risk-free rate on deposits were zero and the bank took investment strategy i_s, the depositors of the bank (assumed to be risk neutral) would require that the deposits pay 1.011 if the bank is solvent. This would give depositors an expected payoff of $0.1(0.9)+0.9(1.011) \approx 1.0$, which is equal to their expected payoff if they invested in risk-free assets. Alternatively, if the bank took investment strategy i_r, a similar analysis would find that depositors would require a payment of approximately 1.0429 to compensate them for the increased chance of the low return.

[6] Restricting the bank to two investment strategies is done mainly for expository purposes. Marshall and Prescott (2001) study a model where the bank can choose both the mean and variance characteristics of its loan portfolio. They find that the two investment strategies that mattered the most for deposit insurance are the low-risk, high-mean strategy and the high-risk, low-mean strategy. Restricting the investment strategies to these two choices is a stand-in for the more complicated problem.

The bank's payoff is the difference between its return and its payment to depositors. In either case, the expected gross return to depositors is 1.0, so the bank's expected payoff is

$$E(r|i) - 1.0. \tag{1}$$

Faced with this tradeoff, the bank would take the socially desirable investment strategy, i_s, because $E(r|i_s) - 1 > E(r|i_r) - 1$.

With Deposit Insurance

Improperly priced deposit insurance may distort the bank's preference-ordering over these choices. To see this distortion, consider the situation where the deposit insurance premium is independent of the bank's investment strategy.[7] Because of deposit insurance, depositors always receive 1.0. With limited liability, the bank's payoff function is $\max\{r - 1 - p, 0\}$, where 1 is the payment to depositors and p is the premium.[8] When the premium is set to zero, the bank's expected utility is

$$\sum_{r \geq 1.0} f(r|i)(r - 1.0) = E(r|i) - 1.0 + \sum_{r < 1.0} f(r|i)(1.0 - r). \tag{2}$$

Compared with equation (1), the bank's payoff without deposit insurance, the bank's utility under deposit insurance contains an additional term. This additional term is sometimes referred to as the value of the deposit insurance put option. It can be considered a put option because it allows the bank to dump its liabilities on the deposit insurer at a strike price of zero. It is valuable because with deposit insurance, risk is not reflected in the price of deposits. The lower rate paid on deposits leads to an increased payoff to the bank, the amount of which is the additional term. In essence it is a transfer from the deposit insurer to the bank; it also illustrates why underpriced deposit insurance can lead to a taste for risk. This last term increases as the expected transfer from the deposit insurer increases.

This taste for risk matters in the example. If premiums are set to zero, the bank prefers the risky strategy despite the higher expected return of the safe strategy. In particular, the return to the bank of the risky strategy is $0.3(0.0) + 0.3(0.05) + 0.4(0.2) = 0.095$, while the corresponding return of the safe strategy is only $0.1(0.0) + 0.6(0.05) + 0.3(0.2) = 0.09$.

---

[7] For early work identifying the risk-taking incentives created by deposit insurance, see Merton (1977) and Kareken and Wallace (1978).

[8] In practice, banks pay any premiums before investing the funds. Throughout this article I assume premiums are paid after the fact and use as our operational definition of a premium a constant payment that is made subject to limited liability. This assumption is made because I do not want to worry about how the deposit insurer invests the premiums it collects. The assumption does not alter the results.

Risk-based premiums can deal with these perverse incentives but only if the deposit insurer observes the investment strategy taken by the bank and makes the premiums dependent on it. Let the insurer index premiums by the bank's risk strategy, p_i, and set premiums to be actuarially fair.[9] The premium level for a given investment strategy i must satisfy

$$\sum_{r \geq 1.0+p_i} f(r|i)p_i + \sum_{1.0 \leq r < 1.0+p_i} f(r|i)(r-1.0) = \sum_{r<1.0} f(r|i)(1.0-r). \quad (3)$$

The left-hand side is the expected value of collected premiums. The second term on the left-hand side reflects the amount of funds collected by the insurer if the bank produces enough to pay depositors but not enough to pay the full amount of the premium. The right-hand side of equation (3) is the expected transfer made by the deposit insurer to depositors. Later it will be convenient to write (3) as

$$\sum_{r \geq 1.0+p_i} f(r|i)p_i = \sum_{r<1.0+p_i} f(r|i)(1.0-r).$$

Under this actuarially fair, risk-based premium schedule, the bank's expected payoff is

$$\begin{aligned}
\sum_{r \geq 1.0+p_i} (r-1.0-p_i) &= E(r|i) - \sum_{r<1.0+p_i} f(r|i)r - \sum_{r \geq 1.0+p_i} f(r|i)1.0 \\
&\quad - \sum_{r \geq 1.0+p_i} f(r|i)p_i \\
&= E(r|i) - \sum_{r<1.0+p_i} f(r|i)r - \sum_{r \geq 1.0+p_i} f(r|i)1.0 \\
&\quad - \sum_{r<1.0+p_i} f(r|i)(1.0-r) \\
&= E(r|i) - 1.0. \qquad (4)
\end{aligned}$$

This equation is identical to equation (1), which describes the expected payoff to the bank under the no deposit insurance case. There is equivalence because in the risk-based deposit insurance premium case, the premiums are set to exactly offset the expected payments made by the deposit insurer. In the context of equation (2), the premiums paid exactly offset the value to the bank of the deposit insurance put option. Consequently, just as in the no deposit insurance case, the bank will choose the safe investment strategy because it has the highest expected return.

[9] Analysis of deposit insurance usually operates under the assumption that actuarially fair deposit insurance is desirable. This mode of operation is based on the view that transfers to or from taxpayers are undesirable. For a deposit insurance model that argues that this view may be incorrect, see Boyd, Chang, and Smith (2001).

In the numerical example, the actuarially fair deposit insurance premium for investment strategy i_s is 0.011. (Recall that in this article the premium is being assessed after the return is realized, and to be consistent with limited liability the bank cannot pay its premium if it produces the low return of 0.9.) The corresponding premium for the i_r investment strategy is 0.0429. With these investment-dependent premiums the expected payoff to the bank of i_s is 0.08, while the corresponding payoff to the bank if it takes i_r is 0.065. Consequently, with risk-based deposit insurance premiums, the bank chooses the socially desirable investment.

This example illustrates the argument behind risk-based deposit insurance premiums. Risk-based premiums control risk because premiums can be made explicitly on the investment strategy, and if they are set to keep deposit insurance fairly priced, the bank faces the true costs of its investment decision. But this result depends on the insurer being able to ascertain just how risky a strategy the bank is taking, which it must be able to do in order to set the premiums properly. It is by no means clear, however, that assessing the bank's strategy is an easy task. As I mentioned earlier, the quality of a bank loan may be hard to determine, let alone the quality of an entire portfolio. Just witness the enormous debate and controversy over how to make the Basle capital regulations reflect risk more accurately.[10] In the next section, I will illustrate just how important the full-information assumption is and how the conclusions change when it is dropped. Those results will form the basis for my argument that risk-based premiums alone cannot control moral hazard.

Private Information

To illustrate the second variation on the environment, where the bank's investment strategy is private information, let us continue with the numerical example. The deposit insurer sets a risk-based premium of 0.011 if the bank takes the safe strategy and 0.0429 if it takes the risky strategy. But to implement this policy, the insurer has to know which strategy the bank takes. For the reasons described above, this knowledge is not easy to ascertain. What if the bank claims it is taking the safe strategy but is actually taking the risky strategy?

I can evaluate this possibility by setting the premium to 0.011, that of the safe strategy, and evaluating the expected payoff to the bank if it takes the risky strategy. Its payoff in this case is $0.3(0) + 0.3(1.05 - 1 - 0.011) + 0.4(1.2 - 1 - 0.011) = 0.0872$. This expected payoff is greater than 0.08, which is

[10] The 1988 Basle Accord assigned risk weights to different classes of assets and then set a minimum capital requirement based on the sum of these risks. There has been widespread dissatisfaction with the Accord because all loans of a particular class, such as Commercial and Industrial loans, are treated as equally risky. A major reconsideration of the Accord is underway right now, and the proposals for reform are based on trying to better ascertain risks at the level of individual loans.

what the bank would get if it took the safe strategy. This evaluation suggests that the insurer cannot use the risk-based premium schedule analyzed above to implement i_s.

Unlike in the previous section, the insurer does not observe the bank's investment strategy and the bank is therefore able to say that it is taking one strategy while it is really taking a different one. Economists say there is *private information* when information relevant to a transaction or a contractual arrangement is known to only one of the participants. In the context of deposit insurance pricing, private information puts limits on the types of pricing schemes that can be used. Economists deal with these limits by requiring contracts, or in this case pricing schemes, to be *incentive compatible*. A deposit insurance pricing scheme and an investment strategy are incentive compatible if under the scheme it is in the bank's best interest to take the investment strategy. In contrast, there is no such requirement in the full-information case. If the bank changes its strategy, the premium level can change with it.

As the above analysis indicates, a fixed premium and the socially desirable investment strategy i_s are not incentive compatible. The insurer can do better, however, if it does not restrict itself solely to premiums but also allows payments to depend on the realized return. More formally, I write these payments as $p(r)$. A deposit insurance premium is a special case of this function in which $p(r)$ equals a constant.[11] With this notation, I can more formally define incentive compatibility.

Definition 1 *A deposit insurance price system $p(r)$ and investment strategy i is incentive compatible if for all alternative investment strategies i'*

$$\sum_r f(r|i) \max\{r - p(r) - 1.0, 0\} \geq \sum_r f(r|i') \max\{r - p(r) - 1.0, 0\}.$$

In words, this definition says that for a given deposit insurance price system $p(r)$, the expected payoff a bank receives from taking investment i must be more than it would receive if it took any other possible investment strategy i'. For example, the safe investment strategy i_s is not incentive compatible when the fixed premium is set to 0.011. The risky investment strategy i_r, however, is incentive compatible for that same premium.

With private information, state-contingent payments may improve upon risk-based premiums (which are not state contingent). To see this, consider the following deposit insurance pricing scheme. If the bank produces the high return, charge it 0.053, and if it produces the middle return, rebate to it 0.01. Of course, no payments are made if the bank produces the low return since the bank fails in this event.

[11] Technically, in this article $p(r)$ is only a constant when the bank has enough funds to pay the premium.

The safe investment strategy is incentive compatible for this deposit insurance pricing system. If the bank chooses the safe investment strategy, it receives 0.08. (The number is unchanged from above since the price schedule was chosen to be actuarially fair.) Furthermore, incentive compatibility holds because the expected payoff to the bank from taking the risky strategy is now only 0.077.

This effect can be seen more formally through an analysis of the *likelihood ratios*. In moral hazard problems with recommended strategy i, the likelihood ratio for a given return r is the probability of r, given alternative investment strategy i' divided by the corresponding probability if the recommended strategy was taken. More formally, the ratio is $\frac{f(r|i')}{f(r|i)}$. Examination of the incentive constraint reveals the following. If $p(r)$ is set high when $\frac{p(r|i')}{p(r|i)}$ is high, a bank that takes i' is punished relatively more than a bank that takes the desired i. Similarly, if $p(r)$ is set low (or even negative) when this fraction is low, a bank that takes i' is rewarded relatively less than a bank that takes the desired i.

In this example, the likelihood ratio (when $i = i_s$ and $i' = i_r$) is high for the high return and low for the middle return. This property of the ratio generates the seemingly paradoxical result that the payment is higher if the highest return is produced.[12] But in this example, a low payment for the high return would give the bank too much of an incentive to take the risky investment strategy.[13] Finally, it is worth noting that the likelihood ratio is high for the low return as well, but because of limited liability the bank cannot make payments to the insurer.

Figure 1 illustrates why this pricing scheme is effective. The solid line depicts the payoff to the bank if it faces a fixed premium. The dashed line with the stars reports the payoff from a pricing schedule that collects all payments from the bank when the bank does very well. Notice how the shapes of the two functions differ. The solid line is convex, which means it rewards risk-taking.[14] The dashed line with the stars, while convex in portions, is basically a concave function. It does not reward risk-taking.

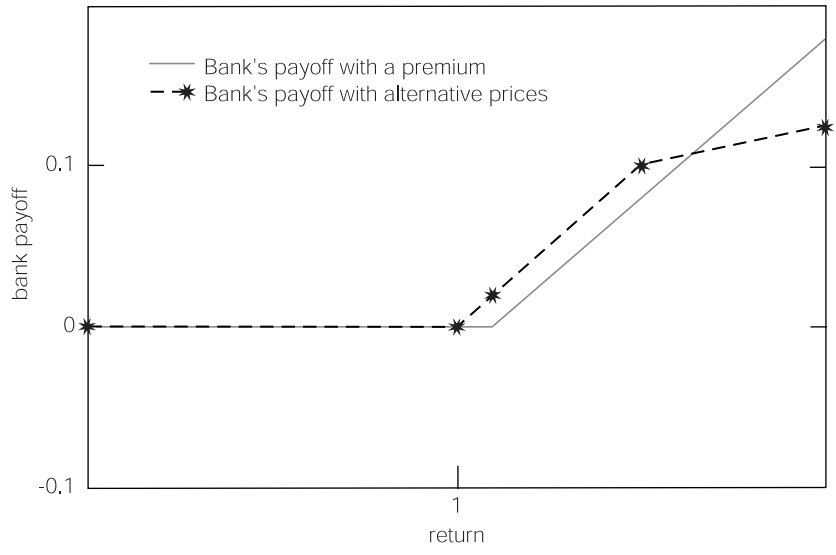
The lesson of this example is that risk-based premiums cannot control moral hazard on their own. Private information requires richer deposit insurance pricing schemes that take advantage of state-contingent pricing. This is not to say that risk-based premiums are not useful but that they are only one component of the entire deposit insurance price system. For example, if

[12] For similar results in the context of bank capital regulations, see Marshall and Prescott (2001) or Prescott (2001).

[13] One potential problem with this pricing scheme is that high returns could also reflect innovation. High payments for high returns would then have the undesirable effect of punishing innovation. The proper balance of these considerations is an open research question.

[14] In the full-information case, this shape did not cause the bank to prefer the risky investment because the premium level could change with investment strategy. Under private information, the premium does not change with the strategy so the convex shape becomes a problem.

Figure 1 Bank's Payoff as a Function of the Return



Notes: The solid line depicts the bank's payoff as a function of the return if it pays a fixed premium of 0.02. The dashed line with the stars represents the bank's payoff for a deposit insurance pricing system that charges no premium but requires a payment if the bank produces a return greater than 1.1. For both payoff functions, the horizontal portion reflects limited liability. Because of limited liability, a bank facing a fixed premium has a convex payoff function. Payoff functions with this shape create a taste for risk. (To see this draw a line between a return on the horizontal portion of the payoff function and a return on the increasing portion. Randomizing over these two returns is preferred to the certain production of the expected amount.) A bank that faces the alternative price schedule has a payoff function that is almost concave, with only a portion being convex. Concave payoff functions create a distaste for risk.

some investment decisions are easy to observe, like the class of investments a bank specializes in, then the analysis will contain elements of both the full information and private information models. In this case, there could be one pricing scheme for banks that specialize in real estate lending and another pricing scheme for banks that hold safe assets like Treasuries. The real estate lending bank might face high premiums plus state-contingent payments, while the Treasury-holding bank might face low premiums and relatively non-state-contingent payments. The pricing scheme is risk based as advocated by proponents of risk-based deposit insurance premiums, but, as my analysis suggests, the pricing scheme would also be state contingent.

Changing the Information Structure

The previous analysis focused on how a price system with state-contingent pricing could improve upon narrow risk-based premium systems. Indeed, the state-contingent price system was successful at implementing the safe, socially desirable investment strategy. The example should not be taken, however, to mean that state-contingent pricing can control all of the moral hazard created by deposit insurance. In many moral hazard problems, the best incentive-compatible contract only partially mitigates the moral hazard.

In this section, I consider the third variation on the environment by providing a private information environment where the insurer can take some costly action that lets it observe some of the private information. This analysis can be used to form the basis for analyzing numerous supervisory activities like safety and soundness exams, audits, and off-site surveillance. As we will see, these activities can play a crucial role in a well-designed deposit insurance pricing system.

To illustrate this principle, I return to the example used in the above section. Now, however, I assume that it costs the bank effort and resources to screen its investment portfolio in order to identify the i_s investment strategy. If the bank does not supply this effort, it cannot take the i_s strategy. The effort cost translates directly into a utility loss to the bank that corresponds to a drop in its payoff of 0.05 units. This loss is not affected by limited liability. The idea is that this loss corresponds to effort by bank management. The bank can choose not to supply the screening effort. If it takes this route, it saves utility but must choose investment strategy i_r. As before, I assume that the socially desirable investment strategy is for the bank to take i_s.[15]

The incentive problem here is more severe than in the previous example. Before it was only necessary to worry that the bank might take the risky strategy. Now, however, it is also necessary to worry that the bank might not screen its portfolio and then take the risky strategy by default. If it does not screen its portfolio, it saves on the utility cost of 0.05. This additional saving is important for the incentive constraints. In particular, the safe investment strategy cannot be implemented with the deposit insurance pricing schedule examined above. Furthermore, this strategy cannot be implemented for any actuarially fair deposit insurance pricing scheme.[16]

---

[15] In making this assumption, I am ignoring the utility cost to the bank in my welfare calculation. This assumption keeps the problem simple.

[16] For the example, an actuarially fair pricing scheme must satisfy

$$0.6p(r_m) + 0.3p(r_h) \quad = \quad 0.01,$$

where $p(r_m)$ is the payment made if the medium return is generated and $p(r_h)$ is the payment made if the high return is generated. The right-hand side is 0.01 because that is the expected payment made by the deposit insurer to the depositors.

For i_s to be incentive compatible, the pricing scheme must satisfy the incentive compatibility

What is the insurer to do? Let us make one last addition to the environment and allow the insurer to spend 0.02 units examining the bank. By examining the bank, the insurer does not observe which investment strategy the bank takes, but it can tell if it expended the effort to properly screen the projects. Observing this effort could be interpreted as examiners checking bank lending procedures or resources devoted to risk management.

If the insurer examines the bank, the problem is identical to that of the previous section except that now the insurer also has to make up the examination cost of 0.02 units from its pricing scheme. It can recover these funds by setting the rebate to zero and raising the charge on the high return to 0.10. Under this deposit insurance pricing and inspection system, it is incentive compatible for the bank to screen and then take the safe investment strategy. The exam prevents the bank from not screening and once it screens, the state-contingent payments convince the bank to take the safe investment strategy. Finally, the deposit insurance price system is actuarially fair (including examination costs), so no resources are transferred in or out of the banking system in expectation.

The key feature of this example is the way in which the examination policy changes the information structure of the bank. In this example, the information is revealed in a straightforward manner. More generally, examinations or other types of supervisory monitoring may only reveal signals that are partially correlated with the true action. Or, supervisors may want to use the information they receive from inexpensive information gathering methods, like balance sheet observations, to decide whether or not they should gather more information using more costly methods like on-site exams. All these possibilities can be added to the framework developed in this article.

2. CONCLUSION

This article argues that risk-based deposit insurance premiums alone cannot control moral hazard in deposit insurance. The examples demonstrate how richer procedures with more complicated pricing schedules and examination procedures can be more useful than risk-based deposit premiums. The critical factor in the analysis is private information.

Interesting parallels to the analysis exist in markets without government insurance. As was discussed earlier, insurance contracts include deductibles and copayments and may allow for audits to control moral hazard.[17] Banks

---

constraint

$$-0.3p(r_m) + 0.1p(r_h) \geq 0.055.$$

Furthermore, the payments are subject to limited liability, which means that $p(r_m) \leq 0.05$ and $p(r_h) \leq 0.2$. A simple graph reveals that there is no pair $(p(r_m), p(r_h))$ that satisfies these four equations.

[17] Experience rating is an important tool used by insurance companies that was not addressed

also take several actions to mitigate the private information of their borrowers. For example, they regularly impose covenants on their borrowers' actions and they often list conditions under which they can call a loan.[18] Just as there is more to the price of a bank loan than the interest rate, there is more to pricing deposit insurance than insurance premium levels.

REFERENCES

Black, Fisher, Merton H. Miller, and Richard A. Posner. 1978. "An Approach to the Regulation of Bank Holding Companies." *Journal of Business* 51: 379–412.

Blinder, Alan S., and Robert F. Wescott. 2001. "Reform of Deposit Insurance: A Report to the FDIC." (March).

Boyd, John H., Chun Chang, and Bruce D. Smith. 2001. "Deposit Insurance: A Reconsideration." Manuscript, Carlson School of Management, University of Minnesota.

Dye, Ronald A. 1986. "Optimal Monitoring Policies in Agencies." *RAND Journal of Economics* 17: 339–50.

Federal Deposit Insurance Corporation. 2000. "Options Paper." (March).

Hart, Oliver D., and Bengt Holmstrom. 1987. "The Theory of Contracts." In *Advances in Economic Theory: Fifth World Congress,* ed. Truman F. Bewley. Cambridge: Cambridge University Press: 71–155.

John, Kose, Teresa A. John, and Lemma W. Senbet. 1991. "Risk-Shifting Incentives of Depository Institutions: A New Perspective on Federal Deposit Insurance Reform." *Journal of Banking and Finance* 15: 895–915.

Kareken, John H., and Neil Wallace. 1978. "Deposit Insurance and Bank Regulation: A Partial-Equilibrium Exposition." *Journal of Business* 51: 413–38.

Keeley, M. C. 1990. "Deposit Insurance, Risk, and Market Power in Banking." *American Economic Review* 80: 1183–1200.

in this article. Indeed, an experience rating may be a partial substitute for state-contingent payments by the bank. This omission was made for simplicity; static models are a lot easier to work with than dynamic ones. Nevertheless, this tool may be very important and deserves to receive more attention than it has received from the bank regulation literature.

[18] For examples and further discussion of the parallels between private lending and how bank regulation should be structured, see Black, Miller, and Posner (1978).

Marshall, David A., and Edward S. Prescott. 2001. "Bank Capital Regulation With and Without State-Contingent Penalties." *Carnegie-Rochester Conference on Public Policy.* Forthcoming.

Merton, Robert C. 1977. "An Analytic Derivation of the Cost of Deposit Insurance Guarantees." *Journal of Banking and Finance* 1: 3–11.

Prescott, Edward S. 1999. "A Primer on Moral-Hazard Models." Federal Reserve Bank of Richmond *Economic Quarterly* 85 (Winter): 47–77.

_____. 2001. "Regulating Bank Capital Structure to Control Risk." Federal Reserve Bank of Richmond *Economic Quarterly* 87 (Summer): 35–52.