On the Size Distribution of Banks

Huberto M. Ennis

In recent years, important changes to the U.S. banking regulatory framework have been introduced that were expected to affect the size distribution of banks. These changes in regulation had a clear objective: to allow for a higher degree of horizontal and vertical integration in the banking industry. While horizontal integration takes place when different firms that are producing the same product merge, vertical integration takes place when firms producing certain inputs merge with the firms that use those inputs.

The Riegle-Neal Interstate Banking and Branching Efficiency Act was passed in September 1994. The act allows banks and bank-holding companies to freely establish branches across state lines. In fact, the act came as the final step in a long process of gradual removal of interstate branching restrictions that took place at the state level during the late eighties and early nineties. This new flexibility in the branching regulation has opened the door to the possibility of substantial geographical consolidation in the banking industry. Indeed, geographical consolidation has always been one of the main channels used to achieve *horizontal* integration at an industry level.

In November 1999 Congress passed the Gramm-Leach-Bliley Financial Services Modernization Act. It allows affiliations among banks, securities firms, and insurance companies, removing many long-standing restrictions over the *horizontal and vertical* integration of firms providing financial services.

Both of these regulatory changes were expected to have substantial effects on the overall structure of the U.S. banking sector, and in particular, on the size

Research Department, Federal Reserve Bank of Richmond, huberto.ennis@rich.frb.org. The author wishes to thank Kevin J. Scotto for excellent research assistance and Ned Prescott for useful conversations on the subject. Kartik Athreya, Tom Humphrey, Pedro M. Oviedo, John Weinberg, Alex Wolman, Jose Wynne, and the seminar participants at NCSU and Duke University provided useful comments on a previous draft. All errors are my own. The views expressed here do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

distribution of banks. Some of these effects are already apparent in the data, and there may be more to come. It is not yet clear if the transition period is over. The question of whether *all* banks will eventually become nationwide banks is still very much unanswered. In other words, is there something special that community banks do which nationwide banks cannot replicate, or are small regional banks simply a consequence of long-lasting and strict government regulations? Even seven years after passage of the Riegle-Neal Act, there are still 7,920 small commercial banks (with less than a billon dollars in assets) representing 95 percent of the total number of banks in the system and holding 20 percent of total deposits. At the same time, there are 82 banks with more than $10 billon in assets that hold 70 percent of total deposits. These statistics indicate that even though some very large banks have already emerged, there are still many small banks with substantial participation in the administration of deposits.

In this article I will present some empirical and theoretical elements that could be used to support the view that the existence of community banks is justified even in an unregulated environment. Although the evidence is still preliminary, some interesting insights about the determinants of the banking industry structure arise from the discussion and can provide guidance for evaluating the future evolution of this important sector of the U.S. financial system.

The objective of the article is twofold. In Section 1, I will review stylized facts associated with the U.S. size distribution of banks and its evolution over the last 25 years. I will also include a brief discussion of the recent changes in U.S. regulation. Then, in Section 2, I will review some theoretical explanations for the coexistence of small and large banks in a competitive unregulated system. Section 3 provides conclusions.

1. SOME STYLIZED FACTS

Review of the Regulation

The Riegle-Neal Act is the final stage of a long process of bank branching deregulation in the United States. In 1975, no state allowed out-of-state bank holding companies to buy in-state banks, only 14 states permitted statewide branching, and 12 states completely prohibited branching. The rest had partial restrictions (for example, in some states a bank could only open branches in the county of its headquarters or in contiguous counties). These restrictions date from the Banking Act of 1933. However, starting in the late 1970s and continuing through the 1980s, all states relaxed their restrictions on both statewide and interstate branching (see Jayaratne and Strahan [1997, Table 1] for a list of the specific dates). Finally, in 1994 the Riegle-Neal Act removed all remaining restrictions on branching throughout the country.

It is probably safe to say that by the mid-1970s, the shape of the size distribution of banks fully reflected the effects of the branching restrictions that had been introduced 40 years earlier. Furthermore, the movement towards removing those restrictions in the 1980s surely explains the subsequent evolution of the distribution.

In the last decade, there has been a strong trend towards higher asset concentration in the industry.[1] One way to explain this trend is to acknowledge that the branching restrictions were probably highly binding while in place. Another and perhaps more interesting explanation is that the trend towards concentration appeared during a period when important technological innovations developed. There is little doubt that technological changes like computers and ATMs can help explain the observed increase in bank asset concentration. In fact, the potential efficiency gains associated with becoming a large high-tech bank may actually explain the political pressure for deregulation (see Broaddus [1998]). Deregulation is, to some degree, an endogenous event.

The fact that deregulation and technological innovation happened simultaneously has made it difficult to disentangle the independent effects of each of these factors on the size distribution of banks. Deregulation was a necessary condition for concentration, but probably not a sufficient one.

In 1999, the U.S. Congress passed another important piece of legislation that may strongly affect the market structure of the banking industry. The Gramm-Leach-Bliley Act created a new institution, the financial holding company, and allowed this new entity to offer banking, securities, and insurance products under one corporate roof. The law is still too recent to allow us to evaluate its long run impact on the financial services industry. However, two years after the law's enactment, there are a large number of banks that have taken advantage of the resulting opportunities for horizontal and vertical integration. Indeed, as of July 2001, 558 financial holding companies have been formed and 19 of the 20 largest banks in the United States now belong to a financial holding company.

The Gramm-Leach-Bliley Act also has provisions intended to increase competition and efficiency in the industry. Making an industry more competitive and efficient can change the flows of entry and exit, the optimal scale of operation, and the possibilities of growth at the firm level. These changes may in turn reshape the long-run size distribution of the surviving firms. However, it is still too early to conduct any conclusive evaluation of the actual effects of these provisions.

---

[1] As of March 2001, there were 18 commercial banks with more than $50 billon in assets; 8 of them had more than 1,000 branches in the United States. (The largest commercial bank, Bank of America, had more than $500 billon in assets and over 4,500 branches in the United States.)

There is another feature of the regulatory environment that can have important implications for the observed size distribution of banks. If the participants in the financial system have the perception that there exists a "too-big-to-fail" bias in the way regulators treat large institutions, then the level of asset concentration in the industry will tend to be higher and the size distribution more skewed to the right (with a disproportionately long right tail). Being a large institution presumably increases the ability of a bank to access the implicit subsidy associated with a too-big-to-fail policy. The existence of this type of policy in the U.S. banking industry is the subject of an ongoing debate (see, for example, Feldman and Rolnick [1997]). Furthermore, and most important for this article, it seems that isolating the effects of this particular policy over the scale of operation of banks can be a very complicated enterprise.

Data

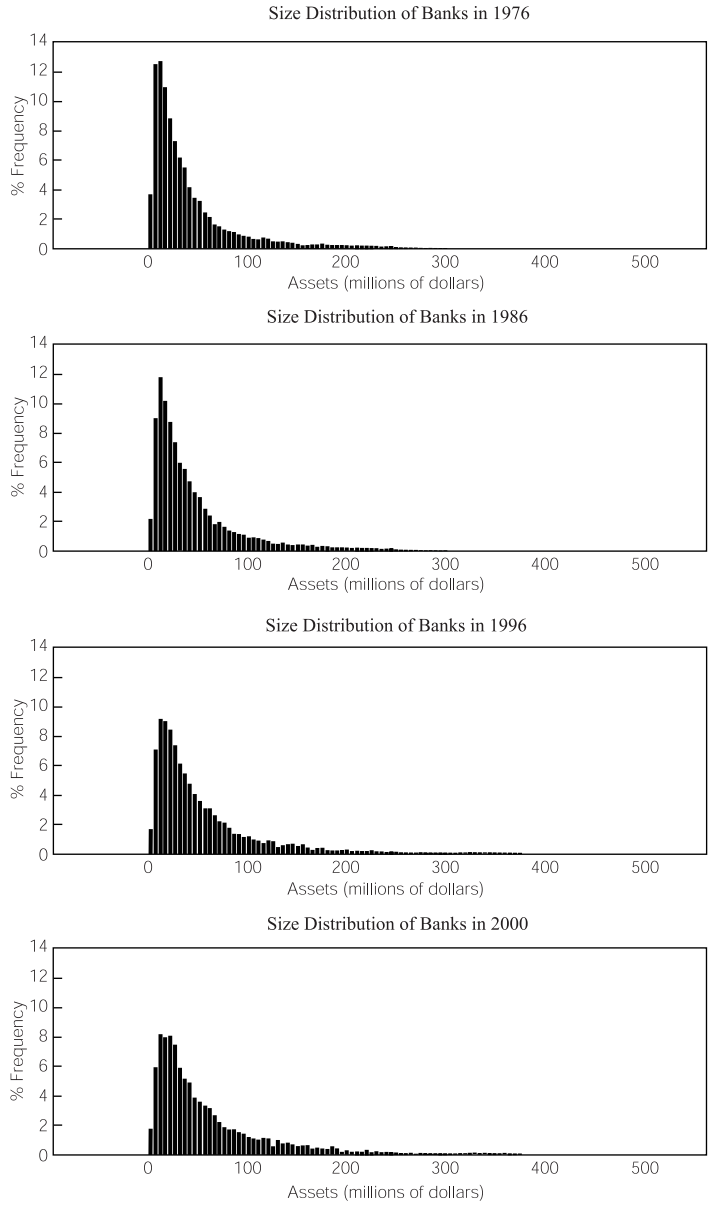
I now will present some statistics to characterize the size distribution of commercial banks in the United States and its evolution since 1976.[2] I use total assets to proxy the size of each bank, and all values are in real terms (dollars of 1982–1984). Figure 1 presents the histogram for the bottom (smallest) 95 percent of the total number of banks in each of four years 1976, 1986, 1996, and 2000. There is a wide range of bank sizes in each year. The distribution has shifted substantially in the last two decades. The average size has more than doubled (see Table 1). The density (frequency) of very small banks has clearly diminished.

Although there is a large number of small banks, the concentration in the industry is relatively high. Asset concentration has also increased in recent years. In Table 1, I report a time series for the Gini Coefficient of the asset size distribution in the industry.[3] The Gini Coefficient is relatively stable during the 1980s (with a value of around 0.84), but increases substantially after 1993 (reaching 0.90 in 2000). A noteworthy observation is that the density of midsize banks has increased. An important factor to have in mind when interpreting this fact is that the total number of banks in the system has been diminishing in the last decade, which means that higher densities do not imply a larger number of banks in certain ranges of the distribution. Figure 2 presents the histograms for banks with less than $400 million in assets (13,452 banks in 1986 and 7,745 in 2000). There seems to be a significant shift of the

[2] The data used here are from the Federal Reserve Bank of Chicago website (http://www.chicagofed.org/economicresearchanddata/data/bhcdatabase/subfiles.cfm).

[3] The Gini Coefficient is a measure of the degree of concentration associated with a given distribution of assets in the industry. It would be approximately equal to one when only 1 percent of the banks (the large banks) hold 99 percent of the assets in the industry and approximately equal to zero when all banks are of the same size.

Figure 1 Histogram of Bank Sizes (by Total Assets) (I)



mass of banks towards the right end of the distribution (although the absolute number of banks has been falling for almost all categories). In other words, compared with the size distribution of banks 20 years ago, today's distribution

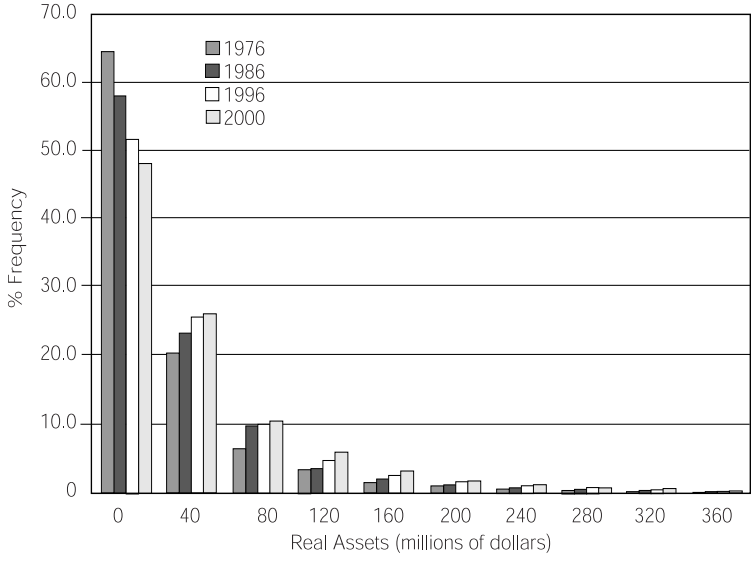
Table 1 Asset Concentration (I)

Year	Gini Coef.	Std. Dev. (Mean)	Number of Banks
1976	0.82	1,828 (140)	14,419
1977	0.83	1,965 (149)	14,417
1978	0.83	2,050 (154)	14,392
1979	0.83	2,077 (153)	14,356
1980	0.83	1,998 (149)	14,426
1981	0.83	1,953 (149)	14,407
1982	0.83	1,959 (155)	14,430
1983	0.83	1,877 (160)	14,420
1984	0.83	1,854 (165)	14,388
1985	0.84	1,921 (174)	14,278
1986	0.84	1,996 (188)	14,059
1987	0.84	1,932 (190)	13,553
1988	0.85	1,889 (199)	12,982
1989	0.85	1,933 (207)	12,572
1990	0.85	1,867 (206)	12,212
1991	0.85	1,883 (209)	11,807
1992	0.84	2,017 (216)	11,363
1993	0.85	2,188 (232)	10,881
1994	0.86	2,509 (256)	10,381
1995	0.87	2,749 (282)	9,875
1996	0.88	3,318 (302)	9,465
1997	0.89	4,055 (339)	9,081
1998	0.89	4,679 (377)	8,713
1999	0.90	5,476 (395)	8,520
2000	0.90	5,861 (427)	8,252

The mean and the standard deviation are in millions of 1982–1984 dollars.

has relatively fewer small banks, and, conditional on being small, banks tend to be larger today than in the past. It is not the case, then, that the very small banks disappearing in large numbers are losing all their market share to the extremely large national institutions. Intermediate-size banks are becoming relatively more important, too. In fact, the reduction in the number of small banks is especially concentrated on banks with less than 120 million dollars in assets, accounting for more than 96 percent of the reduction in the number of banks with less than 400 million (from 12,060 in 1986 to 6,558 in 2000). However, this shift in the relative mass of banks could be a consequence of the

Figure 2 Histogram of Bank Sizes (II)



transition process if very small banks are easier to take over and medium-size banks are simply in transition on their way to becoming larger institutions.

Table 2 further documents the level of concentration in the industry and its evolution over time. Again the table shows that concentration was stable (or slightly increasing) during the eighties and the early nineties and has significantly increased in the second half of the nineties. It is striking to note that the top 1 percent of the banks in the year 2000 own almost 70 percent of the assets (and the top 10 percent own almost 90 percent).

Table 3 presents some measures of the skewness (or asymmetry) of the distribution. In a symmetric distribution, the mean is located at the 50th percentile and the ratio of the mean to the median is 1. The bigger the concentration of assets in a few large banks, the more skewed to the right is the distribution. Indeed, according to the indicators in Table 3, the skewness of the asset distribution of banks has increased substantially during the nineties.

To try to determine the effect of government branching restrictions on the size distribution of banks, one can compare the distribution at the national level with that of a large state like California (Berger, Kashyap, and Scalise 1995). California has had no restrictions on statewide branching since the year 1909. The Gini Coefficient for the size distribution of banks in California was around 0.9 for most of the eighties and nineties, and the percentile location of the mean was around 94 percent. In summary, the concentration and the

Table 2 Asset Concentration (II)

Year	% of Assets (largest 1% of banks)	% of Assets (largest 10% of banks)	Ratio of largest 1% to smallest 40%	Ratio of largest 10% to smallest 40%
1976	55.8	78.1	15.6	21.8
1977	56.0	78.2	15.8	22.1
1978	56.8	78.7	16.4	22.7
1979	58.1	79.3	17.3	23.6
1980	58.1	79.4	17.1	23.4
1981	57.9	79.3	16.9	23.1
1982	57.3	79.2	16.8	23.2
1983	55.9	78.8	16.0	22.6
1984	55.6	79.0	16.2	23.1
1985	55.5	79.7	16.8	24.1
1986	55.4	80.1	17.2	24.8
1987	55.1	80.6	17.5	25.6
1988	54.7	81.1	18.0	26.8
1989	54.6	81.4	18.6	27.8
1990	54.1	81.3	18.2	27.3
1991	53.6	81.2	17.7	26.8
1992	54.0	81.1	17.6	26.5
1993	55.3	82.1	18.9	28.1
1994	56.7	83.5	21.2	31.2
1995	57.3	84.2	22.8	33.4
1996	60.9	85.0	25.8	36.0
1997	66.5	86.4	31.1	40.4
1998	68.0	87.2	33.8	43.4
1999	68.5	87.5	35.5	45.3
2000	70.2	88.2	38.6	48.5

skewness of the size distribution of banks in California during the eighties and nineties was very similar to that observed today for the national numbers.[4]

It is worth mentioning that using California as a benchmark for comparison became a less meaningful exercise after the mid-nineties deregulation of interstate branching. Indeed, in the last three or four years, changes at the national level have had some important indirect effects at the state level. Those

[4] During the seventies, bank-asset concentration in California was even higher (with a Gini Coefficient of around 0.94).

Table 3 Skewness

Year	Percentile Location of Mean	Ratio of Mean to Median
1976	90.6	4.9
1977	90.5	4.9
1978	90.8	5.0
1979	91.2	5.1
1980	91.3	5.2
1981	91.4	5.1
1982	91.4	5.1
1983	91.1	5.0
1984	90.8	5.1
1985	91.0	5.3
1986	91.1	5.4
1987	91.2	5.5
1988	91.3	5.8
1989	91.4	5.9
1990	91.2	5.9
1991	91.2	5.9
1992	91.3	5.8
1993	91.8	6.1
1994	92.1	6.6
1995	92.2	6.9
1996	92.7	7.3
1997	93.4	8.1
1998	93.8	8.5
1999	93.9	8.9
2000	94.2	9.5

effects were not present previously because the branching restrictions made California an isolated market.[5]

The histograms of bank sizes presented in Figure 1 resemble the probability distribution of a lognormal random variable. The lognormal distribution has been important in theoretical and empirical research. One of the most influential theories of the size distribution of firms was introduced by Robert

[5] In recent years, the measures of concentration and skewness for California have suffered large swings due to the fact that large state banks have merged with out-of-state banks and, in the process, have changed the location of their headquarters. (For example, from 1998 to 1999 the Gini Coefficient dropped from 0.92 to 0.84.)

Gibrat in the 1930s (see Sutton [1997]). His theory delivers a precise prediction for the long-run distribution of firm sizes: the lognormal distribution. Two strong assumptions are behind this prediction: (1) the number of firms is stationary and (2) the rate of growth of firms is given by an i.i.d. random variable independent of firm size. If one is willing to accept these assumptions as providing a reasonable representation of the evolution of a particular industry, then one can expect that the distribution of firm sizes will converge to the lognormal distribution.[6] Additionally, the lognormal distribution is a very convenient tool for analytical work. If a variable is lognormal, then the logarithm of that variable has a normal distribution. This means that a simple transformation of the data allows the researcher to apply all the well-known results associated with the normal distribution.

Because of the potential importance of lognormality, in Table 4 I perform some preliminary tests to see how far the U.S. commercial bank data is from delivering the lognormal distribution. The match is not very promising. The distribution of the logarithm of bank asset-size is relatively skewed to the right and has a higher degree of kurtosis (fatter tails or higher "peakedness," or both) than the normal distribution. Since the number of observations for each year is very large (around 10,000) we can safely conclude that these differences are not associated with sampling error: the distributions are significantly different. However, it should be said that during the years under consideration the industry has experienced important changes, and these calculations are not really appropriate as a test of Gibrat's theory (for that we would have to somehow control for the large flow of exit that took place in the industry).

On a related point, Simon and Bonini (1958) show that firm-entry assumptions matter for the determination of the stationary distribution. In particular, they combine Gibrat's firm-growth proportionality assumption with the assumption that new small firms enter the industry at a constant rate, and they show that the long-run size distribution approaches the Yule distribution (which has a fatter right tail than the log-normal).

2. SOME THEORETICAL EXPLANATIONS

There is an extensive literature on the size distribution of business firms that goes back to Gibrat's work during the 1930s. The literature on the size distribution of financial firms, however, is much smaller. In this section, I first

[6] This is actually not hard to see. Denote the size of the firm by x_t and let the i.i.d. random variable ε_t be a proportional rate of growth of the firm size. Then, we have that

$$\log x_t = \log x_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t,$$

and the distribution of $\log x_t$ converges to the normal distribution as $t \to \infty$.

Table 4 Skewness and Kurtosis of the Log Data

Year	α_3 (Skewness)	α_4 (Kurtosis)	Ratio of Mean to Median
1976	1.02	5.80	1.0106
1986	1.13	6.07	1.0123
1996	1.23	6.52	1.0139
2000	1.25	6.89	1.0140
Normal Distribution	0.00	3.00	1.0000

The statistic $\alpha_3 = \mu_3/(\mu_2)^{3/2}$ and $\alpha_4 = \mu_4/(\mu_2)^2$ where μ_i is the ith moment about the mean, which is given by $\mu_i = (1/N)\sum_{j=1}^{N}(x_j - \mu)^i$ (μ is the mean of the distribution of x_js, N is the total number of banks and x_j is the asset size of bank j).

discuss in some detail one possible theory of bank size heterogeneity and then review some complementary theories available in the literature.

Explaining the size distribution of banks is a challenging task. There is always the possibility of extending the explanations used for business firms to the financial sector. Indeed, several of these theories are probably useful for explaining some of the size heterogeneity observed in the banking industry. But it seems that these theories will always be partial insofar as they do not recognize that there are some special characteristics of financial firms that act as the essential determinants of the size distribution of banks. One of these special characteristics is that banks play a role as information managers in the provision of credit. In the next subsection, I present a formal model that uses this characteristic to deliver a theory of the equilibrium heterogeneity of bank sizes.

The more traditional theories of firm size heterogeneity are founded on the notion of an underlying life cycle of firms.[7] The idea is that firms tend to be small at birth, after which they experience partially stochastic growth. This process generates a level of size heterogeneity in the long run that is not too far from the one observed in the business firm data. However, in this view, there is nothing special that small banks are doing which makes them different from large banks; they are just in the process of growing. This description does not seem to be a good representation of the U.S. banking industry, in which there is a large number of small banks that are not growing substantially through time and have no apparent intention of growing. The model presented next tries

[7] See Jovanovic (1982), Hopenhayn (1992), and Ericson and Pakes (1995). These models are modern versions of the traditional Gibrat's theory. They endogenize the growth process of firms and the decision of entry and exit.

to capture this point. It represents an economy where there are two different ways of organizing the production of information services by banks (one with small local banks and the other with large national banks) and these two organizational practices can coexist in equilibrium. In the second subsection, I discuss some alternative explanations of bank size heterogeneity that, in principle, should be taken as complementary to the formal model presented in the first part.

A Simple Model of the Size Distribution of Banks

I study an environment where two types of banks can coexist in equilibrium. On one side there is a large, geographically diversified national bank, with high leverage ratio (i.e., low bank-equity capital), and on the other side there are several small community banks restricting operation to one geographical area (hence not well diversified) and with lower leverage ratios.

The main motivation for the existence of banks in this model is their ability to monitor the behavior of investors with financial needs. Several other possible banking functions, including mobilizing funds and pooling risks, do not play a role in the present model. Banks can monitor investors, but monitoring is costly and not observable by third parties. If the bank is not well diversified, then it has to commit some of its own funds (i.e., hold some capital) so that depositors will become confident that the bank will perform the required monitoring activities. Because of this need for own funds, and because there are some fixed costs associated with becoming a bank, only wealthy individuals choose to become community bankers. The national bank, on the other hand, is well diversified and its owner does not need to commit his or her own funds to the operation. However, running a large institution involves some extra operational costs. Because of the economies of scale associated with the fixed cost of setting up a bank, only one diversified bank exists in equilibrium. Having only one national bank is an extreme situation but of no fundamental importance for the points that I intend to illustrate with the model. A minor extension of the model would allow for the existence of several large banks in equilibrium (for example, by introducing managerial ability, as in Lucas [1978]).[8]

The main idea underlying the model is that there are two possible ways to provide a specific service (in this case, management of information). One way is to run a community bank with high capital ratios and low operating costs and the other way is to run a national bank with low capital ratios and

[8] Another way to generate a bounded optimal size of the diversified banks is to assume that the average cost of monitoring, constant in the present article, is instead increasing in the size of the bank (see Cerasi and Daltung [2000]).

higher operating costs. Both ways can be made equally efficient and hence can coexist in equilibrium.

Two interesting results emerge from the comparison of the equilibria when there is a national bank in the system and when national banks are exogenously ruled out (for example by regulation). First, lower levels of total investment are observed in the equilibrium with no national bank. Second, in the equilibrium with a national bank there are fewer community banks and they tend to be smaller in size. Some of these facts are consistent with the evolution of the U.S. banking industry after branching deregulation (see Section 2).

I turn now to the details of the model.[9] Assume that there are a large number of different geographic (or economic) zones in the economy. There is a continuum of risk-neutral investors living in each zone. For simplicity I assume the population of investors in each zone has size 1. Investors are indexed by the amount of funds they own. Let $\widetilde{\theta} \in [0, 1]$ be the amount of funds owned by investor $\widetilde{\theta}$. We also assume that there is only one investor for each level of $\widetilde{\theta}$. A more general assumption would be needed to obtain a realistic size distribution of banks. At the beginning of the period, agents have to invest (or store) their funds in order to have them back at the end of the period when they will be used to pay for consumption.

Each zone has available a large number of risky investment projects. Each project is associated with an entrepreneur and, when undertaken, can either succeed or fail. We index projects by their productivity when success occurs, $r_A \in [1, 2]$, and projects are uniformly distributed across the possible values of r_A. Success and failure are verifiable, but the value of r_A is private information to the entrepreneur. When the project fails, the return is zero. In other words, project r_A has productivity r_A when success happens. Each project is owned by an entrepreneur that can choose to exert effort in running the project. A project requires I units of funds to be undertaken. If the project is undertaken with effort, the probability of success is given by p_H. The probability of success for projects undertaken with no effort is p_L. We assume that p_H is greater than p_L. Projects within a zone are perfectly correlated (they all fail together) and projects in different zones are independent.[10] We assume that for a project to be undertaken with effort, it has to be monitored by a bank. Finally, assume that only projects undertaken with effort can have a positive net present value. Hence, the incentive compatible allocation is the unique implementable allocation. Assume, for simplicity, that there is a given gross

[9] The model shares some similarities with those used in Holmstrom and Tirole (1997) and Ennis (2001).

[10] In equilibrium all projects will be undertaken exerting effort. The underlying assumption on success correlation is that projects undertaken with no effort fail when projects undertaken with effort fail, as well as some other times (so that $1 - p_L > 1 - p_H$). See Holmstrom and Tirole (1997, footnote 8) for details.

Table 5 Notation

$\widetilde{\theta}$	funds owned by investor $\widetilde{\theta}$
R	gross safe interest rate
r_A	return of project r_A when success
p_H (p_L)	prob. of success with (without) effort
I	size of investment projects (in amount of funds needed)
c	cost of monitoring a project
κ	cost of becoming a bank
δ	cost of diversification
ψ	size of the community banks (number of projects monitored)
ψ_D	size of the diversified bank
I_m	bank capital per project
r_P	interest rate on deposits (deposit interest rate)
$\widehat{r}_A$ (r_A^*)	interest rate on bank loans with (without) branching restrictions
$\widehat{\theta}$ (θ^*)	funds owned by the smallest bank with (without) branching rest.
Θ	total amount of monitors' own funds

interest rate R on funds. We can think of R as the return obtained from a safe storage technology.[11]

Assume monitoring is costly and not observable. Let c be the per-project cost of monitoring. The cost c is in utility terms (it does not deplete available funds). Any investor in the economy can choose to become a monitor. For reasons that will become clear below we can call each of these monitors a bank. To acquire the monitoring technology the agent has to incur a cost κ (in utility terms). Given that an agent has incurred the cost κ, the agent can monitor as many projects as desired as long as he or she incurs the cost c per project being monitored. This makes the market for monitoring services perfectly competitive. The monitor can also choose whether to handle projects in one zone or in a large number of zones.[12] Assume that there is an extra operational cost δ of running an institution (bank) handling projects in more than one zone. Then, we need to consider only two possible levels of diversification: the monitor either specializes in projects from one particular zone or becomes completely diversified.

[11] The following restrictions on fundamental parameters are assumed to hold: $2p_L < R < R + c/I < 2p_H$ and $p_H < R + c/I$.

[12] Specifically we assume that there is a continuum of different zones with total measure of one. See Ennis (2001) for details.

Bank Branching Restrictions

Let us consider first the case where each monitor is exogenously restricted to handle projects from a single zone. An agent with a monitoring technology accepts funds from other agents and invests in projects. These external funds available to the monitor can be called deposits. If a bank only handles projects in one zone, then the bank fails with probability $1 - p_H$, the probability that the projects in the zone fail. Let r_p be the return on deposits when the bank does not fail. By an arbitrage condition we have

$$p_H r_p = R.$$

This condition means that the expected rate of return on deposits in a community bank is equal to the safe interest rate.

It is not hard to see that in equilibrium there is a threshold on the productivity of projects, $\widehat{r}_A$, such that only projects with $r_A \geq \widehat{r}_A$ will be undertaken. Consequently, $2 - \widehat{r}_A$ is the total number of projects undertaken. Because there is perfect competition in the market for monitoring services, the project owners only pay $\widehat{r}_A I$ to the bank in return for a loan of size I. For this reason we can call $\widehat{r}_A$ the loan interest rate. Let ψ be the number of projects handled by a bank. The variable ψ is an indicator of the size of the bank. Agents agree to deposit funds in a bank of size ψ only when the following incentive compatibility condition is satisfied

$$-c\psi + p_H \left[\widehat{r}_A I\psi - r_p(I - I_m)\psi\right] \geq p_L \left[\widehat{r}_A I\psi - r_p(I - I_m)\psi\right], \quad (1)$$

where I_m is the amount of own funds the bank commits per project. This condition says that the return to the banker from monitoring the projects must be greater than the return from not monitoring (given that depositors believe that the bank *will* be monitoring). Because monitors want to handle as many projects as possible, condition (1) holds with equality in equilibrium and determines the equilibrium bank capital per project, $\widehat{I}_m$. For this reason, the banker's return per project must satisfy

$$-c + p_H \left[\widehat{r}_A I - r_p(I - \widehat{I}_m)\right] = \frac{p_L c}{p_H - p_L}.$$

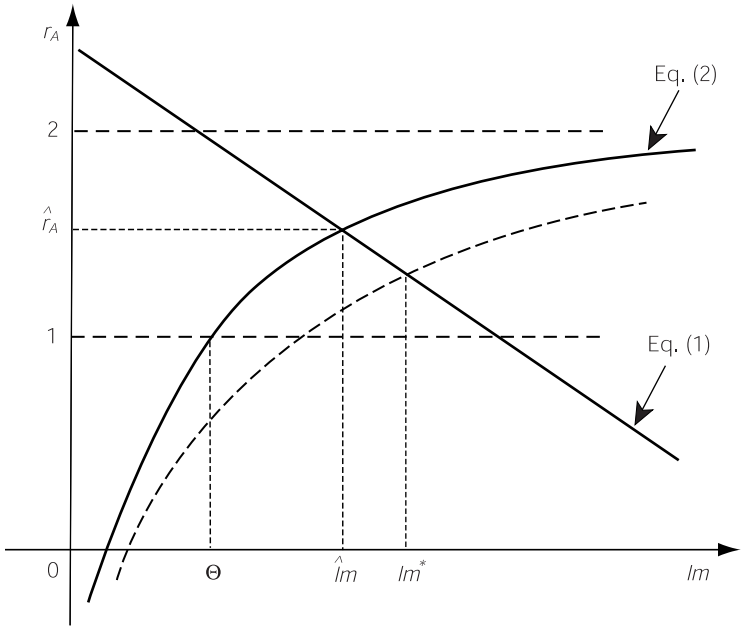
Let Θ be the total amount of own funds committed by monitors in equilibrium. The next paragraph explains how this quantity is determined. Remember that $(2-\widehat{r}_A)$ is the number (measure) of projects undertaken in equilibrium. Then, market clearing for the funds owned by monitors requires that

$$(2 - \widehat{r}_A)\widehat{I}_m = \Theta. \quad (2)$$

This states that the number of projects funded times the amount of the banker's own funds invested per project equals the total amount of banker's funds invested. Given Θ, we can use expressions (1) and (2) to determine the equilibrium values of $\widehat{r}_A$ and $\widehat{I}_m$ (see Figure 3).

Consider now the decision of an investor to become a bank. Note that because of the incentive-compatibility constraint (1) for monitors, the return

Figure 3 Equilibrium Loan Interest Rate



The intersection of the incentive-compatibility constraint for community banks (equation (1)) and the market clearing condition for funds owned by monitors (equation (2)) determine the equilibrium loan interest rate. The dashed locus corresponds to the shift in equation (2) when a diversified monitor is introduced in the model. See Table 5 for notation.

associated with becoming a bank is given by

$$-\kappa + \left(-c + p_H\left[\widehat{r}_A I - r_p(I - \widehat{I}_m)\right]\right)\psi = -\kappa + \frac{p_L c}{p_H - p_L}\psi.$$

As long as the return from becoming a bank is greater than $R\widehat{I}_m\psi$ (the return from safely storing funds), the agent will choose to become a bank. Because the return is increasing with the number of projects monitored, there is a minimum equilibrium size of banks, $\widehat{\psi}$, determined by

$$-\kappa + \frac{p_L c}{p_H - p_L}\widehat{\psi} = R\widehat{I}_m\widehat{\psi}. \tag{3}$$

Since the amount of funds banks commit to each project, $\widehat{I}_m$, is uniform across projects, a particular value of ψ (the size of the bank) is directly associated with a particular value of the wealth of the banker, $\widetilde{\theta}$. This relationship is

given by the following equation

$$\psi = \frac{\widetilde{\theta}}{\widehat{I}_m}. \tag{4}$$

Then, given the value of $\widehat{\psi}$ that solves equation (3), there is a threshold on the amount of funds that an agent has to own in order to become a bank. Call this threshold $\widehat{\theta}$. All agents with $\widetilde{\theta} > \widehat{\theta}$ will become banks, and $1 - \widehat{\theta}$ is the total number of banks in each zone. The total amount of own funds invested by banks in equilibrium is then given by

$$\Theta = \int_{\widehat{\theta}}^{1} \widetilde{\theta} d\widetilde{\theta} = \frac{1}{2}\left(1 - \widehat{\theta}^2\right). \tag{5}$$

Substituting expressions (1) and (5) into equation (2) we obtain what can be thought of as a demand for bank funds, $\widehat{\theta}^d = f^d(I_m)$.[13] Equation (3) implicitly defines a supply of bank funds. By making demand equal supply, we can obtain the equilibrium level of $\widehat{\theta}$ (see Figure 4).

Note that this equilibrium induces a size distribution of banks (monitors) according to equation (4). These banks are all essentially the same type of institution (community banks). The size distribution is a direct consequence of an underlying heterogeneity among bank owners (in terms of own funds) that is exogenous to the model. In what follows we introduce some *endogenous* heterogeneity.

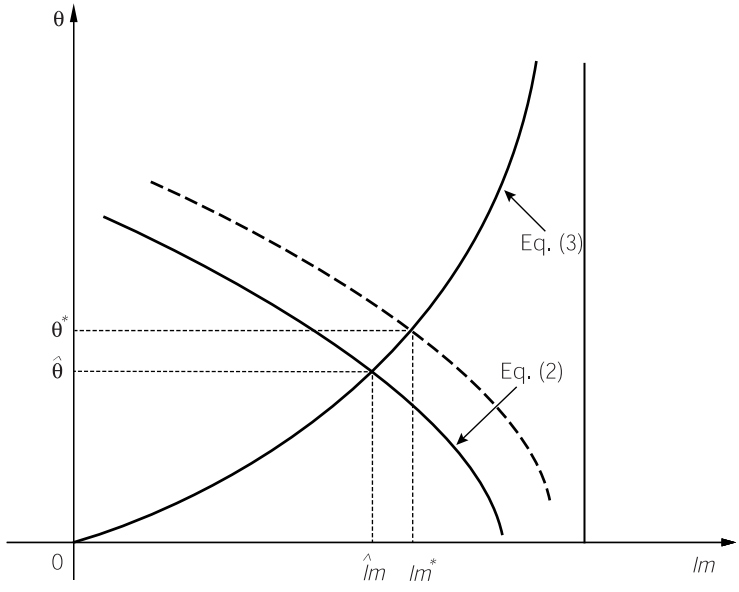
No Bank Branching Restrictions

Consider the case when full diversification is possible, i.e., when we do not restrict banks to handling projects in only one zone. Fully diversified monitors do not face an information problem. The proportion of failed projects ("bad loans") in a monitor's portfolio is observable by third parties, and this proportion reveals the bank's monitoring activities.[14] Anyone can become a well-diversified monitor; no internal funds are needed. However, because there is a fixed cost $\delta + \kappa$ of establishing a diversified bank, economies of scale imply that only one diversified bank will exist in equilibrium. We assume contestability and hence a zero profit condition must hold (see Tirole [1988], 307). Let ψ_D be the number (measure) of projects handled by the

[13] Note that the right-hand side of equation (2) gives us the amount of monitor funds needed to run $(2 - \widehat{r}_A)$ investment projects when $\widehat{I}_m$ monitors funds are needed per project to satisfy the incentive compatibility constraints.

[14] Note that diversification is not originated on risk aversion considerations. All agents are risk neutral in the model. See Diamond (1984) for a related result. In Diamond's model, diversification allows the economy to save on actual monitoring costs. In the model in this article, diversification allows the possibility of running a bank without committing internal funds. No saving of monitoring cost goes on here.

Figure 4 Demand and Supply of Bank Funds



The demand (equation (2)) and supply (equation (3)) of funds owned by monitors determine the number of community banks in the economy, $1-\widehat{\theta}$. The dashed locus represents the shift in equation (2) when a diversified monitor is introduced in the model.

diversified monitor. The zero profit condition is

$$\left[p_H\left(r_A - r_p\right) I - c\right]\psi_D - (\delta + \kappa) = 0. \tag{6}$$

This condition states that the net return per loan in the diversified bank multiplied by the size of the bank (i.e., the total number of loans in the bank) has to equal the fixed cost of setting up the diversified institution. The market clearing condition for monitors' funds is now given by

$$(2 - r_A - \psi_D) I_m = \Theta. \tag{7}$$

That is, the number of projects monitored by community banks multiplied by the amount of bank capital per loan has to equal the total amount of bankers' funds invested. Equations (1), (3), (4), and (5) still hold in equilibrium. For intermediate values of δ a well-diversified bank (monitor) will coexist with the community banks in equilibrium. For notational convenience, I use an asterisk to indicate the value of the variables in the equilibrium with a diversified bank and a hat for the equilibrium with no diversification.

The first important result is that the well-diversified bank is also a large bank, i.e., $\psi_D^* > \psi^*$ (where ψ^* is the size of the smallest community bank).

Table 6 Bank Size and Capital Ratios

Asset Size	1976	1986	1996	2000
$\leq$ 40 million	9.2	9.6	11.7	13.2
$\leq$ 50 billion	7.5	7.6	9.7	9.7
$>$ 50 billion	5.2	5.3	7.9	8.6

To see this, note that if we plug (3) and (6) into equation (1) (holding with equality) we obtain that

$$\frac{\kappa}{\psi^*} = \frac{\delta + \kappa}{\psi_D^*},$$

which implies that $\psi_D^*/\psi^* > 1$. Diversified banks must be larger in order to generate sufficient returns to cover the fixed cost, δ, of lending across different regions.

We turn now to comparing the value of some fundamental variables under the two possible cases: when diversification is ruled out exogenously and when it is not. Think of this comparison as a way to improve our understanding of the long-run implications associated with removing geographic (and possibly other) restrictions on the level of integration in the banking industry.

The second important result is that there are fewer single zone banks when a well-diversified bank is part of the system, i.e., $1 - \theta^* < 1 - \widehat{\theta}$. To see this, note that having $\psi_D > 0$ in equation (7) shifts the demand curve for community bank funds to the right (see Figure 4), increasing the equilibrium threshold to θ^*. It is worth noticing that the banks that are disappearing are the smallest (those for which $\widetilde{\theta} \in [\widehat{\theta}, \theta^*]$). In Figure 4 we can also see that $\widehat{I}_m < I_m^*$. This inequality is the foundation for the following two results.

The third result is that the number of projects undertaken in equilibrium is smaller when there is no diversified bank, i.e., $2 - \widehat{r}_A < 2 - r_A^*$. This result is a direct implication of the fact that equation (1) holds (with equality) in both equilibria and that $\widehat{I}_m < I_m^*$.

Finally, the fourth result is that non-diversified banks tend to become smaller in the equilibrium with a diversified institution. From equation (4), given a value of $\widetilde{\theta}$, a non-diversified bank will hold a smaller number of projects in the equilibrium with the larger I_m, that is, in the equilibrium with the diversified institution.

In terms of the implications for the observed size distribution of banks, using equation (3) we can see that $\psi^* > \widehat{\psi}$, i.e., the smallest community bank is larger when there is a diversified bank. When a diversified bank enters the market, the equilibrium loan interest rate (r_A^*) falls, reducing the profitability of community banks. As a consequence, only larger community banks survive.

Four final remarks seem relevant at this point. First, note that by adjusting the distribution of agents over the level of own funds $\widetilde{\theta}$, one can easily match any given size distribution of community banks. The assumption of a uniform distribution over $\widetilde{\theta}$ is convenient but is in no way necessary. Second, I have assumed that the market for national banks is contestable. This assumption allowed us not to worry about the possibility of monopoly power even though only one national bank exists in equilibrium. Contestability has been challenged on several grounds in the theoretical literature (see Tirole [1988], Chapter 8). The implications of increasing bank-asset concentration on the level of competition in the industry are of major concern to researchers and policymakers. I abstracted from these considerations in the model, but they are probably important and merit further study. Third, note that the model has implications for the amount of bank capital that community and national (diversified) banks would hold in equilibrium. Preliminary analysis of the data shows that, in accordance with the model, small community banks tend to systematically hold higher capital ratios than large national banks (see Table 6). Finally, the size of business firms plays no role in the model presented here. All investment projects are the same size and have the same financing requirements. Empirical studies tend to find that small firms rely more heavily on banks for their financing needs (compared with large firms). The model presented here is too simple to be used to study this last issue. However, below I discuss some complementary theories for which the size of business firms is important.

Other Theoretical Explanations of the Bank Size Distribution

Product Differentiation

It has been well documented that small businesses tend to rely heavily on bank credit (see Bitler, Robb, and Wolken [2001]). Small banks that maintain a long-term relationship with borrowers provide an important share of this credit. For example, Strahan and Weston (1996) document that the market share of small banks in the market for loans to small firms was 35 percent in 1995. This stylized fact can be used as a foundation for a product-differentiation theory of the size distribution of banks.

Banks provide differentiated financial services. For example, a bank could make available standardized loans, for which the approval procedure and the necessary monitoring are systematic and uniform across borrowers, or it could provide customized loan contracts to long-term clients. But, in principle, a single bank could also provide both. Some other factor needs to be introduced to explain the different sizes of banks. One possibility is that there are some technological reasons that make the provision of both types of loans by uniform

size banks inefficient. There are two issues related to this argument that need explaining. First, why are large banks more efficient at providing standardized loans and, second, why are small banks more efficient at relationship lending? The answer to the first question could be in the existence of economies of scope. Usually, standardized loans are more appropriate for large firms because the information required for the loan is more readily available and verifiable. At the same time, large firms tend to demand a wider array of products and services from the bank. In most cases, only large banks can satisfy all those demands efficiently (perhaps as a matter of being able to achieve the optimal scale of production).

The harder question is why large institutions cannot replicate the relationship lending practices of small banks. In fact, Strahan and Weston (1996) find that in 1995 large U.S. banks had a significant participation rate in the market for loans to small businesses (35 percent).[15] Perhaps the question should be rephrased in terms of the difference in bank portfolio shares of small business loans. In 1995, small commercial and industrial loans represented only 3 percent of total assets of large banks as opposed to 9 percent of small banks (see Strahan and Weston [1996]).

One possible explanation for this difference can be found in the combination of two factors: it is harder to monitor lending decisions in large banking organizations, and relationship loans require more discretion by loan officers. As a consequence of these two factors, small banks tend to be more efficient in the provision of this kind of loan. Regardless of the details, what supports this theory is the underlying heterogeneity of business firms. Because there is a size distribution of business firms, there is a size distribution of banks.

This theory, based as it is on a demand for differentiated products, also has implications for the interpretation of the recent changes in the U.S. banking industry. In the long run, a larger share of the market for loans to small firms will probably be held by large banks, but it is also likely that some small banks will continue to exist (due to their relative efficiency in the provision of relationship loans). Finally, it is important to highlight that, according to this theory, the evolution of the size distribution of business firms should directly affect the size distribution of banks. In other words, if technological developments drive the optimal scale of most business firms to become ever larger (Lucas 1978), then the role of small banks in the economy will also tend to decrease with time.

---

[15] In recent years the approach of large banks to small-business lending has experienced important changes. More and more large banks have started to adopt automated underwriting systems based on credit scoring. This allows large banks to make small business loans on a large scale. See Frame, Srinivasan, and Woosley (2001) for an updated account of this new development.

Corporate Governance

Issues in corporate governance of financial institutions are potentially important for explaining the size distribution of banks. Some authors have argued that internal corporate governance tends to be weaker for banks than for other types of corporations (see Prowse [1997]).[16] Here I sketch one theory of bank size that is based on these considerations. The idea is not to provide a definitive explanation of size heterogeneity but to illustrate how weak corporate governance may affect bank-size dispersion.

There are numerous empirical studies documenting that recent bank mergers do not seem to result in large efficiency gains (see, for example, Berger, Demsetz, and Strahan [1999]). The traditional justifications for mergers (for example, economies of scale and scope) have problems accounting for these findings. Some efforts have been made to provide alternative explanations for the tendency of banks to become large. One of these possible explanations is based on imperfect corporate governance and uncertainty about the managerial ability of bank CEOs (see Milbourn, Boot, and Thakor [1999]). This explanation can also provide justification for some of the bank size dispersion observed in the data. In fact, talent heterogeneity among bank CEOs alone could be used to induce a size distribution of banks, as in Lucas (1978). However, the corporate governance story involves information issues that were not present in Lucas's paper.

The main objective of the theory is to explain mergers that do not imply efficiency improvements, which is not so important to the present article; however, the theory's prediction of some size heterogeneity among banks is more germane. Suppose that shareholders do not know how talented the CEO of their bank is, but they would like to better compensate a talented CEO. Since talent is associated with a higher probability of success, the shareholders will use the success rate of the CEO as a proxy for his or her talent. However, not only talented CEOs are successful; some CEOs are just lucky. This brings up a problem: Inferring who is talented is not an easy task. Suppose further that as the bank gets bigger, it becomes harder for the CEO to just get lucky. CEOs who perceive themselves as talented individuals will then tend to prefer to manage large institutions (or make their institutions bigger by completing mergers and acquisitions) because if they eventually become successful, they will more clearly signal their ability and thereby increase their compensation. It can be shown that in this kind of environment, CEOs will tend to generate and manage different sizes of banks according to their perception of their own ability (not known to them with certainty).

[16] For example, government measures regulating bank takeovers, such as the need for prior approval and other potential delays, make the possibility of takeovers a less effective mechanism for disciplining bank managers.

An interesting extension of this theory suggests that there may be a bias towards large organizations in the banking industry. Suppose that the more talented CEOs tend in fact to perceive themselves as more talented (and hence to manage large banks). Suppose also that shareholders have this information and intend to use it in their compensation decisions. Less talented individuals may then choose to manage large institutions just to avoid revealing that they are actually not in the group of talented managers.[17]

3. CONCLUSIONS

This article provides an overview of some empirical and theoretical issues associated with the existence of a nondegenerate size distribution of banks in the United States. I review a number of theories of bank size heterogeneity, and I concentrate on those theories that tend to explain the small-banks phenomenon not as a transitory situation but as the result of an explicit equilibrium choice. This explanation seems to be in accord with the empirical facts described in the first part of the article. The size distribution of banks tends to be relatively more skewed to the right than life-cycle-of-firms theories predict. In other words, the mass of banks is highly concentrated around the range of small asset size. The theories reviewed in this article could help explain this fact.

But it is also true that 50 years of heavy regulation in the banking industry, and branching restrictions in particular, have played a major role in shaping the size distribution of banks in the United States. Deregulation is still very recent, and it may well be that the transition to a new banking industry structure is not over yet. For example, the banking system in Canada, which has never had branching restrictions, has mainly large banks with numerous branches across the country. The question remains, will the U.S. system converge to the Canadian model of banking? One possibility is that the final industry structure will be influenced by initial conditions even after the transition period is over. For example, community banks, having existed for some time, may have generated a demand for their services that will persist. If this is the case, then the market structure of the U.S. banking system and the Canadian system will continue to be different even after their regulatory frameworks have fully converged.

---

[17] Bliss and Rosen (2001) study the relationship between bank mergers and CEOs' compensation in a sample of megamergers that took place between 1986 and 1995. They find significant evidence supporting the hypothesis that asset growth (especially via mergers) tends to increase CEOs' compensation. They also find that this effect tends to motivate acquisition decisions by CEOs. (CEOs with a higher proportion of stock-based compensation tend to be less likely to engage in an acquisition.)

REFERENCES

Berger, Allen N., Anil K. Kashyap, and Joseph M. Scalise. 1995. "The Transformation of the U.S. Banking Industry: What a Long, Strange Trip It's Been." *Brookings Papers on Economic Activity* 2: 55–201.

_____, Rebecca S. Demsetz, and Philip E. Strahan. 1999. "The Consolidation of the Financial Services Industry: Causes, Consequences, and Implications for the Future." *Journal of Banking and Finance*. 23 (February):135–94.

Bitler, Marianne P., Alicia M. Robb, and John D. Wolken. 2001. "Financial Services Used by Small Businesses: Evidence from the 1998 Survey of Small Business Finances." *Federal Reserve Bulletin*. (April): 183–205.

Bliss, Richard T., and Richard J. Rosen. 2001. "CEO Compensation and Bank Mergers." *Journal of Financial Economics*. 61 (July): 107–38.

Broaddus, J. Alfred, Jr. 1998. "The Bank Merger Wave: Causes and Consequences." Federal Reserve Bank of Richmond *Economic Quarterly* 84 (Summer): 1–11.

Cerasi, Vittoria, and Sonja Daltung. 2000. "The Optimal Size of a Bank: Costs and Benefits of Diversification." *European Economic Review* 44 (October): 1701–26.

Diamond, Douglas. 1984. "Financial Intermediation and Delegated Monitoring." *Review of Economic Studies* 51: 393–414.

Ennis, Huberto M. 2001. "Loanable Funds, Monitoring and Banking." *European Finance Review* 5: 79–114.

Ericson, Richard, and Ariel Pakes. 1995. "Markov-Perfect Industry Dynamics: A Framework for Empirical Work," *Review of Economic Studies* 62 (January): 53–82.

Feldman, Ron J., and Arthur J. Rolnick. 1997. "Fixing FDICIA. A Plan to Address the Too-Big-To-Fail Problem." Federal Reserve Bank of Minneapolis *Annual Report*.

Frame, W. Scott, Aruna Srinivasan, and Lynn Woosley. 2001. "The Effect of Credit Scoring on Small-Business Lending." *Journal of Money, Credit and Banking* 33 (August): 813–25.

Holmstrom, Bengt, and Jean Tirole. 1997. "Financial Intermediation, Loanable Funds, and The Real Sector." *Quarterly Journal of Economics* 62 (August): 663–91.

Hopenhayn, Hugo A. 1992. "Entry, Exit, and Firm Dynamics in Long Run Equilibrium." *Econometrica* 60 (September): 1127–50.

Jayaratne, Jith, and Philip Strahan. 1997. "The Benefits of Branching Deregulation" Federal Reserve Bank of New York *Economic Policy Review* (December): 13–29.

Jovanovic, Boyan. 1982. "Selection and the Evolution of Industry," *Econometrica* 50 (May): 649–70.

Lucas, Robert E., Jr. 1978. "On the Size Distribution of Business Firms." *The Bell Journal of Economics* 9 (Autumn): 508–23.

Milbourn, Todd T., Arnoud W. A. Boot, and Anjan V. Thakor. 1999. "Megamergers and Expanded Scope: Theories of Bank Size and Activity Diversity" *Journal of Banking and Finance* 23 (February): 195–214.

Prowse, Stephen. 1997. "Corporate Control in Commercial Banks." *The Journal of Financial Research* 20 (Winter): 509–27.

Simon, Herbert A., and Charles P. Bonini. 1958. "The Size Distribution of Business Firms." *The American Economic Review* 48 (September): 607–17.

Strahan, Philip E., and James Weston. 1996. "Small Business Lending and Bank Consolidation: Is There Cause for Concern?" Federal Reserve Bank of New York *Current Issues in Economics and Finance* 2 (March): 1–6.

Sutton, John. 1997. "Gibrat's Legacy." *Journal of Economic Literature* 35 (March): 40–59.

Tirole, Jean. 1988. *The Theory of Industrial Organization*. Cambridge, Mass.: MIT Press.

A Primer on Optimal Monetary Policy with Staggered Price-Setting

Alexander L. Wolman

If the monetary authority can make a binding promise concerning future monetary policy, what policy should it follow? If the monetary authority cannot make such a promise, how should it behave on a period-by-period basis? How one answers these questions reflects one's beliefs about how the economy works, what monetary policy is feasible, and how outcomes should be evaluated. Economists work with explicit models of the economy and impose explicit welfare criteria in order to answer questions about optimal policy. This approach facilitates reasoned debate. If one disagrees with a policy being advocated, that disagreement necessarily reflects disagreement with the economic model, the welfare criterion, or assumptions about institutional features of policy.

Conditional on one model of how the economy works that is currently popular and is based on optimizing behavior by households and firms, I will discuss three different notions of optimal monetary policy.[1] Because economists have not reached a consensus about the appropriate model of the macroeconomy to be used for monetary policy analysis, this article cannot provide definitive answers to questions about optimal policy.

The distinctive feature of the model is that firms do not continuously adjust the prices of goods they sell. Instead, the price of any individual good changes only periodically. Such price stickiness is observed for many goods (see the

The author thanks Mike Dotsey, Andreas Hornstein, Tom Humphrey, Bennett McCallum, and Pierre Sarte for helpful comments and discussions. This article does not necessarily represent the views of the Federal Reserve Bank of Richmond or any branch of the Federal Reserve System.

[1] Parts of this article summarize research described in more detail in King and Wolman (1996 and 1999), and parts can serve as background for Khan, King, and Wolman (2000 and 2001). Goodfriend and King (1999 and 2000) contains much complementary discussion of monetary policy in the type of model used in this article.

survey in Wolman [2000]) and forms an important channel through which monetary policy can affect aggregate economic outcomes.

As in the work of Kydland and Prescott (1977) and Barro and Gordon (1983), policy surprises have the potential to improve welfare under certain conditions; there is a time-consistency problem for monetary policy. In contrast to that work, however, here the current-period policy problem is affected by the nature of future policy because the model involves multiperiod pricing.

Unlike much work on optimal monetary policy, the model in this article does not generate welfare losses associated with the area under the money demand curve; that is, there are no shoe leather costs of inflation.[2] This modeling choice allows us to focus on issues related to staggered price-setting. The same approaches described here can easily be applied to models in which there are both staggered price-setting and shoe leather costs of inflation. Khan, King, and Wolman (2000) contains such an application.

The model delivers two important results regarding optimal policy. First, it is beneficial for the monetary authority to be able to make binding promises about its future behavior. As in Kydland and Prescott (1977) and Barro and Gordon (1983), if binding promises are not feasible, we should expect a relatively high inflation rate. Under current arrangements, it is not possible for the Federal Reserve to make binding promises. If it were possible, the model prescribes that in the long run, inflation should either be zero or just slightly positive, depending on whether the policy objective is present value welfare or steady state welfare.

1. THE MODEL ECONOMY

As is standard in modern economic models, the fundamental assumptions of this model concern who the agents are, their preferences and endowments, the technology to which they have access, and the market structure. Here there are three types of agents: households, firms, and the government. I will give an overview of the model before describing it in mathematical terms.

There is a large number of identical households. Households are assumed to live forever and to obtain utility from leisure time and from consuming goods produced by firms. Consumption and leisure in the present are preferred to the same amount of consumption and leisure in the future. Households' endowments consist of one unit of time in each period and ownership of the firms. Time is allocated between labor supply to firms—in exchange for wage income—and leisure. Households also demand government-supplied money in an amount identical to their nominal consumption spending.

[2] Shoe leather costs of inflation are the resource costs individuals and firms incur in order to economize on holdings of currency, which does not bear interest. See Bailey (1956) and Friedman (1969).

Firms produce consumption goods using a technology that relies only on labor. Their objective is to maximize the present value of profits distributed to households, and to achieve this objective they set prices and hire labor. It is also assumed that firms may only adjust their prices every two periods (I assume that a period is three months). This shortcut imposes price stickiness of a magnitude arguably lower than that observed in the United States. As for market structure, in this model there is a large number of firms producing under conditions of monopolistic competition. Each firm produces a distinct good and faces a constant elasticity demand curve, with the elasticity common across firms.

The government's sole role is to supply money. Money enters the economy in the form of lump-sum transfers from the government to consumers.

Households

Households' preferences for consumption (c_t) and leisure (l_t) in the present and future are given by a concave utility function $u\,(c, l)$ and a discount factor $\beta < 1$:

$$\sum_{t=0}^{\infty} \beta^t u\,(c_t, l_t)\,, \tag{1}$$

where the subscript t indexes time. For the examples below, I will specify the utility function to be

$$u\,(c, l) = \ln c + \chi \cdot l, \tag{2}$$

where $\chi > 0$ is a fixed parameter. The household's total consumption is an index of consumption of a unit measure of different goods. In keeping with the market structure and pricing behavior described below, we need to keep track of only two types of goods, each with measure 1/2: those with prices set in the current period, indexed by 0, and those with prices set in the previous period, indexed by 1:

$$c_t = c\left(c_{0,t}, c_{1,t}\right) \equiv \left(\frac{1}{2} \cdot c_{0,t}^{\frac{\varepsilon-1}{\varepsilon}} + \frac{1}{2} \cdot c_{1,t}^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{\varepsilon}{\varepsilon-1}}. \tag{3}$$

This consumption index implies that households have constant elasticity demands for each individual good:

$$c_{j,t} = \left(\frac{P_{j,t}}{P_t}\right)^{-\varepsilon} c_t, \;\; j = 0, 1, \tag{4}$$

where $P_{j,t}$ is the nominal price of a good with price set in period $t - j$, and P_t is the price index, given by

$$P_t = \left(\frac{1}{2} P_{0,t}^{1-\varepsilon} + \frac{1}{2} P_{1,t}^{1-\varepsilon}\right)^{1/(1-\varepsilon)}. \tag{5}$$

For derivations of (4) and (5), see the appendix to Wolman (1999).

The household's budget constraint requires that consumption spending not exceed the sum of wage income and firms' profits (D_t):

$$P_t c_t \leq P_t w_t n_t + D_t, \tag{6}$$

where w_t is the real wage. The household's time constraint requires that its labor supply and leisure time not exceed its endowment of one unit of time:

$$l_t + n_t \leq 1, \tag{7}$$

where total labor supply (n_t) is the sum of labor supplied to the two types of firms,

$$n_t = \frac{1}{2} \cdot \left(n_{0,t} + n_{1,t}\right). \tag{8}$$

The sum is multiplied by 1/2 because $n_{j,t}$ is labor hired per j-type firm, and half the firms are of each type. Households also demand government-supplied money in an amount identical to their nominal consumption spending:

$$M_t = P_t c_t, \tag{9}$$

where M_t is nominal money balances.

The key equation from the household's side of the model is optimal labor supply, which can be determined by maximizing $u\left(c_t, l_t\right)$ subject to the budget constraint and time constraint:

$$w_t \cdot u_c\left(c_t, l_t\right) = u_l\left(c_t, l_t\right), \tag{10}$$

and, for the preferences in (2),

$$\frac{w_t}{c_t} = \chi. \tag{11}$$

If the household were to marginally increase the quantity of labor it supplies, the utility-denominated value of additional labor income would be exactly offset by the utility loss associated with the decrease in leisure time.

Firms

There is a continuum of monopolistically competitive firms, each producing a unique differentiated product. As mentioned above, firms set their price for two periods—which implies $P_{1,t} = P_{0,t-1}$—and half of the firms adjust their price in any given period. This pattern is known as staggered price-setting. Each firm has the same production technology:

$$c_{j,t} = n_{j,t}, \;\; j = 0, 1. \tag{12}$$

When firms adjust, they choose a price that will maximize their present discounted profits over the two periods for which the price is fixed. As owners of the firms, consumers instruct firms to value current and future profits using

the marginal utility of income, which is $u_c(c_t, l_t)$ for the current period and $u_c(c_{t+1}, l_{t+1})$ for the next period.

The optimal price satisfies

$$\frac{P_{0,t}}{P_t} = \left(\frac{\varepsilon}{\varepsilon - 1}\right) \cdot \left(\frac{u_c(c_t, l_t)\, c_t w_t + \beta u_c(c_{t+1}, l_{t+1})\, c_{t+1} w_{t+1} \pi_{t+1}^{\varepsilon}}{u_c(c_t, l_t)\, c_t + \beta u_c(c_{t+1}, l_{t+1})\, c_{t+1} \pi_{t+1}^{\varepsilon-1}}\right), \quad (13)$$

where the inflation rate in period $t+1$ is given by $\pi_{t+1} \equiv P_{t+1}/P_t$. This expression makes the marginal present discounted value of profits zero for an adjusting firm.[3] Firms would like to charge a price that is a constant markup of $\varepsilon/(\varepsilon - 1)$ over marginal cost, and in this model marginal cost is equal to the real wage. Because the price level may change over time and prices are set for two periods, it is generally impossible for a firm to achieve this ideal markup in every period. Firms do the best they can, which typically means setting a markup higher than $\varepsilon/(\varepsilon - 1)$ in periods when they adjust. If they were to choose the ideal markup, they would maximize profits in the first period of each price cycle, but in the next period profits would be low because inflation would erode the firm's real price. The profit function is concave, so it is optimal to sacrifice some profits in both periods rather than maximizing profits in one period. Another way to view the optimal price is that it achieves the ideal markup with respect to a quasi-weighted average of marginal cost in the two periods of the price cycle.

Equilibrium

In equilibrium, the wage rate and the prices of individual consumption goods are such that households maximize present discounted utility with their labor supply and consumption demand decisions, and firms maximize present discounted profits with their price-setting decisions. In addition, equilibrium requires a specification of monetary policy.

We can use many of the equations stated above to eliminate variables so that a complete description of equilibrium reduces to as few as two difference equations in the two variables $c_{0,t}$ and $c_{1,t}$. One of those equations is firms' optimal pricing equation (13), and the others are determined by monetary policy. The appendix shows how the optimal pricing equation can be expressed in terms of a relationship between $c_{0,t}$, $c_{1,t}$, $c_{0,t+1}$, and $c_{1,t+1}$. That relationship can be summarized by a function $x()$ which is defined in the appendix:

$$0 = x\left(c_{0,t}, c_{1,t}\right) + \beta x\left(c_{1,t+1}, c_{0,t+1}\right). \quad (14)$$

[3] A detailed derivation is provided in Wolman (1999). See also Yun (1996). Note that in Wolman (1999), the factor Δ discounts *nominal* profits, so that the different exponents on inflation in that article are offset by a different discount factor (i.e., the expressions are equivalent).

The sum $x\left(c_{0,t}, c_{1,t}\right) + \beta x\left(c_{1,t+1}, c_{0,t+1}\right)$ is the present discounted marginal profit of a firm choosing its price in period t, and optimal pricing behavior requires present discounted marginal profit to be zero. In (13) I solved this equation for the optimal price $P_{0,t}/P_t$, but it will be easier to work with in the form shown in (14). The function $x\,(a, b)$ is the equilibrium value of current marginal profit (with respect to price) for a firm that sells quantity a, when prices and demands are such that half of all firms sell quantity a and the other half sell quantity b. Specifically, $x\left(c_{0,t}, c_{1,t}\right)$ is the marginal profit in period t of a firm setting its price in period t; that firm sells quantity $c_{0,t}$, as do all other firms setting their price, while firms charging a price set in the previous period sell quantity $c_{1,t}$. The form of the other equation(s) needed to characterize equilibrium is determined by monetary policy.

Two Distortions to Summarize Outcomes

Consumption and leisure are the fundamental variables that households in the model care about. Thus, good monetary policy makes consumption and leisure behave in ways that households like. It is important to stress that in this model the central bank does not directly control consumption and leisure. Nonetheless, the central bank's choices regarding the money supply affect some relative prices in the model—in particular, the relative prices of goods set in the current and previous periods—and these relative prices in turn affect consumption and leisure.

By focusing on households' preferences and the technology for producing goods, we see there is an optimal allocation that serves as a useful benchmark for policy. Analyzing this outcome leads us to define two measures of distortions, summarizing deviations from the optimal allocation. The relative price distortion effectively makes the economy operate inside its production possibility frontier. The markup distortion acts as a tax on labor input, placing the economy at an inefficient point on the production possibility frontier. If the two distortions are eliminated, the optimal allocation is attained. The monetary policy problem can then be thought of as minimizing these distortions.

The optimal allocation is referred to as the first-best. To find the first-best, maximize utility subject to the constraints imposed by factor endowments and the production technology:

$$\max_{c,l,c_0,c_1,n_0,n_1} u\,(c, l)\,,$$

subject to (3), (7), (8), and (12). Note that this problem is entirely static; the only dynamic element of the model involves price-setting, and the first-best overcomes price stickiness. The six first-order conditions to this problem make it clear that $c_0 = c_1 = c$ and $n_0 = n_1 = n$. The solution for consumption and leisure is a constant pair, which I will denote c^{fb}, l^{fb}, implicitly given by

the following equations:

$$\begin{aligned} u_c\left(c^{fb}, 1-c^{fb}\right) &= u_l\left(c^{fb}, 1-c^{fb}\right), \\ l^{fb} &= 1-c^{fb}. \end{aligned} \tag{15}$$

In the first-best, the technology for producing aggregate output is identical to the technologies for producing individual outputs (12). The marginal product of aggregate labor in the production of aggregate output is thus 1.0. In addition, the marginal rate of substitution between consumption and labor (u_l/u_c) equals the marginal product of labor. This condition is reflected in (15).

Referring back to the full staggered price-setting model, there are two ways in which outcomes can deviate from the first-best. First, the implicit technology for producing aggregate output may not be identical to the technologies for producing individual outputs. That is, $c_t < n_t$ because $c_0 \neq c_1$. In addition, the marginal rate of substitution may not be equated to the marginal product of labor. That is, $u_c\left(c^{fb}, 1-c^{fb}\right) \neq u_l\left(c^{fb}, 1-c^{fb}\right)$, or equivalently $w_t \neq 1$.

Deviations from the equality of consumption and labor input will be denoted by ρ_t and referred to as the relative price distortion:

$$\rho_t \equiv \frac{n_t}{c_t} = \frac{1-l_t}{c_t}. \tag{16}$$

Deviation of the real wage from unity will be denoted by μ_t and referred to as the markup of price over marginal cost:

$$\mu_t \equiv \frac{1}{w_t} = \frac{u_c\left(c_t, l_t\right)}{u_l\left(c_t, l_t\right)}. \tag{17}$$

Because the real wage is equal to real marginal cost in this model, (17) shows that the markup is simply the inverse of real marginal cost. These two equations allow us to move between working in terms of consumption and leisure and working in terms of the relative price distortion and the markup. In other words, given consumption and leisure, (16) and (17) allow us to compute the two distortions, and given the two distortions, the same two equations allow us to compute consumption and leisure implicitly.

To see why the relative price distortion is so named, replace the denominator of (16) with the consumption aggregator, replace the numerator with $\frac{1}{2}\left(n_{0,t}+n_{1,t}\right) = \frac{1}{2}\left(c_{0,t}+c_{1,t}\right)$, and divide numerator and denominator by $c_{1,t}$:

$$\rho_t = \left(\frac{1}{2}\right)^{\frac{-1}{\varepsilon-1}} \frac{\left(\frac{c_{0,t}}{c_{1,t}}+1\right)}{\left(\left(\frac{c_{0,t}}{c_{1,t}}\right)^{\frac{\varepsilon-1}{\varepsilon}}+1\right)^{\frac{\varepsilon}{\varepsilon-1}}}.$$

Finally, replace the consumption ratios with price ratios, using (4):

$$\rho_t = \left(\frac{1}{2}\right)^{\frac{-1}{\varepsilon-1}} \left(\left(\frac{P_{0,t}}{P_{1,t}}\right)^{-\varepsilon} + 1\right)\left(\left(\frac{P_{0,t}}{P_{1,t}}\right)^{1-\varepsilon} + 1\right)^{\frac{-\varepsilon}{\varepsilon-1}}. \qquad (18)$$

From (18), if prices of both types of good are the same ($P_{0,t} = P_{1,t}$), the relative price distortion is eliminated, meaning $c_t = n_t$. Furthermore, knowing the ratio $\frac{c_{0,t}}{c_{1,t}}$ (or $\frac{P_{0,t}}{P_{1,t}}$) is sufficient for determining the relative price distortion.

In a flexible price model with perfect competition, the markup and the relative price distortion would both equal unity ($\rho_t = \mu_t = 1$). With flexible prices but the monopolistic competition structure of the model, the relative price distortion would still be unity, but the markup would exceed unity ($\mu_t = \frac{\varepsilon}{\varepsilon-1} > 1$): with monopoly power, firms set prices above marginal cost. With staggered price-setting as well as monopolistic competition, the relative price distortion and the markup generally exceed unity. But with staggered price-setting, monetary policy can affect these distortions. Policy's leverage over the relative price distortion is straightforward: the more variability in the price level, the greater the relative price distortion.

There are two components to policy's leverage over the markup. First, the markup is affected by the level of inflation even if the inflation rate is constant; this relationship will be detailed in Section 3. Second, with some prices predetermined, policy surprises will affect the level of real activity and the markup. This second mechanism is conventional, yet quite complicated. Suppose that instead of the pricing structure in our model, all prices were reset every period, but in every period they were chosen before any other information was revealed. Then it would be obvious that surprise increases in money raise output, simply from the money demand equation (9). The price level would be treated as fixed, so increases in money would correspond to increases in consumption. In our model, however, only some prices are predetermined. So, again referring to the money demand equation, whether a surprise monetary expansion results in an expansion in output depends on how the firms that set their price in the current period respond to the expansion. And, because those firms set their price for two periods, their response depends on their expectations about the behavior of price setters and monetary policy in the next period. Extending this line of reasoning, one can see that the entire path of future policy matters.

2. DIFFERENT APPROACHES TO OPTIMAL POLICY

Even for a particular explicit model of the private macroeconomy, there is not just one reasonable notion of optimal monetary policy. The reasons for this ambiguity were suggested at the outset; they involve disagreements about the appropriate welfare criterion and about the nature of policy institutions. A welfare criterion is used to rank different policies, and the rankings will

generally differ according to the welfare criterion. Institutions matter because they affect the range of feasible policies.

I will discuss three common notions of optimal monetary policy as they apply to the staggered pricing model. The first and second differ according to whether the welfare criterion is steady state welfare or present value welfare, whereas the second and third differ according to an institutional assumption that changes the set of feasible policies. In each case, I ignore shocks to the economy, so the only changes that occur over time are due to monetary policy.

Perhaps the simplest notion of optimal policy is the idea of an optimal steady state inflation rate: view the monetary authority as choosing the inflation rate, and ask what constant inflation rate is best. With a constant inflation rate, outcomes in each period are identical. Because there is no uncertainty in the model, and there are no fundamental state variables that affect the set of feasible outcomes, it seems sensible to require that the central bank pick a constant inflation rate. However, it would also be interesting to check whether in fact the policymaker would choose a constant inflation rate. Thus, a second natural approach to optimal policy involves assuming that the policymaker has the same welfare criterion as households. One can then ask how inflation would behave over time: would it be constant, and would it end up at the optimal steady state inflation rate?

Both of these approaches to optimal policy maintain that the policymaker (central bank) can credibly promise (commit to) how it will behave in current and future periods. This is not a trivial assumption. For instance, the Fed does not make explicit, detailed promises as to how it will behave in the future. Factors such as reputation may make for implicit commitments in the Fed's behavior, but these factors are probably not strong enough to make either of the first two approaches to optimal policy practically relevant in the current institutional environment. I therefore consider a third notion of optimal policy, which maintains the assumption that the policymaker's welfare criterion is identical to the representative agent's, and further assumes that policy cannot commit in any way to future actions. Each period, the central bank acts in the best interests of society. The central bank cannot dictate what its behavior will be in the next period, but it foresees what form that behavior will take.

These three approaches to optimal policy do not have the same implications. In the first two cases, the results differ in an interesting qualitative dimension, although quantitatively they are quite close. When the welfare criterion is steady state welfare, the central bank will choose a small but positive level of inflation, as this reduces the markup distortion. In contrast, when the central bank is allowed to behave in a time-varying manner and still can commit to future policies, optimal policy leads in the long run to zero inflation. This pair of results has an analogue in the golden rule versus modified golden rule idea in growth theory, which is explained in Section 4.

If, as in the third case, the central bank cannot commit, the nature of optimal policy changes more dramatically. The steady state equilibrium in this case involves rather high inflation. Loosely speaking, because the central bank in the current period cannot affect the behavior of the next period's central bank, it takes what it can get in the current period. The central bank chooses to exploit firms that set their price in the past, as a way of trying to drive down the markup.

3. OPTIMAL POLICY I: STEADY STATE WELFARE

Our first notion of optimal policy has as its criterion the steady state level of welfare. Examining steady state welfare as a function of inflation answers the question, What average level of inflation should a central bank target? This is a sensible policy question, though not the most obvious one to ask in the context of models like the one used here. I will return to this point in Section 4. To determine the optimal steady state inflation rate, I will first derive steady state allocations for an arbitrary inflation rate.

In a steady state equilibrium, all real variables are constant. Nominal prices grow at a constant rate, the steady state inflation rate. If π is the steady state inflation rate, we find the steady state values of all variables by eliminating time subscripts on the real variables in the equilibrium conditions and setting $\pi_t = \pi$. From the consumption aggregator, the demand function, and the optimal pricing equation ((3), (4), and (13)), we can derive an expression for the real wage in steady state:

$$w^{ss} = \left(\frac{\varepsilon - 1}{\varepsilon}\right) \cdot \left(\frac{1}{2} + \frac{1}{2}\pi^{\varepsilon-1}\right)^{1/(\varepsilon-1)} \cdot \left(\frac{1 + \beta\pi^{\varepsilon-1}}{1 + \beta\pi^{\varepsilon}}\right). \tag{19}$$

In addition, from (4) for $j = 0$ and 1,

$$c_0^{ss}/c_1^{ss} = \pi^{-\varepsilon}. \tag{20}$$

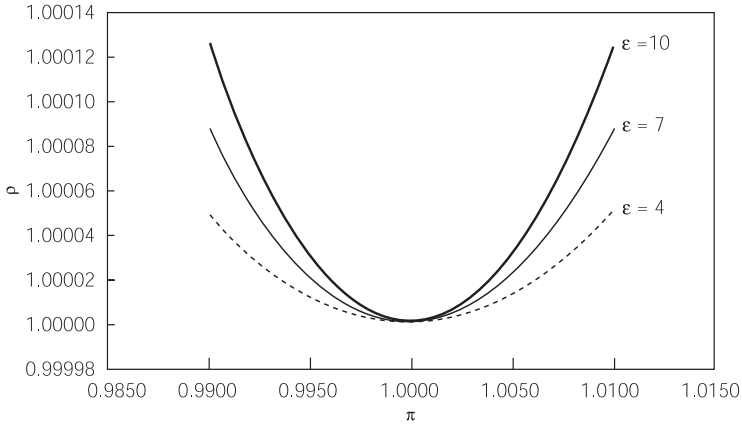
The level of consumption is implicitly determined by substituting the previous two expressions into the consumer's labor supply equation (28):

$$\frac{u_l\left(c_1^{ss} \cdot \pi^{-\varepsilon}, c_1^{ss}\right)}{u_c\left(c_1^{ss} \cdot \pi^{-\varepsilon}, c_1^{ss}\right)} = \left(\frac{\varepsilon - 1}{\varepsilon}\right) \cdot \left(\frac{1}{2} + \frac{1}{2}\pi^{\varepsilon-1}\right)^{1/(\varepsilon-1)} \left(\frac{1 + \beta\pi^{\varepsilon-1}}{1 + \beta\pi^{\varepsilon}}\right). \tag{21}$$

For a given specification of preferences ($u(c, l)$, which determines the functions u_c and u_l), (20) and (21) allow us to compute c_0^{ss}, c_1^{ss} and thus c^{ss}, l^{ss} and the two distortions as functions of the steady state inflation rate. Imposing $u(c, l) = \ln(c) + \chi l$, the steady state levels of consumption and leisure are

$$c^{ss}(\pi) = \left(\frac{\varepsilon - 1}{\varepsilon\chi}\right) \cdot \left(\frac{1}{2} + \frac{1}{2}\pi^{\varepsilon-1}\right)^{1/(\varepsilon-1)} \left(\frac{1 + \beta\pi^{\varepsilon-1}}{1 + \beta\pi^{\varepsilon}}\right)$$

Figure 1 Steady State Relative Price Distortion as a Function of Inflation



The variable π is the gross quarterly inflation rate.

and

$$l^{ss}(\pi) = 1 - \frac{1}{2} \cdot c^{ss}(\pi) \cdot \left(\frac{1}{2} + \frac{1}{2}\pi^{\varepsilon-1} \right)^{\frac{-\varepsilon}{\varepsilon-1}} \cdot (1 + \pi^{\varepsilon}).$$

The steady state relative price distortion and markup for this example are, respectively,

$$\rho^{ss} = \left(\frac{1}{2} \right)^{\frac{-1}{\varepsilon-1}} \frac{\left(\pi^{-\varepsilon} + 1 \right)}{\left(\pi^{1-\varepsilon} + 1 \right)^{\frac{\varepsilon}{\varepsilon-1}}} \tag{22}$$

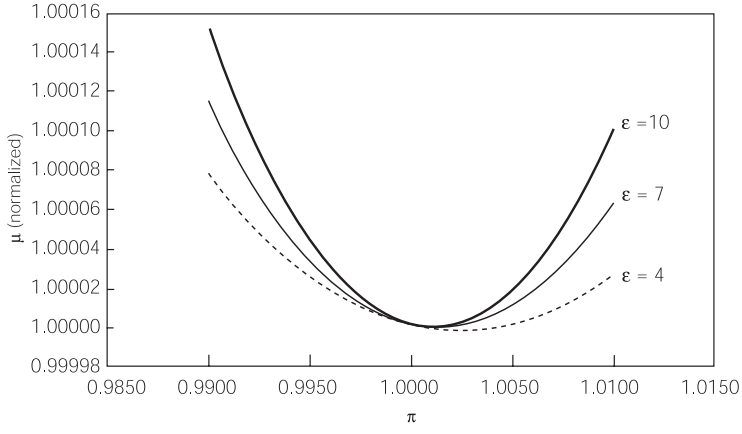
and

$$\mu^{ss} = \left(\frac{\varepsilon}{\varepsilon - 1} \right) \cdot \left(\frac{1}{2} + \frac{1}{2}\pi^{\varepsilon-1} \right)^{1/(1-\varepsilon)} \cdot \left(\frac{1 + \beta\pi^{\varepsilon}}{1 + \beta\pi^{\varepsilon-1}} \right). \tag{23}$$

Figures 1 and 2 plot the steady state markup and relative price distortions as functions of the steady state inflation rate. The relative price distortion is minimized—and in fact eliminated—at zero inflation, whereas the markup is minimized but not eliminated at a very low positive inflation rate. Zero inflation eliminates the relative price distortion because it results in all nominal prices being constant and equal. It is less obvious why a low positive inflation rate should minimize the markup.

To understand the relationship between steady state inflation and the markup, it is helpful to begin by using the price level definition (5) to write

Figure 2 Steady State Markup Distortion as a Function of Inflation, Normalized to $\mu(1.0) = 1.0$



The variable π is the gross quarterly inflation rate.

the markup as

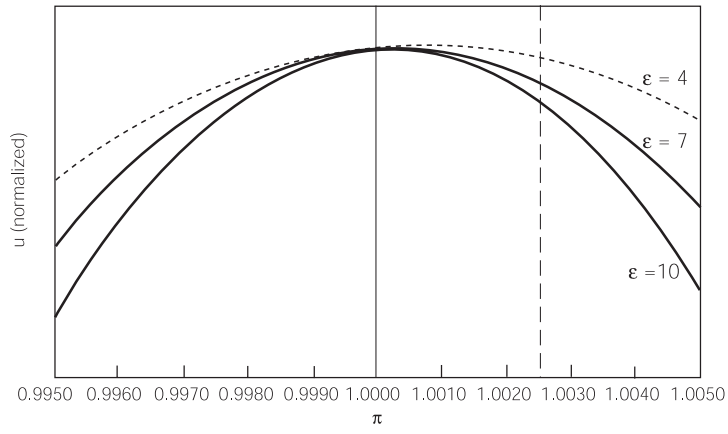
$$\mu^{ss} = \left(\frac{1}{2} \left(\frac{P_0}{MC} \right)^{1-\varepsilon} + \frac{1}{2} \left(\frac{P_1}{MC} \right)^{1-\varepsilon} \right)^{\frac{1}{1-\varepsilon}}, \tag{24}$$

where MC is nominal marginal cost. The effect of inflation on the markup depends on the effect of inflation on the markups charged by the two types of firms. In steady state, the markup of type one firms is simply the markup charged by type zero firms divided by the inflation rate:

$$\frac{P_1}{MC} = \frac{P_0}{MC} \cdot \frac{1}{\pi},$$

because $P_1 = P_0/\pi$. Using (23) and (24) it is easy to show that the markup of type zero firms is increasing in the inflation rate (except at high rates of deflation). This is also intuitive: to protect themselves from the real price erosion caused by inflation in the next period, firms set a higher markup when they do adjust. Thus, whether the aggregate markup increases in inflation depends on whether the effect of inflation on P_0/MC is strong enough to overwhelm the erosion effect of inflation on P_1/MC. It turns out that for very low inflation, the erosion effect dominates, so the aggregate markup falls with inflation. But for even moderately high inflation, the effect of inflation on adjusting firms' prices is strong enough that the aggregate markup rises.

Figure 3 Steady State Welfare as a Function of Inflation



The variable π is the gross quarterly inflation rate.

Returning now to optimal policy, the optimal steady state problem is

$$\max_{\pi} u\left(c^{ss}(\pi), l^{ss}(\pi)\right).$$

Using the expressions for steady state consumption and leisure, we could analytically characterize the optimal steady state inflation rate. It is more revealing, though, to look at a picture. Just as Figures 1 and 2 plot the markup and relative price distortions as functions of the steady state inflation rate, Figure 3 plots steady state welfare as a function of inflation. The vertical dashed line indicates the inflation rate that minimizes the markup (for $\varepsilon = 4$), and the vertical solid line indicates the inflation rate that minimizes the relative price distortion (zero inflation). As we should expect, the welfare maximizing inflation rate is between the two that each minimize a distortion. It is notable that the optimal steady state inflation rate is positive in this model, but it is only slightly positive—at approximately 0.4 percent per year—for $\varepsilon = 4$.

4. OPTIMAL POLICY II: PRESENT VALUE WELFARE WHEN THE POLICYMAKER CAN KEEP ITS PROMISES

Steady state welfare is an intuitively appealing welfare criterion, but it is not the only natural choice for our model. Present discounted welfare, defined by (1), is another natural welfare criterion. I will henceforth refer to this simply as *optimal policy*. If there is an initial period to the policy problem, then optimal policy may not involve choosing a steady state—if it did, we would

get the same answer as that in the previous section. In fact, under present value welfare maximization, inflation does vary over time, and it converges to a rate in which steady state welfare is lower than was found in Section 3 in the best steady state (King and Wolman [1999] first presented this result). Period welfare during the transition to steady state is higher than it is in the best steady state. Note that the optimal policy problem is viewed as being solved once and for all in an initial period, and initial periods are more important than later periods precisely because the future is discounted.

To find the inflation behavior that maximizes welfare, I use what is known in optimal taxation problems as the primal approach. Under the primal approach, instead of searching for the policy behavior that maximizes welfare, one first searches for the optimal allocations that are feasible for the policymaker and then (as a secondary step not provided here) determines a rule that would achieve those allocations.

Recall the discussion of equilibrium above. There I showed that an equilibrium was described by at least two difference equations: one representing optimal price-setting and the other(s) depending on policy. Here the additional difference equation(s) must be generated by optimality conditions for policy. To find those optimality conditions, write down the policy problem as maximizing present discounted utility,

$$\sum_{t=0}^{\infty} \beta^t u\left(c\left(c_{0,t}, c_{1,t}\right), l\left(c_{0,t}, c_{1,t}\right)\right),$$

subject to the optimal price-setting condition being satisfied in each period,

$$0 = x\left(c_{0,t}, c_{1,t}\right) + \beta x\left(c_{1,t+1}, c_{0,t+1}\right), \ t = 0, 1, \ldots$$

by choice of sequences for $c_{0,t}$ and $c_{1,t}$. The Lagrangian for this problem is

$$\begin{aligned} L \ = \ & \sum_{t=0}^{\infty} \beta^t u\left(c\left(c_{0,t}, c_{1,t}\right), l\left(c_{0,t}, c_{1,t}\right)\right) \\ & + \sum_{t=0}^{\infty} \beta^t \phi_t \left[x\left(c_{0,t}, c_{1,t}\right) + \beta x\left(c_{1,t+1}, c_{0,t+1}\right)\right]. \end{aligned}$$

The first order conditions are as follows: for $c_{0,t}$ when $t = 0$:

$$u_c\left(\right)\left(\frac{c_{0,t}}{c_t}\right)^{-\frac{1}{\varepsilon}} - \frac{1}{2} u_l\left(\right) + \phi_t x_1\left(c_{0,t}, c_{1,t}\right) = 0;$$

for $c_{0,t}$ when $t = 1, 2, \ldots$:

$$u_c\left(\right)\left(\frac{c_{0,t}}{c_t}\right)^{-\frac{1}{\varepsilon}} - \frac{1}{2} u_l\left(\right) + \phi_t x_1\left(c_{0,t}, c_{1,t}\right) + \phi_{t-1} x_2\left(c_{1,t}, c_{0,t}\right) = 0;$$

for $c_{1,t}$ when $t = 0$:

$$c_{1,t} : u_c\,() \left(\frac{c_{1,t}}{c_t}\right)^{-\frac{1}{\varepsilon}} - \frac{1}{2} u_l\,() + \phi_t x_2\left(c_{0,t}, c_{1,t}\right) = 0;$$

and for $c_{1,t}$ when $t = 1, 2, \ldots$:

$$u_c\,() \left(\frac{c_{1,t}}{c_t}\right)^{-\frac{1}{\varepsilon}} - \frac{1}{2} u_l\,() + \phi_t x_2\left(c_{0,t}, c_{1,t}\right) + \phi_{t-1} x_1\left(c_{1,t}, c_{0,t}\right) = 0.$$

In these expressions, $u_c\,()$ ($u_l\,()$) refers to the partial derivative of the utility function with respect to consumption (leisure), and $x_j\,()$ refers to the partial derivative of the function $x\,()$ with respect to its j^{th} argument. The first-order conditions for c_0 and c_1 in period zero have a different form than the first-order conditions in all later periods. This reflects the fact that in period zero there is no previous policy commitment to be honored, so policy takes advantage of preset prices to expand output.[4] After period zero, the first-order conditions for $c_{0,t}$ and $c_{1,t}$ contain a term that involves the optimal pricing condition for the *previous* period $(t - 1)$.

It might seem odd that the optimality conditions for some period $t > 0$ should take into account a response of firms in the previous period. The explanation is that policy is determined in period zero for all subsequent periods. Firms choosing their price in period $t - 1$ act not only in response to current variables, but also in anticipation of period t variables, and the period zero policymaker takes this effect into account when choosing period t variables. If prices were set for more than two periods, policy with regard to period t variables would affect behavior more than one period in advance, and there would be more than one initial period for which the optimality conditions differed from their eventual form.

The fact that the first-order conditions for optimal policy take on a different form in period zero than in all other periods is indicative of a time-consistency problem: the optimal behavior to which the policymaker committed in period zero would not be maintained if the policymaker were allowed to reoptimize in a later period. That later period would effectively become a new "period zero," differing from all subsequent periods. But if reoptimization were believed to be a possibility, the policy problem would not make sense as written above. Firms would not believe that a binding policy commitment had been made, and the policymaker would be unable to determine anything but the current period outcome. I return to the lack-of-commitment scenario in Section 6. For now I maintain the assumption that the policymaker can credibly commit to his or her behavior into the infinite future.

[4] For a more detailed discussion of this type of optimization problem, see Kydland and Prescott (1980) and Marcet and Marimon (1999).

The difference equations described by the first-order conditions for $c_{0,t}$ and $c_{1,t}$, together with the pricing constraint, describe a complicated dynamic system. However, King and Wolman (1999) show that this system has a particularly simple limit point with unique local dynamics, namely zero inflation. In simple terms, if the policymaker's optimality criterion is present discounted utility, then he or she will choose a path that approaches zero inflation in the long run.

It is somewhat surprising that optimal policy approaches a steady state with lower welfare than may be attainable in steady state. However, this result has an analogue in growth theory that can help us understand what is going on. The result from growth theory involves what are known as the golden rule and the modified golden rule. Under the golden rule, the stock of capital is that which supports the highest possible steady state consumption. However, a planner maximizing present value utility would not choose to build up or maintain this level of capital stock. Instead, the planner would choose to accumulate less capital; while this would lead to lower consumption in the long run, in the short run consumption could be higher as the excess capital was converted to consumption goods. Because present consumption is assumed to be preferred to future consumption, such a plan is optimal (see Blanchard and Fischer [1989, 45] and McCallum [1996, 49]).

A similar phenomenon occurs in the staggered pricing model. In the long run, the economy will approach a steady state with a higher markup and hence lower consumption than is feasible. However, this is optimal because the transition path generates higher consumption and a lower markup than can be achieved in the optimal steady state described in the previous section.[5]

The long-run limiting behavior under optimal policy with commitment corresponds to what Michael Woodford (1999) has called a *timeless perspective*. Under the timeless perspective, the policymaker behaves each period in the way he or she would have promised to behave if asked to commit in the long-distant past. Woodford advocates that the long-run limiting behavior under the full-commitment solution be adopted *in every period* by policymakers who can commit, in part because that behavior leads to stationary outcomes over time: period zero is not treated as special. However, optimality of the long-run limit is inextricably linked to the high welfare levels in transitional start-up periods; specifically, the long-run limit is optimal only as part of an entire path that includes the start-up periods.[6] If commitment is feasible and

[5] Unpublished work by the author suggests that the transition path can be complicated, for example displaying nonmotonic behavior of the markup. It is clear, however, that there are some periods during the transition in which the markup is lower and utility is higher than in the optimal steady state.

[6] Woodford conducts his analysis in a different model, where the steady state is unaffected by monetary policy. The distinction between the golden rule and modified golden rule *steady states* thus does not exist in Woodford (1999).

the optimality criterion requires that policy be stationary, then optimal policy in the current model is represented by the optimal steady state of Section 3. Rotemberg and Woodford (1999) take this approach.[7]

5. OPTIMAL POLICY III: PRESENT VALUE WELFARE WITHOUT COMMITMENT

Maintaining the natural welfare criterion of the previous section, suppose that no promises are credible. Policy cannot commit to future actions, and thus the current policymaker cannot affect expectations about future outcomes. Effectively, there is a new policymaker each period. One can again use the primal approach to study this problem. The current period policymaker will choose $c_{0,t}$ and $c_{1,t}$ subject to the constraint imposed by optimal price-setting, with the location of this constraint determined by the expected levels of $c_{0,t+1}$ and $c_{1,t+1}$. Once the problem has been thus expressed, it is straightforward to interpret the policymaker's choice variables as the two distortions (markup and relative price), rather than the two consumption levels.

If policy can commit to future actions, optimal policy varies over time. In the initial period the policymaker optimally takes advantage of preset prices by expanding output. Pricing behavior anticipates all future actions, so no surprises are possible after the initial period. Nonetheless, the dynamic path does not immediately reach the long-run limit. In contrast, when the policymaker cannot commit, the current period does not differ from any other period: *every* period is an initial period. The policy problem is stationary, and this leads me to look for a stationary equilibrium with discretionary optimization.[8]

Before discussing the details of characterizing equilibrium, I will briefly relate the analysis to Barro and Gordon (1983), with which some readers may be familiar. There, equilibrium was determined by analyzing how current policy responded to expectations about current policy. In contrast, I will analyze how current variables (which are determined by current policy subject to the pricing constraint) optimally respond to the expected *future* variables (which determine the location of the pricing constraint). A fixed point of this relationship is a steady state equilibrium with discretionary policy. In Barro and Gordon's model, as long as one abstracts from reputational considerations, there is no reason for future policy to play a role in equilibrium because prices are not set for multiple periods. Furthermore, all relevant expectations

[7] For an interesting and detailed discussion of the timeless perspective, see Dennis (2001).

[8] Without commitment, there may be many equilibria. We describe the unique Markov-Perfect equilibrium (see Krusell and Rios-Rull [1999]), meaning the equilibrium that is determined by the economy's natural state variables. Because there are no state variables in our model, the Markov-Perfect equilibrium is a steady state. Khan, King, and Wolman (2001) discuss a variant of this model where prices are set for three periods and there is thus one natural state variable. They find multiple equilibria.

are determined simultaneously with the policy action. Herein, on the other hand, future policy directly affects current behavior, and the timing is more complicated. The policymaker takes as given the prices set by firms in the previous period, but current price-setting firms make their decisions *after* the policymaker. These differences dictate using a forward-looking approach to solve the model.

I express the model in terms of variables in the current period (for example c_0 and c_1) and variables in the next period (c_0' and c_1'). A stationary equilibrium under discretionary optimal policy (i.e., no commitment) consists of scalars v^*, c_0^*, and c_1^*, which solve (P1) when $v' = v^*$, $c_0' = c_0^*$, and $c_1' = c_1^*$:

$$v = \max_{c_0, c_1} \left\{ u\left(c\left(c_0, c_1\right), l\left(c_0, c_1\right)\right) + \beta \cdot v' \right\} \tag{P1}$$

subject to

$$0 = x\left(c_0, c_1\right) + \beta x\left(c_1', c_0'\right) \tag{25}$$

$$c_0', c_1', v' \text{ given.}$$

In principle, it is straightforward to work with this formulation. However, it is easier to develop intuition by transforming the problem so that the choice variables are the two distortions introduced earlier, instead of c_0 and c_1. To express c_0 and c_1 implicitly as functions of μ and ρ, use (16) and (17). The appendix derives $c_1/c_0 = \Gamma(\rho)$, $c_0 = \Omega(\mu, \rho)$, and hence $c_1 = \Gamma(\rho) \cdot \Omega(\mu, \rho)$. Problem (P1) can then be written

$$v = \max_{\mu, \rho} \left\{ \widetilde{u}\left(\mu, \rho\right) + \beta \cdot v' \right\} \tag{P1$'$}$$

subject to

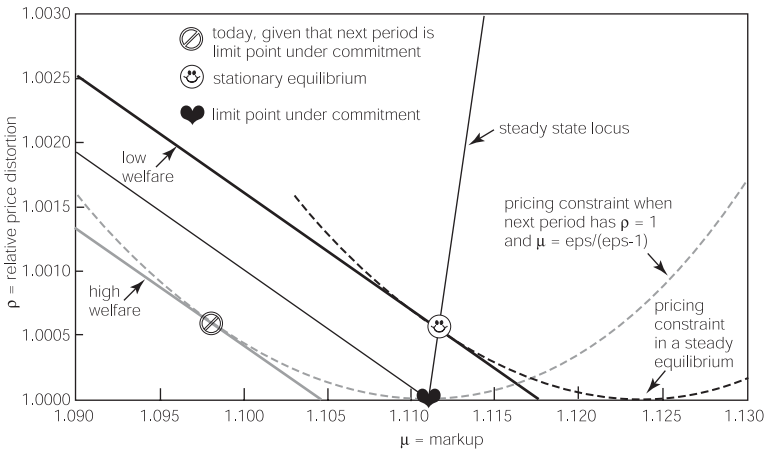
$$0 = \widetilde{x}\left(\mu, \rho\right) + \beta \widehat{x}\left(\mu', \rho'\right) \tag{26}$$

$$\rho', \mu', v' \text{ given.}$$

It is easy to analyze this transformed problem graphically.

Figure 4 illustrates the nature of (P1′) and its equilibrium for the same example used above. The markup is on the horizontal axis, and the relative price distortion is on the vertical axis. The figure is similar to the indifference curve/budget constraint graphs common in microeconomics with three important exceptions. First, welfare is increasing toward the origin because the two distortions can be thought of as bads not goods. Second, unlike in textbook examples from microeconomics where the budget constraint is a fixed line, here there are many constraints (a continuum, with only three constants shown in the figure) because there are many possible outcomes in the next period, and outcomes next period determine the location of the constraint (μ' and ρ' appear in (26)). Finally, recall from microeconomics that optimal behavior implies a tangency between the budget constraint and an indifference curve. The multiple constraints here create the possibility of multiple tangencies, so

Figure 4 Stationary Equilibrium without Commitment



in order to identify the equilibrium among these tangencies (assuming the equilibrium is unique), one needs more information. The locus of steady state points provides this information.

The mildly concave downward sloping curves are indifference curves associated with the indirect utility function in (P1′); welfare is increasing toward the origin, at which point both distortions would be eliminated. The indifference curves slope downward because both objects are bads; if the markup rises, the relative price distortion must fall for welfare to be unchanged.

The convex parabolic curves are pricing constraints for three different assumptions about the future markup and relative price distortion (μ' and ρ'). To explain the shape of the pricing constraints, I will focus on the flat points that occur for each constraint at $\rho = 1$. When $\rho = 1$, it is also the case that $c_0 = c_1 = c$ and $P_0 = P_1 = P$: all firms charge the same price and sell the same quantity. For the constraint corresponding to a given pair (μ', ρ'), the $\rho = 1$ point reveals the current markup (and thus real marginal cost) for which a relative price equal to unity maximizes the present value of profits. Now vary real marginal cost either up or down; from equation (13) the optimal price will change, necessarily moving away from the price set by firms in the previous period. With the two types of firms charging different prices, the relative price distortion is necessarily greater than one. The further marginal cost moves from the level associated with no relative price distortion, the greater the relative price distortion that must be accepted. This explains why the relative price distortion rises along every constraint as we move away from the point where $\rho = 1$.

The curve that slices through the middle of the figure—apparently positively sloped—is the locus of points that correspond to steady state equilibria; to compute this locus, I impose a steady state and vary the money growth rate, tracing out the (ρ, μ) locus that results. Along the steady state locus, moving upward usually corresponds to raising the steady state inflation rate. At higher rates of inflation, the dispersion—and hence distortion—in relative prices increases. The average markup also usually rises with the steady state inflation rate: higher markups by adjusting firms more than offset the increased markup erosion experienced by nonadjusting firms.[9]

A policymaker in the current period takes as given firms' expectations about the future. These expectations determine the relevant pricing constraint, and the policymaker chooses levels of the current distortions such that an indifference curve is tangent to the relevant pricing constraint. At such a point, the rate at which firms' behavior allows the policymaker to trade off the markup against the relative price distortion is equated to the rate at which the policymaker's welfare function trades off the markup against the relative price distortion. The former rate is given by the slope of the pricing constraint, and the latter is given by the slope of the indifference curve. As indicated above, for arbitrary expectations about the future, a tangency point does not represent an equilibrium. Given the future outcome, a single tangency point does represent optimal policy and private sector equilibrium in the current period, but there is no guarantee that the future outcome taken as given is an equilibrium. If the future outcome and the current outcome are identical, then we have an equilibrium: the outcome taken as given in the future is found to be optimal in the current period, and because the future looks just like the present, that outcome will be optimal in the future.

Suppose firms expect that in the next period the markup will be at its static level $\left(\mu' = \varepsilon/(\varepsilon - 1)\right)$ and the relative price distortion will be eliminated $\left(\rho' = 1\right)$. This is the outcome in a steady state with zero inflation, and it is also the long-run limit point under optimal policy with commitment. This point is helpful in understanding the nature of equilibrium even though it is not itself an equilibrium. The current period policymaker then faces the dashed pricing constraint, which passes through the steady state locus at $\mu = \varepsilon/(\varepsilon - 1)$ and $\rho = 1$. It is feasible for the current policymaker to achieve the same outcome

[9] Some readers will correctly infer from Figures 1 and 2 that the steady state locus in Figure 4 is nonmonotonic and has a second branch not shown in the figure. The argument goes as follows. As the steady state inflation rate falls from high levels, the markup and the relative price distortion fall together, but at a low positive inflation rate shown in Figure 2, further decreases in inflation lead to higher markups, whereas the relative price distortion continues to fall until inflation turns into deflation. There is a small downward sloping portion of the steady state locus that corresponds to the low inflation region where the relative price distortion is rising with inflation and markup is falling with inflation. There is also a second branch of the locus—upward sloping—that lies to the right of the branch in Figure 4. One can show that no steady state equilibrium lies on this branch.

expected in the future. Because that outcome is the limit point of optimal policy under commitment, it is indicated with a heart in Figure 4. Because the relative price distortion is eliminated, the pricing constraint passes through this point with zero slope. Immediately this point can be ruled out as an equilibrium because the indifference curves have negative slope everywhere. A policymaker contemplating the heart outcome would see that he or she could do better by accepting some relative price distortion in exchange for a lower markup. Specifically, a policymaker facing $\mu' = \varepsilon/(\varepsilon - 1)$ and $\rho' = 1$ would choose the point marked with a slashed circle; it is on the same pricing constraint but tangent to an indifference curve with higher welfare. This tangency is not an equilibrium, though, because it implies a different outcome in the current period than in the future.[10]

A steady state discretionary equilibrium is a point *on the steady state locus* at which an indifference curve is tangent to a constraint. The point marked with a smile is the unique steady state equilibrium. The relative price distortion is fairly high, so the pricing constraint is steeply downward sloping and at this point is tangent to an indifference curve. It is feasible for the policymaker to reduce the markup from this point, but doing so would require an increase in the relative price distortion big enough to make welfare fall.[11] The high relative price distortion corresponds to a high inflation rate (around 15 percent annually). Comparing the steady state equilibrium under discretion to the heart-shaped point, which represents the limiting behavior under commitment, it is clear that in the long run the economy is worse off without commitment. Even though the discretionary policymaker acts in society's best interest, society would be better off if the policymaker could credibly commit to future policy actions.

6. CONCLUSIONS

There has been an explosion of research in recent years on sticky-price models with optimizing agents (see Taylor [1998]). At least three notions of optimal monetary policy are natural in these models: the optimal steady state inflation rate; the path that maximizes present value welfare when policy can commit to future actions; and the equilibrium that occurs when each period's policymaker

[10] To be clear, equilibrium does not necessarily imply constant outcomes over time. In the current model, however, with optimal discretionary policy there are no exogenous forces leading to changing outcomes over time. Thus, a Markov-Perfect equilibrium involves constant outcomes.

[11] Peter Ireland (1997) shows that when all firms set their price before the policymaker moves and for just one period, in an otherwise similar model there is no interior Markov-Perfect equilibrium. Because all firms charge the same price, unexpected monetary expansions have no cost to the policymaker in Ireland's model. No matter how high an inflation rate is expected, the policymaker would always choose to bring down the markup by making inflation even higher. Here there is an interior Markov-Perfect steady state because higher inflation exacerbates the relative price distortion.

maximizes present value welfare, but no policymaker can commit to future actions.

In one model where nominal factors affect real allocations only because of staggered price-setting, I obtain the following results. The optimal steady state inflation rate is slightly positive (less than 1 percent) because a very small amount of constant inflation decreases the economy's average markup. A policy that maximizes present value welfare under commitment leads toward zero inflation, for the higher markup that will result in the long run is preceded by a lower markup in early periods that are weighted more heavily. When policy cannot commit, the inflation rate that results from optimizing behavior is quite high, on the order of 15 percent. Corresponding to the higher inflation rate without commitment is a lower level of welfare for the representative agent. If staggered price-setting in fact represents the primary channel through which monetary policy affects real variables, the results in this article indicate the value to society of institutions that allow the monetary authority to credibly commit to future behavior.[12]

APPENDIX

1. EQUILIBRIUM PRICING CONSTRAINT

From the optimal pricing condition (13) we can derive (14), an equation in $c_{0,t}$, $c_{1,t}$, $c_{0,t+1}$, and $c_{1,t+1}$ only. The first step is to use the demand function for $c_{0,t}$ to write the left hand side of (13) in terms of $c_{0,t}$ and $c_{1,t}$:

$$P_{0,t}/P_t = \left(c_{0,t}/c\left(c_{0,t}, c_{1,t}\right)\right)^{-1/\varepsilon},$$

where the function $c\left(c_{0,t}, c_{1,t}\right)$ is given by the consumption aggregator (3). For the right hand side, again use the consumption aggregator to eliminate c_t and c_{t+1}, and use the time constraint to write leisure $\left(l_{t+j}\right)$ in terms of $c_{0,t+j}$ and $c_{1,t+j}$:

$$l\left(c_{0,t+j}, c_{1,t+j}\right) = 1 - n\left(c_{0,t+j}, c_{1,t+j}\right) = 1 - \frac{1}{2}\left(c_{0,t+j} + c_{1,t+j}\right). \quad (27)$$

[12] Practically speaking, a specific commitment about the nature of future policy could never be completely credible. However, feasible institutional arrangements can tie the policymaker's hands somewhat, decreasing the severity of the time consistency problem. For a comparative study of institutional arrangements for monetary policy in various countries, see Bernanke et al. (1999).

Next, use the labor supply equation (10) to write the real wage in terms of c_0 and c_1:

$$w\left(c_{0,t}, c_{1,t}\right) = u_l\left(c\left(c_{0,t}, c_{1,t}\right), l\left(c_{0,t}, c_{1,t}\right)\right) / u_c\left(c\left(c_{0,t}, c_{1,t}\right), l\left(c_{0,t}, c_{1,t}\right)\right), \tag{28}$$

and use the demand functions for c_0 and c_1 to write the inflation rate in terms of current and past c_0 and c_1:

$$\begin{aligned} \pi_{t+1}\left(c_{0,t}, c_{1,t}, c_{0,t+1}, c_{1,t+1}\right) &\equiv P_{t+1}/P_t = \frac{P_{0,t}/P_t}{P_{0,t}/P_{t+1}} = \\ \frac{P_{0,t}/P_t}{P_{1,t+1}/P_{t+1}} &= \left(\frac{c_{0,t}/c\left(c_{0,t}, c_{1,t}\right)}{c_{1,t+1}/c\left(c_{0,t+1}, c_{1,t+1}\right)}\right)^{-1/\varepsilon}. \end{aligned} \tag{29}$$

Substituting all of these relationships into (13) yields

$$\left(\frac{c_{0,t}}{c\left(c_{0,t}, c_{1,t}\right)}\right)^{-1/\varepsilon} = \left(\frac{\varepsilon}{\varepsilon - 1}\right) \frac{N\left(c_{0,t}, c_{1,t}, c_{0,t+1}, c_{1,t+1}\right)}{D\left(c_{0,t}, c_{1,t}, c_{0,t+1}, c_{1,t+1}\right)}, \tag{30}$$

where

$$\begin{aligned} N_t \equiv\ & u_l\left(c\left(c_{0,t}, c_{1,t}\right), l\left(c_{0,t}, c_{1,t}\right)\right) \cdot c\left(c_{0,t}, c_{1,t}\right) + \\ & \beta \cdot u_l\left(c\left(c_{0,t+1}, c_{1,t+1}\right), l\left(c_{0,t+1}, c_{1,t+1}\right)\right) \cdot c\left(c_{0,t+1}, c_{1,t+1}\right) \cdot \\ & \left(\frac{c_{0,t}/c\left(c_{0,t}, c_{1,t}\right)}{c_{1,t+1}/c\left(c_{0,t+1}, c_{1,t+1}\right)}\right)^{-1} \end{aligned}$$

$$\begin{aligned} D_t \equiv\ & u_c\left(c\left(c_{0,t}, c_{1,t}\right), l\left(c_{0,t}, c_{1,t}\right)\right) \cdot c\left(c_{0,t}, c_{1,t}\right) + \\ & \beta \cdot u_c\left(c\left(c_{0,t+1}, c_{1,t+1}\right), l\left(c_{0,t+1}, c_{1,t+1}\right)\right) \cdot c\left(c_{0,t+1}, c_{1,t+1}\right) \cdot \\ & \left(\frac{c_{0,t}/c\left(c_{0,t}, c_{1,t}\right)}{c_{1,t+1}/c\left(c_{0,t+1}, c_{1,t+1}\right)}\right)^{\frac{1-\varepsilon}{\varepsilon}}. \end{aligned}$$

Next, multiplying both sides of (30) by $(\varepsilon - 1) \cdot \left(c_{0,t}/c\left(c_{0,t}, c_{1,t}\right)\right) \cdot D_t$, and collecting terms according to whether they contain β factors, we arrive at

$$\begin{aligned} 0 = \quad & c_{0,t} \cdot \left[(\varepsilon - 1) \cdot u_{c,t} \cdot \left(\frac{c_{0,t}}{c\left(c_{0,t}, c_{1,t}\right)}\right)^{-1/\varepsilon} - \varepsilon \cdot u_{l,t}\right] + \\ & \beta \cdot c_{1,t+1} \cdot \left[(\varepsilon - 1) \cdot u_{c,t+1} \cdot \left(\frac{c_{1,t+1}}{c\left(c_{0,t+1}, c_{1,t+1}\right)}\right)^{-1/\varepsilon} - \varepsilon \cdot u_{l,t+1}\right]. \end{aligned} \tag{31}$$

This yields (14) in the text, with

$$x(a,b) \equiv a \cdot \left[\begin{array}{c} (\varepsilon - 1) \cdot u_c\left(c(a,b), l(a,b)\right) \cdot \left(\frac{a}{c(a,b)}\right)^{-1/\varepsilon} \\ -\varepsilon \cdot u_l\left(c(a,b), l(a,b)\right) \end{array}\right]. \tag{32}$$

(Note that $c(a, b) \equiv c(b, a)$ and $l(a, b) \equiv l(b, a)$ and that in (31) I have used abbreviated objects such as $u_c\left(c\left(c_{0,t}, c_{1,t}\right), l\left(c_{0,t}, c_{1,t}\right)\right)$ by writing them as $u_{c,t}$.)

2. c_0 AND c_1 AS FUNCTIONS OF THE TWO DISTORTIONS

The definitions of the two distortions immediately imply

$$\mu = a/w = a \cdot \frac{u_c\left(c\left(c_0, c_1\right), l\left(c_0, c_1\right)\right)}{u_l\left(c\left(c_0, c_1\right), l\left(c_0, c_1\right)\right)} \tag{33}$$

and

$$\rho = 2^{\frac{1}{\varepsilon-1}} \cdot \frac{\left(1 + c_1/c_0\right)}{\left(1 + \left(c_1/c_0\right)^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{\varepsilon}{\varepsilon-1}}}. \tag{34}$$

From (34), it is clear that the ratio c_1/c_0 depends only on ρ. That is, $c_1/c_0 = \Gamma(\rho)$, where $\Gamma(\rho)$ is the function defined implicitly by (34). Substitute this function into (33) to get

$$\mu = a \cdot \frac{u_c\left(c\left(c_0, \Gamma(\rho) \cdot c_0\right), l\left(c_0, \Gamma(\rho) \cdot c_0\right)\right)}{u_l\left(c\left(c_0, \Gamma(\rho) \cdot c_0\right), l\left(c_0, \Gamma(\rho) \cdot c_0\right)\right)}, \tag{35}$$

which implicitly gives c_0 as a function of μ and ρ. That is, $c_0 = \Omega(\mu, \rho)$, and hence $c_1 = \Gamma(\rho) \cdot \Omega(\mu, \rho)$.

REFERENCES

Bailey, Martin J. 1956. "The Welfare Cost of Inflationary Finance." *The Journal of Political Economy* 64 (April): 93–110.

Barro, Robert, and David B. Gordon. 1983. "A Positive Theory of Monetary Policy in a Natural-Rate Model." *Journal of Political Economy* 91 (August): 589–610.

Bernanke, Ben S., Thomas Laubach, Frederic S. Mishkin, and Adam S. Posen. 1999. *Inflation Targeting: Lessons From the International Experience*. Princeton N.J.: Princeton University Press.

Blanchard, Olivier Jean, and Stanley Fischer. 1989. *Lectures on Macroeconomics*. Cambridge, MA: MIT Press.

Dennis, Richard. 2001. "Pre-Commitment, the Timeless Perspective, and Policymaking from Behind a Veil of Uncertainty." Manuscript, Federal Reserve Bank of San Francisco.

Friedman, Milton. 1969. "The Optimum Quantity of Money," in *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine Publishing Company.

Goodfriend, Marvin, and Robert G. King. 2001. "The Case for Price Stability." Working Paper 01-2. Federal Reserve Bank of Richmond.

_____. 1997. "The New Neoclassical Synthesis and the Role of Monetary Policy." In *NBER Macroeconomics Annual,* ed. Ben Bernanke and Julio Rotemberg. Cambridge, Mass.: MIT Press: 231–95.

Ireland, Peter. 1997. "Sustainable Monetary Policies." *Journal of Economic Dynamics and Control* 22 (November): 87–108.

Khan, Aubhik, Robert G. King, and Alexander L. Wolman. 2000. "Optimal Monetary Policy." Working Paper 00-10. Federal Reserve Bank of Richmond.

_____. 2001. "The Pitfalls of Monetary Discretion." Working Paper 01-8. Federal Reserve Bank of Richmond.

King, Robert G., and Alexander L. Wolman. 1996. "Inflation Targeting in a St. Louis Model of the 21st Century." *Federal Reserve Bank of St. Louis Review* 78 (May/June): 83–107.

_____. 1999. "What Should the Monetary Authority do When Prices are Sticky?" In *Monetary Policy Rules,* ed. John B. Taylor. Chicago: University of Chicago Press: 349–98.

Krusell, Per, and Jose Victor Rios-Rull. 1999. "On the Size of U.S. Government: Political Economy in the Neoclassical Growth Model." *American Economic Review* 89 (December): 1156–81.

Kydland, Finn, and Edward C. Prescott. 1980. "Dynamic Optimal Taxation, Rational Expectations and Optimal Control." *Journal of Economic Dynamics and Control* 2 (February): 79–91.

_____. 1977. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy* 85 (June): 473–92.

Marcet, Albert, and Ramon Marimon. 1999. "Recursive Contracts." Manuscript. http://www.iue.it/Personal/Marimon/recursive399.pdf.

McCallum, Bennett T. 1996. "Neoclassical vs. Endogenous Growth Analysis: An Overview." Federal Reserve Bank of Richmond *Economic Quarterly* 82 (Fall): 41–72.

Rotemberg, Julio J., and Michael Woodford. 1999. "Interest Rate Rules in an Estimated Sticky Price Model." In *Monetary Policy Rules,* ed. John B. Taylor. Chicago: University of Chicago Press: 57–119.

Taylor, John B. 1999. "Staggered Wage and Price Setting in Macroeconomics." In *Handbook of Macroeconomics,* vol. 1B, eds. John B. Taylor and Michael Woodford. Amsterdam: Elsevier Science B.V.: 1009–50.

Wolman, Alexander L. 1999. "Sticky Prices, Marginal Cost, and the Behavior of Inflation." Federal Reserve Bank of Richmond *Economic Quarterly* 85 (Fall): 29–48.

_____. 2000. "The Frequency and Costs of Individual Price Adjustment." Federal Reserve Bank of Richmond *Economic Quarterly* 86 (Fall): 1–22.

Woodford, Michael. 1999. "Commentary: Monetary Policy at Zero Inflation." In *New Challenges for Monetary Policy*. Kansas City: Federal Reserve Bank of Kansas City: 277–316.

Yun, Tack. 1996. "Nominal Price Rigidity, Money Supply Endogeneity, and Business Cycles." *Journal of Monetary Economics* 37 (July): 345–70.

International Pricing in New Open-Economy Models

Margarida Duarte

Recent developments in open-economy macroeconomics have progressed under the paradigm of nominal price rigidities, where monetary disturbances are the main source of fluctuations. Following developments in closed-economy models, new open-economy models have combined price rigidities and market imperfections in a fully microfounded intertemporal general equilibrium setup. This framework has been used extensively to study the properties of the international transmission of shocks, as well as the welfare implications of alternative monetary and exchange rate policies.

Imperfect competition is a key feature of the new open-economy framework. Because agents have some degree of monopoly power instead of being price takers, this framework allows the explicit analysis of pricing decisions. The two polar cases for pricing decisions are *producer-currency pricing* and *local-currency pricing*. The first case is the traditional approach, which assumes that prices are preset in the currency of the seller. In this case, prices of imported goods change proportionally with unexpected changes in the nominal exchange rate, and the law of one price always holds.[1] In contrast, under the assumption of local-currency pricing, prices are preset in the buyer's currency. Here, unexpected movements in the nominal exchange rate do not affect the price of imported goods and lead to short-run deviations from the law of one price.

The author would like to thank Michael Dotsey, Thomas Humphrey, Yash Mehra, and John Walter for helpful comments. The views expressed in this article do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] The law of one price states that, absent barriers to trade, a commodity should sell for the same price (when measured in a common currency) in different countries.

Empirical evidence using disaggregated data suggests that international markets for tradable goods remain highly segmented and that deviations in the law of one price are large, persistent, and highly correlated with movements in the nominal exchange rate, even for highly tradable goods. Moreover, there is strong evidence that the large and persistent movements that characterize the behavior of real exchange rates at the aggregate level are largely accounted for by deviations in the law of one price for tradable goods.

In this article I make use of a simplified version of a two-country model where the two markets are segmented, allowing firms to price discriminate across countries, and where prices are preset in the consumer's currency. This model generates movements in the real exchange rate in response to unexpected monetary shocks, which are a result of the failure of the law of one price for tradable goods. I then compare this model to a version in which prices are preset in the producer's currency and examine the implications of these two alternative price-setting regimes for several key issues.

The price-setting regime determines the currency of denomination of imported goods and the extent to which changes in exchange rates affect the relative price of imported to domestic goods and the international allocation of goods in the short run. That is, different pricing regimes imply different roles for the exchange rate in the international transmission of monetary disturbances. As we shall see, this assumption has very striking implications for several important questions, namely real exchange rate variability, the linkage between macroeconomic volatility and international trade, and the welfare effects of alternative exchange rate regimes, among others.

While generating deviations from the law of one price that are absent from models assuming producer-currency pricing, the assumption of local-currency pricing still leaves important features of the data unexplained. The key role of this assumption in the properties of open-economy models suggests that it is necessary to keep exploring the implications of alternative pricing structures in open-economy models.

In Section 1, I review the empirical evidence on the behavior of real exchange rates and on international market segmentation and pricing. In Section 2, I present the model with local-currency pricing and explore the main implications of this pricing assumption. The final section concludes.

1. SOME EVIDENCE ON REAL EXCHANGE RATES

I first review some empirical evidence on the behavior of real exchange rates using aggregate data. I then turn to a review of the evidence on the sources of movements in real exchange rates.

Real Exchange Rates and PPP

The real exchange rate between two countries represents the relative cost of a common reference basket of goods. For two countries, say the United States and Japan, the real exchange rate is given by

$$\frac{P_{US}}{eP_{JP}},$$

where P_{US} and P_{JP} represent the American and Japanese price levels (measured in terms of dollars and yen, respectively) and where e denotes the nominal exchange rate (defined as the dollar price of one yen).[2]

The theory of purchasing power parity (PPP) predicts that real exchange rates should equal one, or at least show a strong tendency to quickly return to one when they differ from this value. The fundamental building block of PPP is the law of one price: due to arbitrage in goods markets, and absent barriers to trade, similar products should sell in different countries for the same price (when converted in the same currency). Large international price differentials would be only temporary, as profit-maximizing traders would quickly drive international goods prices back in line. Therefore, if arbitrage in goods markets ensures that the law of one price holds for a sufficiently broad range of individual goods, then aggregate price levels (when expressed in a common currency) should be highly correlated across countries.[3]

Because aggregate prices are reported as indices rather than levels, most empirical work has tested the weaker hypothesis of relative PPP, which requires only that the real exchange rate be stable over time.[4] Figure 1 shows the log changes in the CPI-based dollar-yen real and nominal exchange rates and the relative price level. In this figure, which is typical for countries with floating exchange rates and moderate inflation, it clearly stands out that short-run deviations from PPP are large and volatile. In the short run, movements in the real exchange rate mimic those in the nominal exchange rate, with no offsetting movements in the relative price level. Not surprisingly, early empirical work based on simple tests of short-run PPP produced strong rejections of this hypothesis for moderate inflation countries.[5] However, these studies did not allow for any dynamics of adjustment to PPP and therefore did not address the validity of PPP as a medium- or long-run proposition.

---

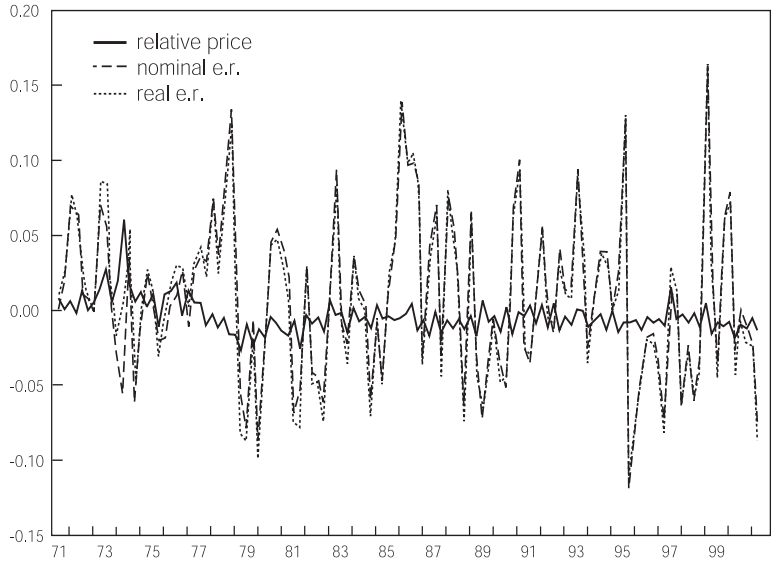
[2] Suppose that the United States and Japan have the same price levels when measured in their respective currencies (for example, $P_{US} = P_{JP} = 1$) and that the nominal exchange rate is two (that is, two dollars are required to buy one yen). Then, the Japanese price level is two when measured in dollars and the real exchange rate between the United States and Japan is 0.5.

[3] For a thorough exposition of the evolution of the PPP theory of exchange rates, see Humphrey and Keleher (1982, chapter 11).

[4] In other words, relative PPP requires only that changes in relative price levels be offset by changes in the nominal exchange rate.

[5] See, for example, Frenkel (1981) or Krugman (1978).

Figure 1 Nominal and Real Exchange Rates and Relative Price Changes between Japan and the United States



Sources: Federal Reserve Board of Governors, Bloomberg, and IMF International Financial Statistics

The conventional explanation for the failure of short-run PPP is the presence of nominal price rigidities. If the short-term volatility of nominal exchange rates were due mostly to monetary and financial disturbances, then nominal price stickiness would translate these disturbances into short-run fluctuations in the real exchange rate. If this were true, however, we should observe a substantial convergence to PPP in one to two years, as the adjustment of prices and wages takes place. Purchasing power parity, therefore, would be reestablished in the medium to long run.[6]

An extensive body of empirical literature has tested the hypothesis of long-run PPP by looking at the mean-reverting properties of real exchange rates. As is well known, it has proved rather difficult to find evidence supporting convergence of real exchange rates to PPP even in the long run.[7]

Most earlier empirical studies, which used only post-Bretton Woods data, found it difficult to reject the hypothesis that bilateral real exchange rates for

[6] See Stockman (1987) for an alternative, equilibrium view of exchange rates.

[7] For a survey of this literature and a more complete list of references, see Froot and Rogoff (1995) or Rogoff (1996).

industrialized countries follow a random walk under floating exchange rates.[8] But if PPP deviations are very persistent, then it may be difficult to distinguish empirically between a random walk model and a slow mean-reversion model for the real exchange rate, especially when this variable is highly volatile. As shown in Frankel (1986), the post-Bretton Woods period may simply be too short to reliably reject the random walk hypothesis. To overcome this problem of low power in tests of the random walk hypothesis, Frankel used an extended data set (annual data for the dollar-pound exchange rate from 1869 to 1984) and rejected the random walk model in favor of a mean-reverting model for the real exchange rate. His point estimate for the rate of decay of real exchange rate deviations was 14 percent per year, which implies a half-life of PPP deviations of 4.6 years. Other studies that test convergence to PPP using long-horizon data sets tend to find values for the half-life of PPP deviations between three to five years.[9]

An alternative way to increase the power of unit root tests is to expand the number of countries in the sample and to perform panel tests of convergence to PPP. Frankel and Rose (1996), for example, use a panel set of annual data from 1948 to 1992 for 150 countries. They estimate half-lives for PPP deviations of about four years. Other studies using panel data sets report similar estimates. Interestingly, these estimates are also similar to those obtained using long-time series data sets.

In brief, studies using aggregate data provide strong evidence that deviations from PPP are highly volatile and persistent. Consensus estimates suggest that the speed of convergence to PPP is roughly 15 percent per year, implying a half-life of PPP deviations of about four years. As we shall see next, a look at disaggregated data will provide us with a much richer analysis of the sources of PPP deviations.

The Law of One Price: Market Segmentation and International Pricing

As I pointed out earlier, the idea underlying PPP is that the law of one price holds for a wide range of individual goods. It has long been recognized, however, that even for highly tradable goods and at different levels of aggregation,

[8] Typically, the real exchange rate, q_t, is assumed to follow a linear $AR(1)$ specification,

$$q_t = \rho q_{t-1} + \epsilon_t,$$

where $\epsilon_t \sim N\left(0, \sigma^2\right)$. This specification means that the adjustment of PPP deviations is both continuous and of constant speed, regardless of the size of the deviation. Given this specification, the *convergence speed* is given by $\lambda = 1 - \rho$ and the *half-life of deviations* is given by $H = \frac{\ln 0.5}{\ln \rho}$.

[9] Mussa (1986) demonstrates that real exchange rates tend to be much more volatile under floating than under fixed exchange rate regimes. Therefore, these long-horizon data sets mix data from different regimes, which exhibit different properties for the real exchange rate. See Lothian and Taylor (1996) for a response to this criticism.

deviations in the law of one price are large, persistent, and highly correlated with movements in the nominal exchange rate.[10]

One possible explanation for the failure of the law of one price is that international markets are segmented by physical distance, like different markets within a country. Engel and Rogers (1996), however, show that both the distance and the physical border between countries are significant in explaining the variation in prices of similar goods across different U.S. and Canadian cities. They find that price dispersion is much higher for two cities located in different countries than for two equidistant cities in the same country. In fact, the effect of the border is estimated to be equivalent to a distance of 1780 miles between cities within one country. Engel and Rogers also show that nominal price stickiness accounts for a large portion of the border effect, suggesting that prices are sticky in the local currency and that changes in the exchange rate lead to deviations in the law of one price.

Not only are failures of the law of one price significant but, as recent evidence suggests, they also play a dominant role in explaining the behavior of real exchange rates. Engel (1999) measures the proportion of U.S. real exchange rate movements that can be accounted for by movements in the relative prices of nontraded goods. Engel decomposes the CPI real exchange rate into two components: a weighted difference of the relative price of nontraded- to traded-goods prices in each country, and the relative price of traded goods between the countries. If tradables, as a category, closely followed the law of one price, then all variability in the real exchange rate would be explained by movements in the first component. However, Engel finds that movements in the relative price of nontraded goods appear to account for almost none of the movement in U.S. real exchange rates, even at long time horizons. Instead, nearly all the variability can be attributed to movements in the relative price of tradables. This finding strongly suggests that consumer markets for tradable goods are highly segmented internationally and that movements in the international relative price of consumer tradables are very persistent.[11] Moreover, given the high volatility of nominal exchange rates, these findings indicate that consumer prices of most goods (either imported or domestically produced) seem to be sticky in domestic currency terms.

An alternative approach to studying the relationship between exchange rates and goods prices is examining how firms in a industry (or country) pass through changes in exchange rates to export prices.[12] Knetter (1989, 1993)

[10] See, for example, the empirical studies in Isard (1977), Giovannini (1988), or Engel (1993).

[11] One should note, however, that at the consumer level, even highly tradable goods embody a large nontradable component.

[12] Exchange rate pass-through is the percentage change in local currency import prices resulting from a 1 percent change in the exchange rate between the exporting and importing countries.

measures the degree of price discrimination across export destinations that is associated with exchange rate changes for U.S., U.K., German, and Japanese industry-level data. He finds that the amount of exchange rate pass-through differs considerably depending on the country and industry. Goldberg and Knetter (1997) provide an extensive survey of the literature and find that local currency prices of foreign products do not respond fully to exchange rate changes. While the response varies by industry, on average exchange rate pass-through to U.S. import prices is only about 50 percent after one year, mainly reflecting changes in destination-specific markups on exports.

In brief, there is strong evidence that international markets for tradable goods remain highly segmented and that deviations from PPP are largely accounted for by movements in the relative price of tradable goods across countries. At the consumer level, exchange rate pass-through to import prices is virtually zero (suggesting that consumer prices are sticky in domestic currency). At the producer level, however, exchange rate pass-through is generally positive, but substantially below one.

Transaction Costs and the Adjustment of PPP and Law of One Price Deviations

Some recent empirical tests of long-run PPP and the law of one price have abandoned the conventional framework, which assumes a linear autoregressive process for the price differential. Instead, these studies have started to look into nonlinear models of price adjustment, where the speed at which price differentials die out depends on the size of the deviation itself.

This alternative framework for the empirical analysis of price differentials is motivated by the observation that commodity trade is not costless. Persistent deviations from the law of one price are implied as an equilibrium feature of models with transaction costs, for deviations will be left uncorrected as long as they are sufficiently small relative to the shipping cost.[13]

The simplest econometric model that implements the notion of a nonlinear adjustment for price differentials assumes that the process is well described by a random walk for small deviations (that is, when deviations are within a "band of inaction") and an autoregressive process for large deviations (that is, when deviations are outside the band).[14] Taylor (2001) shows that the

[13] See Dumas (1992) and Ohanian and Stockman (1997) for equilibrium models of exchange rate determination in the presence of transaction costs. Obstfeld and Rogoff (2000a) argue for the importance of transaction costs in explaining several puzzles in international macroeconomics.

[14] Specifically, the price differential q_t follows the process

$$q_t = \begin{cases} c + \rho\left(q_{t-1} - c\right) + \varepsilon_t & \text{if } q_{t-1} > c, \\ q_{t-1} + \varepsilon_t & \text{if } c \geq q_{t-1} \geq -c, \\ -c + \rho\left(q_{t-1} + c\right) + \varepsilon_t & \text{if } -c > q_{t-1}, \end{cases}$$

improper use of linear models when the true model is nonlinear may produce a large bias towards finding a low speed of convergence.[15] Intuitively, a linear model will fail to support convergence to PPP if the true model is nonlinear and the process spends most of the time in the random-walk band. Using both monthly data from the 1920s and annual data spanning two centuries, Michael, Nobay, and Peel (1997) reject the linear adjustment model in favor of a nonlinear model and provide strong evidence of mean-reverting behavior for PPP deviations for every exchange rate considered.[16]

2. INTERNATIONAL PRICING IN NEW OPEN-ECONOMY MACROECONOMIC MODELS

The common starting point for most of the recent research in open-economy models with price rigidities is the model developed in Obstfeld and Rogoff (1995).[17] This model explores the international monetary transmission mechanism in a general equilibrium setup characterized by nominal price rigidities, imperfect competition, and incomplete asset markets.

Obstfeld and Rogoff's model does not generate deviations from the CPI-based purchasing power parity. This feature reflects the fact that preferences are identical across countries and that all goods are freely tradable, with prices set in the seller's currency. In this model, there is complete pass-through of exchange rate changes to import prices, implying that the law of one price always holds for all goods and that the real exchange rate is constant.

Motivated by the empirical evidence on the sources of real exchange rate fluctuations, several recent papers have extended Obstfeld and Rogoff's framework in order to allow for pricing-to-market[18] and deviations from the law of one price. This class of models assumes that home and foreign markets are segmented, which allows imperfectly competitive firms to price discriminate between home and foreign consumers.[19] Consumers' inability to arbitrage price differentials between countries is exogenous, possibly reflecting arbitrarily high transportation costs at the consumer level. In addition to market

---

where $\varepsilon_t \sim N\left(0, \sigma^2\right)$. This process is parametrized by ρ, the autoregressive coefficient for deviations from the band's edge, and c, which determines the amplitude of the band of inaction.

[15] Taylor (2001) also addresses the problem of temporal aggregation and shows that the use of relatively low-frequency data may also produce large biases in these estimates.

[16] See also Obstfeld and Taylor (1997) and Taylor, Peel, and Sarno (2001).

[17] This work extends the model in Svensson and van Wijenbergen (1989), an endowment two-country dynamic general equilibrium model, where monopolistic competitive firms set prices one period in advance and asset markets are complete.

[18] Strictly speaking, the term *pricing-to-market* refers to the ability of firms to engage in third-degree price discriminations across different export destinations. In its current use, however, the term has come to include the additional assumption that firms set their prices in advance in the local currency of the buyer.

[19] See Betts and Devereux (1996) for the initial contribution.

segmentation, this class of models also assumes that prices are sticky in each country's local currency. That is, firms set prices in advance in the buyer's currency, as opposed to the standard assumption that prices are set in the seller's currency.[20]

I next outline a basic model in which firms set prices in advance in the local currency of the buyer (or pricing-to-market). The model is then used to explore the main implications of pricing-to-market.

A Simple Model of Local-Currency Pricing

There are two countries, home and foreign. Households in each country consume a continuum of differentiated goods, which are indexed by i, $i \in [0, 1]$. A fraction n of these goods is produced by firms located in the home country, and the remaining fraction $1 - n$ is produced by firms located in the foreign country.[21]

Home and foreign households have identical preferences. In the home country, these preferences are defined by

$$U = \sum_{t=0}^{\infty} \beta^t u\left(c_t, \frac{M_t}{P_t}, 1 - l_t\right).$$

The term c_t represents the agent's total consumption. It is an index given by

$$c_t = \left[\int_0^1 c_t(i)^{\frac{\theta-1}{\theta}} di\right]^{\frac{\theta}{\theta-1}}, \tag{1}$$

which aggregates the consumption of all differentiated goods, $c_t(i)$. The parameter θ is the elasticity of substitution between any two differentiated goods, and for values of θ greater than 1, different goods are imperfect substitutes in consumption. Besides consumption, the consumer's momentary utility also depends on leisure, $1 - l_t$, and real money balances held during the period, $\frac{M_t}{P_t}$, where M_t are nominal balances and P_t is the home country consumption price index.

Let $p_t(i)$ represent the home currency price of good i. Given these prices, P_t represents the minimum expenditure necessary to buy one unit of composite good c. The price index corresponding to c is given by

$$P_t = \left[\int_0^1 p_t(i)^{1-\theta} di\right]^{\frac{1}{1-\theta}}. \tag{2}$$

[20] In these models, the firm's choice of invoice currency is exogenous. See Devereux and Engel (2001) for a recent contribution to the literature in which exporting firms can also choose the currency in which they set export prices. They find that exporters will generally wish to set prices in the currency of the country that has the most stable monetary policy.

[21] The fractions n and $1 - n$ also represent the sizes of the home and foreign countries, respectively.

Given the aggregate consumption index (1), the household's optimal allocation of consumption across each of the differentiated goods yields the demand functions

$$c_t(i) = \left(\frac{p_t(i)}{P_t}\right)^{-\theta} c_t,\ i \in [0, 1]. \tag{3}$$

Note that home demand functions for foreign goods $c_t(i)$, $i \in [n, 1]$, do not depend on the nominal exchange rate. As we shall see, this follows from the fact that the home price of foreign goods is denominated in the home currency.

As outlined above, home and foreign markets are segmented, effectively allowing firms to price discriminate across the two markets. Therefore, home firm i, $i \in [0, n]$, will choose separately the price for its good in the home country, $p_t(i)$, and in the foreign country, $p_t^*(i)$, in order to maximize its total profits. By assumption, these prices are denominated in the buyer's currency. That is, $p_t(i)$ is denominated in home currency and $p_t^*(i)$ is denominated in foreign currency.

Home firm i operates the production function $y_t = l_t(i)$, where $l_t(i)$ represents hours worked, and period t profits are given by $\pi_t(i) = p_t(i)\, c_t(i) + e_t p_t^*(i)\, c_t^*(i) - w_t(c_t(i) + c_t^*(i))$. The term w_t is the real wage rate and the nominal exchange rate, e_t, converts the revenues from sales in the foreign country into home currency. Profit maximization is made subject to the firm's production function and home and foreign demand functions for its good (equation (3) and the analogous expression for the foreign consumer).

When nominal prices are flexible, home firm i sets its prices as

$$p_t(i) = e_t p_t^*(i) = \frac{\theta}{\theta - 1} w_t,$$

i.e., the optimal pricing function rule for each firm is to set its price in each market as a constant markup over marginal cost.[22] Therefore, the law of one price holds for each good, even though firms have the ability to price discriminate across markets. The model with flexible prices does not generate deviations from PPP.[23]

Next suppose that firms set prices in advance at a level that achieves the optimal markup in the absence of shocks. Firms cannot adjust prices within the period in response to shocks, accommodating ex-post demand at the preset prices. Prices adjust fully after one period. As before, firms are assumed to set prices in the local currency of sale. Therefore, in this case, unanticipated changes in the exchange rate lead to deviations in the law of one price. In this model, deviations from PPP result only from deviations from the law of

[22] In this monopolistic competition framework, markups are constant, precluding the analysis of possible effects of exchange rates on markups.

See Bergin and Feenstra (1998) for a pricing-to-market model with translog preferences that departs from the monopolistic competition framework.

[23] This is a result of assuming that the elasticities of demand are identical in both markets.

one price, i.e., from movements in the relative price of similar goods across countries.

The Transmission of Monetary Shocks

When prices are preset in the buyer's currency, an unexpected depreciation of domestic currency has no expenditure-switching effect in the short run. In response to the exchange rate change firms are assumed to keep foreign currency export prices fixed, allowing their foreign markups to adjust. Since consumer demand functions do not depend on the nominal exchange rate and exchange rate pass-through to consumer prices is zero on impact, changes in this variable are dissociated, on impact, from allocation decisions.

In response to an unexpected positive shock to the home money supply, the nominal exchange rate immediately depreciates. Since prices only respond after one period and are denominated in the buyer's currency, the adjustment in the nominal exchange rate translates into a real depreciation and does not affect the relative price of home and foreign goods in either country. Thus, the increase in total consumption in the home country associated with the positive money shock is brought about by an increase in consumption of both domestic and foreign goods in the same proportion, as equation (3) shows. If, instead, prices were set in the seller's currency, the increase in the nominal exchange rate would lead to an immediate increase in the home currency price of foreign goods ($e_t p_t(f)$, where $p_t(f)$ is now denominated in foreign currency), while the price of home goods in the home country, $p_t(h)$, would remain unchanged. Similarly, nominal depreciation would reduce the foreign currency price of home goods ($\frac{p_t^*(h)}{e_t}$, with $p_t^*(h)$ denominated in home currency), while leaving the price of foreign goods in the foreign country, $p_t^*(f)$, unchanged. Thus, in this case, the positive money shock would decrease the relative price of home to foreign goods on impact in both countries[24] and both agents would substitute consumption towards home goods and away from foreign goods. Thus, having prices set in the buyer's currency eliminates, on impact, the expenditure switching effect associated with unexpected changes in the nominal exchange rate; the absence of this effect in turn influences the international transmission of monetary disturbances.

Without pricing-to-market, monetary disturbances tend to generate high positive comovements of consumption across countries and large negative comovements of output. In response to a positive money shock in the home country, the real exchange rate (i.e., the relative price of consumption across

[24] With seller's currency, this relative price in the home country would be $\frac{p_t(h)}{e_t p_t(f)}$, where $p_t(f)$ is now preset in units of foreign currency. An unexpected rise in e_t lowers this relative price.

countries) remains constant, leading to the large positive comovement of consumption across countries. Consumption increases in both countries, reflecting the increase in real money balances in the home country and the decline in the consumer price index in the foreign country. At the same time, foreign goods become more expensive relative to home goods and both agents substitute consumption towards home goods and away from foreign goods. Therefore, in response to this expenditure-switching effect, production shifts away from the foreign country to the home country, implying a negative comovement of output across countries.

With pricing-to-market, a positive money shock in the home country is associated with a real exchange rate depreciation, which leads the comovement of consumption across countries to fall. In this case, however, the relative price of home to foreign goods is left unchanged and the elimination of the expenditure switching effect increases the comovement of output across countries.

Implications of Local-Currency Pricing for Two-Country Models

Several recent papers have explored the implications of incomplete short-run exchange rate pass-through for a series of wide-ranging questions in international economics. Since the nature of international pricing has a crucial effect on the international transmission of monetary disturbances, this assumption substantially affects the business-cycle properties of open-economy models, the welfare properties of alternative exchange rate regimes, and the characterization of optimal monetary and exchange rate policies. I now highlight some of these issues.

Chari, Kehoe, and McGrattan (2000) calibrate a stochastic pricing-to-market model and investigate whether the interaction of staggered prices with money shocks can account for the observed behavior of real exchange rates.[25] They show that their model is successful in generating real exchange rates that are as volatile as in data, but not as persistent. Since in a monopolistic competition framework unexpected money shocks do not generate movements in the real exchange rate beyond the periods of (exogenously-imposed) nominal stickiness, this model is not able to generate sufficiently persistent real exchange rates.[26]

---

[25] See Kollmann (1997) for a calibrated small open economy in which both wages and prices are sticky.

[26] See Bergin and Feenstra (2001) for an exploration of the volatility and persistence properties of real exchange rates in a model with translog preferences and intermediate inputs that generates endogenous persistence. In a closed economy setup, Dotsey and King (2001) build a model with structural features that substantially reduce the elasticity of marginal cost with respect to output, generating greater endogenous persistence. The implication of this model's features for the behavior of real exchange rates in a two-country model is still an open question.

The business-cycle properties of different exchange rate regimes are explored in Duarte (2001) in a calibrated pricing-to-market model. Baxter and Stockman (1989) and Flood and Rose (1995) show that, following a change from pegged to floating exchange rate systems, countries with moderate inflations experience a systematic and sharp increase in the variability of the real exchange rate, while the behavior of other macroeconomic variables remains largely unaffected by the change in regime. This puzzling evidence can be accounted for in a model with prices set one period in advance in the local currency of the buyer. By eliminating the expenditure-switching effect of exchange rates in the short run, this model predicts a sharp increase in the volatility of the real exchange rate following a change from fixed to flexible exchange rates, without generating a similar pattern for the volatilities of output, consumption, or trade flows.

Devereux and Engel (1998) compare the welfare properties of fixed and flexible exchange rate systems in an explicitly stochastic setting. Under uncertainty, firms incorporate a risk premium in their pricing decision, which affects the equilibrium prices that are chosen. This effect on equilibrium prices in turn has an impact on expected output and consumption levels and ex-ante welfare levels. Devereux and Engel show that the exchange rate regime influences not only the variance of consumption and output, but also their average values, and that the optimal exchange rate regime depends crucially on the nature of pricing. They find that under producer-currency pricing there is a trade-off between floating and fixed exchange rates, while floating exchange rates always dominate fixed exchange rates under consumer-currency pricing.

The nature of currency pricing also has substantial implications for the welfare effects of monetary policy and international policy coordination. Since consumer import prices do not respond in the short run to changes in the exchange rate, pricing-to-market models predict that unexpected currency depreciations are associated with an *improvement* of the country's terms of trade, rather than with the deterioration that occurs with producer-currency pricing. For example, if the dollar depreciates and consumer prices are sticky (in the local currency), then the dollar price paid in the United States for imported goods remains the same, while the price of American exported goods rises when translated into dollars. Betts and Devereux (2000) show that this effect of domestic monetary expansions on the terms of trade raises domestic welfare at the expense of foreign welfare. That is, expansionary monetary policy is a "beggar-thy-neighbor" instrument. This result contrasts sharply with the prediction from a model with PPP, where a surprise monetary expansion in one country raises welfare in both countries (Obstfeld and Rogoff 1995).

Obstfeld and Rogoff (2000b) argue that the positive relation between exchange rate depreciations and terms of trade implied by pricing-to-market models is at odds with the empirical evidence. They present some evidence supporting the conventional idea that currency depreciations cause the terms

of trade to deteriorate. The role of the degree of exchange rate pass-through in the allocative effect of exchange rate changes and the importance of this mechanism in the properties of open-economy models show that it is crucial to explore the implications of new open-economy models with more realistic pricing assumptions. In particular, it is important to study the implications of models that can distinguish the apparent zero exchange rate pass-through at the consumer level from the clearly positive (but smaller than one) exchange rate pass-through at the producer level. In a recent contribution to the literature, Corsetti and Dedola (2001) introduce labor intensive distribution services in an otherwise standard two-country model with preset wages. They show that the law of one price fails to hold at both producer and consumer levels and that monetary shocks may result in expenditure switching effects.

3. CONCLUDING REMARKS

This article focuses on the implications of alternative international price-setting regimes in open-economy models that incorporate nominal price rigidities and monopolistic competition. Most of the recent research in this field has progressed under the assumption of either producer-currency pricing or consumer-currency pricing. Since the nature of price setting determines the effect of exchange rate changes on the relative price of imported to domestic goods in the short run, the price-setting assumption determines the role of exchange rates in shifting consumer allocation decisions across countries. Therefore, the international monetary transmission mechanism differs markedly under these two alternatives, yielding very different predictions for many substantial issues in international economics.

Assuming that prices are set in advance in the consumer's currency allows for short-run deviations in the law of one price for tradable goods, which occur in response to unexpected changes in the exchange rate. These deviations in turn generate movements in the real exchange rate, as is suggested by recent empirical evidence. Pricing-to-market models have been able to replicate a number of key international business-cycle properties, both for floating exchange rate periods and across alternative regimes.

In pricing-to-market models, exchange rate pass-through to consumer import prices is zero in the short run. This feature of the model implies that exchange rate depreciations and the terms of trade are positively correlated, a relation that is not supported by the data.

While the data suggests that exchange rate pass-through at the consumer level is indeed close to zero, it is clearly positive (but incomplete) at the producer level. Given the crucial role played by the international price-setting regime in the international transmission mechanism of monetary disturbances, it is clearly important to explore the implications of distinct exchange rate pass-throughs at the consumer and producer levels.

REFERENCES

Baxter, Marianne, and Alan Stockman. 1989. "Business Cycles and the Exchange Rate Regime: Some International Evidence." *Journal of Monetary Economics* 23 (May): 377–400.

Bergin, Paul, and Robert Feenstra. 2001. "Pricing-to-Market, Staggered Contracts, and Real Exchange Rate Persistence." *Journal of International Economics* 54 (August): 333–59.

Betts, Caroline, and Michael B. Devereux. 1996. "The Exchange Rate in a Model of Pricing to Market." *European Economic Review* 40 (April): 1007–21.

_____. 2000. "Exchange Rate Dynamics in a Model of Pricing-to-Market." *Journal of International Economics* 50 (February): 215–44.

Chari, V. V., Patrick Kehoe, and Ellen McGrattan. 2000. "Can Sticky Prices Models Generate Volatile and Persistent Real Exchange Rates?" Federal Reserve Bank of Minneapolis Research Department Staff Report 223.

Corsetti, Giancarlo, and Luca Dedola. 2001. "International Price Discrimination and Macroeconomics." Manuscript, Yale University and Banca d'Italia.

Devereux, Michael, and Charles Engel. 1998. "Fixed versus Floating Exchange Rates: How Price Setting Affects the Optimal Choice of Exchange Rate Regime." NBER Working Paper 6867.

_____. 2001. "Endogenous Currency of Price Setting in a Dynamic Open Economy Model." Manuscript, University of Wisconsin, Madison, and University of British Columbia.

Dotsey, Michael, and Robert King. 2001. "Pricing, Production, and Persistence." NBER Working Paper 8407.

Duarte, Margarida. 2000. "Why Don't Macroeconomic Quantities Respond to Exchange Rate Variability? Comparing Fixed and Floating Exchange Rate Systems." Manuscript, Federal Reserve Bank of Richmond.

Dumas, Bernard. 1992. "Dynamic Equilibrium and the Real Exchange Rate in a Spatially Separated World." *The Review of Financial Studies* 5: 153–80.

Engel, Charles. 1993. "Real Exchange Rates and Relative Prices: An Empirical Investigation." *Journal of Monetary Economics* 32 (August): 35-50.

———————. 1999. "Accounting for U.S. Real Exchange Rate Changes." *Journal of Political Economy* 107 (June): 507–38.

———————, and John Rogers. 1996. "How Wide Is the Border?" *American Economic Review* 86 (December): 1112–25.

Flood, Robert, and Andrew Rose. 1995. "Fixing Exchange Rates: A Virtual Quest for Fundamentals." *Journal of Monetary Economics* 36 (December): 3–38.

Frankel, Jeffrey. 1986. "International Capital Mobility and Crowding-out in the U.S. Economy: Imperfect Integration of Financial Markets or of Goods Markets?" In *How Open Is the U.S. Economy?*, ed. R. W. Hafer. Lexington, Mass.: Lexington Books: 33–67.

———————, and Andrew Rose. 1996. "A Panel Project on Purchasing Power Parity: Mean Reversion Within and Between Countries." *Journal of International Economics* 40 (February): 209–24.

Frenkel, Jacob. 1981. "The Collapse of Purchasing Power Parities During the 1970's." *European Economic Review* 16 (May): 145–65.

Froot, Kenneth, and Kenneth Rogoff. 1995. "Perspectives on PPP and Long-Run Real Exchange Rate Rates." In *Handbook of International Economics*, vol. 3, ed. K. Rogoff and G. Grossman.

Giovannini, Alberto. 1988. "Exchange Rates and Traded Goods Prices." *Journal of International Economics* 24 (February): 45–68.

Goldberg, Pinelopi, and Michael Knetter. 1997. "Goods Prices and Exchange Rates: What Have We Learned?" *Journal of Economic Literature* 35 (September): 1243–72.

Humphrey, Thomas, and Robert Keleher. 1982. *The Monetary Approach to the Balance of Payments, Exchange Rates, and World Inflation*. New York: Praeger Publishers.

Isard, Peter. 1977. "How Far Can We Push the 'Law of One Price'?" *American Economic Review* 67 (December): 942–48.

Knetter, Michael. 1989. "Price Discrimination by U.S. and German Exporters." *American Economic Review* 79 (March): 198–210.

———————. 1993. "International Comparisons of Pricing-to-Market Behavior." *American Economic Review* 83 (June): 473–86.

Kollmann, Robert. 1997. "The Exchange Rate in a Dynamic-Optimizing Current Account Model with Nominal Rigidities: A Quantitative Investigation." IMF Working Paper 97/7.

Krugman, Paul. 1978. "Purchasing Power Parity and Exchange Rates: Another Look at the Evidence." *Journal of International Economics* 8 (August): 397–407.

Lane, Philip. 2000. "The New Open Economy Macroeconomics: a Survey." Trinity Economic Paper Series, Paper No. 3.

Lothian, James, and Mark Taylor. 1996. "Real Exchange Rate Behavior: the Recent Float From the Perspective of the Past Two Centuries." *Journal of Political Economy* 104: 488–509.

Michael, Panos, Robert Nobay, and David Peel. 1997. "Transaction Costs and Nonlinear Adjustment in Real Exchange Rates: An Empirical Investigation." *Journal of Political Economy* 105 (June): 862–79.

Mussa, Michael. 1986. "Nominal Exchange Rate Regimes and the Behavior of Real Exchange Rates: Evidence and Implications." *Carnegie-Rochester Conference Series on Public Policy* 25: 117–214.

Obstfeld, Maurice, and Kenneth Rogoff. 1995. "Exchange Rate Dynamics Redux." *Journal of Political Economy* 103 (June): 624–60.

_____. 2000a. "The Six Major Puzzles in International Macroeconomics: Is There a Common Cause?," NBER Working Paper 7777.

_____. 2000b. "New Directions for Stochastic Open Economy Models." *Journal of International Economics* 50 (February): 117–53.

Obstfeld, Maurice, and Alan Taylor. 1997. "Nonlinear Aspects of Goods-Market Arbitrage and Adjustment: Heckscher's Commodity Points Revisited." *Journal of the Japanese and International Economies* 11 (December): 441–79.

Ohanian, Lee, and Alan Stockman. 1997. "Arbitrage Costs and Exchange Rates." Manuscript, University of California, Los Angeles, and University of Rochester.

Rogoff, Kenneth. 1996. "The Purchasing Power Parity Puzzle." *Journal of Economic Literature* 34 (June): 647–68.

Stockman, Alan. 1987. "The Equilibrium Approach to Exchange Rates." Federal Reserve Bank of Richmond *Economic Review* 73 (March/April): 12–30.

Svensson, Lars, and Sweder van Wijnbergen. 1989. "Excess Capacity, Monopolistic Competition, and International Transmission of Monetary Disturbances." *Economic Journal* 99 (September): 785–805.

Taylor, Alan. 2001. "Potential Pitfalls for the Purchasing-Power-Parity Puzzle? Sampling and Specification Biases in Mean-Reversion Tests of the Law of One Price." *Econometrica* 69 (March): 473–98.

Taylor, Mark, David Peel, and Lucio Sarno. 2001. "Nonlinear Mean-Reversion in Real Exchange Rates: Toward a Solution to the Purchasing Power Parity Puzzles." *International Economic Review* 42 (4): 1015–42.

Should Banks Be Recapitalized?

Douglas W. Diamond

When a nation's banks experience major losses, depositors, the markets, and regulators respond. The market responds by making it difficult for the bank to raise funds. Depositors may rush to withdraw funds from the banks. The regulators respond by closing banks, guaranteeing their liabilities, or recapitalizing them. One or more of these outcomes is inevitable. This article studies the effects of the regulatory choice on various parties in the economy.

The most obvious choice that regulators make is whether to let banks fail. Does their inability to raise sufficient private capital indicate that they are not viable or produce future services that are worth less than their cost, and thus should be closed? Only if the government, depositors, and borrowers were first allowed to jointly renegotiate would the inability to restructure indicate that the banks are not viable. This article analyzes the effects and desirability of recapitalizing banks with public funds, with a brief discussion of the implications of recapitalization for the current situation in Japan.

In many countries, including Japan, there is a very deep government safety net and substantial regulation (see Ito and Sasaki [1998] and Hogarth and Thomas [1999] for discussions of bank capital structure in Japan). So one approach would be to ignore the markets and analyze bank recapitalization as a bargaining situation between banks and regulators. However, there is

The author is Merton H. Miller Distinguished Service Professor of Finance, University of Chicago, Graduate School of Business, and consultant to the Research Department of the Federal Reserve Bank of Richmond. Any opinions expressed are those of the author and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. This is a revised version of "Should Japanese Banks Be Recapitalized," originally published in the May 2001 issue of *Monetary and Economic Studies* (Bank of Japan). This research was conducted while I was a Visiting Scholar at the Institute for Monetary and Economic Studies of the Bank of Japan. I am very grateful for many helpful comments on this topic. I wish to thank Takayoshi Hatayama, Tom Humphrey, Hiroshi Nakaso, Nobuyuki Oda, Ned Prescott, Raghu Rajan, Martin Schulz, Yoshinori Shimizu, Masaaki Shirakawa, John Walter, John Weinberg, and Tatsuya Yonetani for this help.

legislation in Japan that will limit deposit insurance (see Nakaso [1999]) and require prompt corrective action from undercapitalized banks, as is now required in the United States. Around the world, the discipline of banks relies to some extent on market incentives. As a result, it is important to study the effects of bank capital on how much banks will be able to raise in the market. Even with total deposit insurance, the banks will need to consider the effects of their credit rating on the other lines of business they can provide. If the level of capital is below the minimum necessary to stay in business (and this minimum will actually be enforced), then banks will need to do whatever it takes to increase their capital to the minimum. This "whatever it takes" type of bank behavior could have undesired effects on the economy.

I focus on the effect of bank recapitalization on banks and their existing borrowers. The effect on future borrowers (new business development) is ignored on the basis that new banks, other recapitalized banks, or even foreign banks could provide such new *relationship-based* funding without a subsidized recapitalization of the majority of existing banks. Recapitalizing a large number of banks is desirable only if it protects the value of existing relationship lending and human capital in banks and firms. If the reason to have a well-capitalized banking system is to ensure that new relationships can be established, this can be achieved by recapitalizing a few of the best banks. The analysis here points out that the recapitalization, and its extent, can result in transfers between banks and borrowing firms that can go in either direction. This result occurs because bank capital influences the bargaining between a bank and its borrowers. In addition, recapitalization can have efficiency effects by influencing a bank's decision whether to foreclose on its defaulted loans.

The amount of current bank capital affects the behavior of a bank when it is required to have a minimum amount of capital in order to remain in business. The same effect occurs when the threat of closure due to low capital comes from market participants who may not provide capital or from potentially uninsured depositors who may withdraw deposits, as in Diamond and Rajan (2000a), summarized in Diamond and Rajan (2001c).[1]

The remainder of the article has the following structure. Section 1 outlines the basic argument, without technical details. Section 2 discusses the effects of a bank's capital on its behavior. Section 3 discusses the effect of bank capital on the way that banks treat their borrowers and on the endogenous payments made by borrowers. Section 4 discusses the policy choice tradeoffs in choosing how much capital to provide. Section 5 argues that banks without lending relationships and those with nonviable borrowers should not be recapitalized. Section 6 concludes the article.

[1] In addition, see Diamond and Rajan (2000b) for an extension to understand the role of short-term debt in the East Asian financial crisis of 1997.

1. A SKETCH OF THE REASONS FOR AND AGAINST RECAPITALIZATION

The effect of bank capital on bank behavior and borrower welfare depends on certain characteristics of the borrower and of the bank. The relevant characteristic of the bank is the presence or absence of relationship lending. Relationship lending implies that the lender has a special skill in evaluating a borrower or in committing to providing a long-term financing policy that a new lender cannot provide. Many discussions of the *keiretsu* system in Japan stress the importance of relationship lending (see Hoshi, Kashyap, and Scharfstein [1991], Aoki, Patrick, and Sheard [1994], and Hoshi and Patrick [2000]). One expects that relationship lending is most important for loans to firms rather than to consumers and when the anticipated response to a potential default is renegotiation rather than immediate foreclosure of collateral.

Applying the model introduced in Diamond and Rajan (2000, 2001a), I define a relationship lender as one whose knowledge allows it to induce the borrower to make larger future payments. As a result, a relationship lender can lend more today than other lenders and is less inclined to foreclose on a loan because it can collect more in the future. However, if the relationship lender is in financial trouble, it may be unable to provide these larger loans or loan extensions. The relationship lender's special loan collection skill makes loans illiquid and hard to sell or borrow against. If there is not relationship lending, then a bank's financial situation has no effect on the borrower. Another lender can replace an undercapitalized bank, and the undercapitalized bank can either sell the loan or accept a payment that the borrower raises from borrowing elsewhere. Only when relationship lending is important is the financial health of particular bank lenders of critical importance to their borrowers and to the economy as a whole.

The characteristics of a bank's borrowers also partly determine the effect and desirability of providing subsidized capital to a distressed bank. The relevant borrower characteristic is the viability of its business. A business is viable if it can commit to paying the relationship lender more (in present value) than the lender can raise by foreclosing today. A viable borrower should not lose access to credit, and it will not lose its access to credit from its bank if the bank is well capitalized. A nonviable borrower should lose access to credit, and in many cases a bank will cut off credit to such a borrower independent of its capital position. I argue that the only case where a subsidized recapitalization may be justified is when the undercapitalized bank is one with lending relationships and viable borrowers. In all other cases, recapitalization is a government subsidy without social value. Table 1 summarizes the results. A more detailed version of this table is presented in Table 2 in the conclusion.

Table 1 Desirable and Undesirable Forms of Recapitalization

	Financially distressed bank with a relationship borrower	**Financially distressed bank without a relationship borrower**
Borrower has the best use of the collateral (and is thus viable)	Main case analyzed. Subsidized capital may be socially desirable.	No reason to recapitalize.
	A very small recapitalization may be worse than no recapitalization at all.	No effect on borrowers of too small a recapitalization.
Borrower does not have the best use of the collateral (and is thus not viable)	No reason to recapitalize unless banks are reluctant to foreclose due to effect on accounting bank capital.	No reason to recapitalize.
	A very small recapitalization just sufficient to avoid this reluctance is a good policy.	No effect on borrowers of too small a recapitalization.

Relationship Lending

A bank with a valuable lending relationship can induce its borrowers to make larger payments than other lenders. The relationship lender has what I call a specific loan collection skill. If a bank has specific loan collection skills, other lenders can collect only a fraction of collectable future loan proceeds (see Diamond and Rajan [2001a]). As a result, the bank's relationship-based loans are illiquid. In addition, this source of illiquidity makes it more difficult for the bank to raise capital than deposits. It turns out that only a fraction of the present value of future relationship-based loan collections is capitalized in the market prices of the bank's non-deposit capital. The higher the capital ratio, the greater the rents absorbed by the bank. The results on relationship borrowers may apply to *keiretsu* loans based on long-standing relationships. The results do not apply, for example, to simple real estate mortgage loans, where repayment incentives come only from the threat of sale in the market of the real estate collateral.[2]

[2] These are nonrelationship loans, discussed in Section 5.

Effects of Bank Capital on Bank Behavior

A bank's capital structure directly influences its ability to raise funding for its relationship loans. Because higher capital implies higher rents to bankers, a high level of required capital reduces the sum of the values of deposits plus capital that a bank can raise from outside investors. Such a limitation on a bank's ability to fund its loans can indirectly influence its behavior toward borrowers—a bank that cannot raise sufficient capital may limit its ability to make or renew loans to its relationship borrowers.

Consider a bank that has developed a lending relationship with a viable borrower. Results in Diamond and Rajan (2000) show that the level of capital influences the horizon over which a relationship lender will operate when a borrower's loans are risky. A well-capitalized bank will operate with a long horizon, while an undercapitalized capital bank will be forced to try to immediately meet its capital requirement. If a bank can get a larger immediate payment by forcing foreclosure, it may have to do so even if it yields a smaller present value than would allowing a borrower more time to pay. An undercapitalized bank will be unwilling to wait to collect loans over the long run. It may liquidate the borrower's collateral when a better-capitalized bank would let the borrower continue to operate. In addition, because it is prone to liquidate, an undercapitalized bank may be able to extract very large payments from its relationship borrowers. In effect, such a bank conducts an auction for the right not to be liquidated.

An undercapitalized bank's incentive to liquidate comes from its need to reduce its portfolio of illiquid loans. This will satisfy a capital requirement imposed by the market: for example, the need to avoid the threat of a run by depositors. If the capital requirement is imposed by regulators and is based on regulatory book capital, then an offsetting effect may dominate. Even if foreclosure produces a larger present value than extending the loan, it may lead to a loss relative to the book value of the loan. For very low levels of book capital, relevant to some banks in Japan, the bank would not foreclose or accept a partial payment because it would cause a write down in book capital that would lead the bank to be closed. In this case, the bank would not foreclose on any loans. I defer discussion of this "evergreening" effect (where the loan is like a tree that is green even when frozen in the dead of winter) until the analysis with market value accounting is complete.

The effects of bank capital identified here are on banks with relationship loans to viable borrowers. This approach implies that banks without such loans should be allowed to fail. The explicit discussion of this case is deferred to Section 5, after I have provided further details regarding the types of recapitalization that may be in the public interest.

2. FOUNDATIONS FOR THE LINK BETWEEN RELATIONSHIPS, ILLIQUIDITY, AND BANK CAPITAL

I consider a bank with a collateralized loan to a single representative borrower. There are three dates, 0, 1, and 2, and riskless interest rates are zero. There are three types of agent: a borrower who needs funds for a project, a banker who is a relationship lender and has special skills in collecting loans from the borrower, and an outside investor who has no loan collection skills. Outside investors can hold deposits or non-deposit capital issued by the bank. They can hold loans, but they have no skill in collecting the loans.

The borrower has substantial bargaining power with the bank, and can make take-it-or-leave-it offers to reschedule payments to the bank. As a result, the bank cannot force the borrower to pay more than the value for which it can liquidate the collateral. This is assumed only for simplicity. So long as the amount that a lender can collect is an increasing function of the value the lender obtains from liquidation, qualitatively similar results will follow. The next section describes the bank's negotiations with the borrower.

Negotiations between the Bank and the Borrower

In Section 3, I examine the effects of the bank's financial position on its dealings with the borrower. It is useful here to describe the dealings between bank and borrower when there is no such constraint and the banker is negotiating unconstrained (as if negotiating for his or her own personal account). As in Hart and Moore (1994), I consider financial contracts that specify that the borrower owns the machinery and has to make a payment to the banker, failing which the banker will get possession of the collateral and the right to use it as he or she pleases. So a contract specifies repayments P_t (for dates $t = 1$ and $t = 2$) that the borrower is required to make at date t, as well as the assets the bank may liquidate if the borrower defaults. Using some notation that I will not use again until Section 3, the bank can liquidate the collateral for $X_1 \geq 0$ at date 1 or $X_2 \geq X_1 \geq 0$ at date 2. The borrower's project produces cash of $C_1 \geq 0$ at date 1 and $C_2 \geq X_2 \geq 0$ at date 2 if not liquidated on or before these dates. The borrower has cash before date 1 production of $C_0 \geq 0$.

The source of friction in the model is that any agent can commit explicitly to contributing specific skills to a specific venture only in the spot market—not in advance, but just before production is to occur. As a result, just before production the borrower may attempt to renegotiate the terms of the loan that was agreed to in the past, using the threat of withholding his or her human capital from production this period (and the promise of committing to produce now if a new agreement is reached). Without the borrower's human capital, no current cash flow is produced (apart from the value of liquidation).

Bargaining between bank and borrower just before the borrower is due to produce takes the following form; the borrower offers an alternative payment

to the one contracted in the past and commits to contribute his or her human capital if the offer is accepted (and not to contribute it if rejected). The banker can (1) accept the offer, (2) reject the offer and choose to liquidate the project immediately, or (3) if the bargaining occurs on or before date 1, reject the offer (implying that the borrower does not produce this period), retaining the option to liquidate at date 2. The game gives all the bargaining power to the borrower, apart from the banker's ability to exercise control rights to liquidate. If the borrower's offer is accepted, the borrower contributes his or her human capital, and the offered payment is made.

Example 1 *Suppose that it is just before date 2, and the borrower promised to pay* $P_2 = C_2$. *The borrower knows the banker can obtain* $X_2 < C_2$ *by liquidating the collateral. As a result, the borrower offers to pay only* X_2 *and the banker,who cannot do any better by refusing, accepts. Note that lenders other than the bank would have no ability to liquidate the project for a positive amount. As a result, they would not be able to enforce any repayment. The banker's specific skills enable him or her to collect more, so I will refer to these skills as collection skills. In this example, in order to be collectable using the bank's threat to liquidate for* X_2, *a contract must specify a promised payment* $P_2 \leq X_2$.

Until Section 3, I will assume that all payments specified in the loan contracts can be collected by the bank using its threat to liquidate the collateral, and I will not further analyze the negotiation between bank and borrower. This allows study of the bank's ability to fund itself and to satisfy capital requirements before examining the effect on the bank's actions toward borrowers.

Relationship Lending

When the bank is a relationship lender, it is the only lender that can force the borrower to repay the maximum value. Other lenders can collect less. For simplicity only, I assume that other lenders would collect zero if they attempted to collect the loan (all results follow when other lenders could collect a positive but smaller amount than that collected by the bank). As a result, a loan would be worthless without some access to the bank's loan collection skills.

A relationship lender cannot raise the full present value that it can collect from the borrower by issuing capital (i.e., non-demandable claims) today. This is because the relationship lender's specific skills are needed to extract repayment from the borrower. The only sanction available to outside capital holders is to dismiss the bank and replace it with one that cannot collect anything from the borrower. So, the original relationship lender can, and will, appropriate a rent for its specific skills. For application to banks with many employees, one can interpret the relationship lender's rent as excessive employment of bankers. Assuming that, in bargaining, the relationship lender

extracts half the additional amount recovered from the borrower, it will keep a rent of one-half and only pass on the other half to outside holders of capital.[3]

The implications of the banker's ability to negotiate for rent from outside investors are best illustrated by an example. Suppose that the relationship lender can collect P_2 from the borrower and has raised exclusively non-deposit capital from outside investors. If the banker threatens to quit and not to collect the loan, both the banker and outside capital holders get zero. If the banker collects the loan, however, their total surplus is P_2. I assume that in bargaining, they divide surplus equally. As a result, outside investors who hold capital know that the banker will renegotiate if he or she promises to pay them more than $\frac{1}{2}P_2$. Thus the most that the banker can raise in non-deposit capital is $\frac{1}{2}P_2$.

The relationship lender can sell the loan or issue capital against it for only a fraction of present value of the payments that it can collect. If there were no relationship, and anyone could collect the full amount of the loan, it would be liquid: the bank could issue capital up to the full value of the loan or sell it for the full amount. With such a liquid loan, outside capital holders would replace the banker or sell the loan unless the banker's rent was zero, and the banker would not be able to threaten to earn a rent.

Discipline from the Threat of a Bank Run

Suppose instead that the banker finances illiquid loans by issuing uninsured demand deposits. These cannot be renegotiated next period without triggering a run, which removes the loan from the banker's control (see Diamond and Rajan [2001a]). Because of the "first come, first served" aspect of uninsured demand deposits, no depositor would want to make a concession if the bank still had assets. Each depositor could force the bank to sell assets to pay in full (until the bank runs out of assets). And once the loan is sold, the banker can earn no rents. The banker will always pay deposits if feasible. If the level of deposits and capital is set when it is known that the banker can collect exactly P_2 from a borrower, the problem with a riskless loan's illiquidity can be solved: set deposits equal to P_2 and capital equal to zero. The banker will pay out the full P_2.

When the bank's capital structure combines deposits and capital, the bank's ability to commit to pay outside investors is in between an all-capital

[3] Note that unlike in the bargaining between the borrower and the banker, where the latter has no bargaining power, here outside investors have some bargaining power. In practice, we would typically have interior amounts of bargaining power in both situations. I assume the borrower has all the bargaining power in negotiations with the banker only to simplify notation.

bank and an all-deposit bank. Returning to Example 1 illustrates the commitment ability of a bank with deposits less than the value of the amount it can collect with its relationship skills.

Example 1 (continued) *Suppose that the banker can collect P_2 from the borrower and has $D_2 < P_2$ deposits outstanding, with outside investors holding 100 percent of the residual equity capital claim on the bank. Without the banker, capital holders are able to collect nothing from the borrower. So capital holders will not be able to avoid a run if the banker quits, and will get zero. The net amount of surplus available to capital and the banker if the bank does use its skills to collect the loan is $P_2 - D$. Since neither can get any of the surplus without the other's cooperation, and because they divide the surplus equally, each gets $\frac{1}{2}(P_2 - D_2)$. As a result, $\frac{1}{2}(P_2 - D_2)$ will be absorbed by the banker as rent. The remainder, $\frac{1}{2}(P_2 - D_2)$ is the maximum value of outside equity capital on the bank. When added to the D_2 that the bank can commit to pay to depositors, the total that the bank can commit to pay to outside investors is $\frac{1}{2}(P_2 - D_2) + D_2 = \frac{1}{2}(P_2 + D_2)$.*

The problem with capital is that it does not provide the banker as hard a budget constraint as demand deposits. The higher the capital-to-deposit ratio, the higher the rent collected by the banker. However, when loans are risky, a positive level of capital is needed to avoid the costs of a high probability of bank failure. With a positive level of capital needed, the illiquidity problem will remain. The problem is that demand deposits are a very rigid form of financing. This is good in that it disciplines the banker and enables him to commit to pay out. It is bad if there is sufficient uncertainty in bank asset values because a drop in bank asset values will precipitate a run, disintermediating the banker, and further reducing their value. Capital can act as a buffer in such cases because, unlike deposits, its value adjusts to underlying asset values. If there is a reduction in the amount that the bank's borrowers can pay, the bank can raise additional capital so long as it can commit to giving a normal rate of return to investors. However, only a fraction of the amount that the banker can collect on the loan can be committed to pay to outside holders of capital.

Rather than introduce uncertainty that leads to the need for some capital, I will instead look at the effects of using some capital to fund the bank under certainty. This will illustrate the qualitative effects of bank capital on bank behavior. Specifically, when there is uncertainty, Diamond and Rajan (2000) show that the optimal capital structure for the bank may involve some capital in addition to demand deposits. In the rest of the article, I will assume there is a capital requirement for banks, specified by regulatory authorities.

The capital requirement is as follows. A bank can raise new capital at any time, but will be closed if the market value of its capital falls below a fraction γ of the market value of its capital plus the par value of its deposits. This

capital requirement limits the amount that the bank can raise, as illustrated in the following example.

Example 1 (continued) *If a bank has a relationship loan that pays P_2, and if its capital requirement is just met, it must be that $\gamma = \frac{\frac{1}{2}(P_2 - D_2)}{\frac{1}{2}(P_2 + D_2)}$ where the numerator on the right hand side is the date 2 value of capital, and the denominator is the value of capital plus deposits (the market value of external claims on the bank). This implies that date 1 deposits satisfy $D_2 = \frac{(1-\gamma)}{1+\gamma} P_2$. Therefore, the total amount that can be pledged to investors at date 1 out of the amount the bank collects from borrowers at date 2 is $\frac{1}{2}(P_2 + D_2)$, which, on substituting for D_2, works out to $\frac{P_2}{1+\gamma}$. Since the total amount paid by the borrower is P_2, the bank absorbs $\frac{\gamma}{1+\gamma} P_2$ in rent, an amount increasing in γ. More generally, we will see that only a fraction $\frac{1}{(1+\gamma)}$ of the total date-t value of the bank can be pledged to outsiders at date $t-1$. The banker absorbs the remaining amount as rent because the capital requirements impose the constraint that $D_2 \leq \frac{(1-\gamma)}{1+\gamma} P_2$.*

Discipline from the Threat of Closure due to Capital Requirements

An effect similar to the threat of runs occurs with insured deposits if deposits are insured and the deposit insurer requires prompt corrective action to enforce a minimum level of capital (and sticks by this threat to close the bank unless it raises sufficient capital in the market). When the deposit insurer and the remainder of the government are prohibited from providing subsidized capital to the bank, the bank is under the same incentives as the threat of a run, and rents are an increasing function of the amount of capital required. Consider a bank with a given level of capital. If it incurs losses beyond a given amount, its uninsured depositors will run, closing the bank. If the same loss leads regulators to close the bank, then the incentives are identical.

Suppose that the bank is closed if the market value of the capital that it has or it raises is below a fraction γ of the market value of its total external liabilities (market value of capital plus deposits valued at par). The market value (the maximum value that can be pledged to outside investors other than the banker) of its total external liabilities will be less than the total present value of what the bank can collect on its loans (plus any other assets) because the relationship loan collection skill makes its loans illiquid. All of the valuation is as if the commitment came from the threat of a run on uninsured deposits: the maximum total that a bank could commit to pay to outsiders (deposits plus capital) is again $\frac{P_2}{1+\gamma}$ if the bank must meet its capital requirement.

Enforced minimum capital requirements make insured deposits a hard "budget constraint" on bankers by committing the deposit insurer not to allow excess rents to the bankers. An all-capital structure provides no discipline

because there is no threat of closure, but once there are some deposits, a required level of capital provides discipline by forcing closure if the bank's total value paid to outside claimants falls sufficiently. Although this is consistent with other views of minimum capital requirements as providing discipline to bankers by committing regulators to close insolvent banks, it provides a somewhat different perspective. If the level of capital above the minimum is too much above the minimum level, the banker will be free to appropriate rents and excessive costs from capital, to the extent that the bankers provide a loan collection service not available elsewhere. Excess capital only influences the rents of banks that do relationship lending when capital owners are free to replace bankers with poor lending performance.

Without a minimum capital requirement, the regulator can allow the bank to operate with negative capital and to raise additional insured deposits to cover excessive costs. As a result, the deposit insurer could in principle give an unlimited subsidy to banks. Such a deposit insurer would be forced to make as large a concession as an all-capital bank (and probably would make an even larger concession).

With a minimum capital requirement, rents are limited. If the deposit insurer must close the bank if capital is too low, and cannot provide capital of its own, then there is no negotiation with the deposit insurer that will yield the bank a larger concession than just negotiating with capital holders. Negotiations must then be with capital holders. Capital holders will make concessions, but not the depositors or their insurer. The value that can go to outsiders as a whole is the value of deposits plus one-half the excess over this amount that the banker can collect. If the deposits exceed what the banker can collect, then the bank fails, and the borrower pays the deposit insurer one-half the amount that the banker could collect (if less was paid, the deposit insurer would hire the banker to collect for a fee of one-half the amount collected), and the deposit insurer covers the rest. Notice that at date 2 (representing the long run) I assume that the borrower has this much cash. I make no such assumption about date 1 in Section 3.

Capital Value over Two Periods

Consider a relationship loan with payments P_1 at date 1 and P_2 at date 2. Suppose that the banker can actually collect these amounts (the borrower has this much cash at each date and the bank can force the borrower to pay this much). No other lender can force the borrower to pay (it can collect only zero).

We showed above that if the bank is to meet its date 2 capital requirement, the maximum date 2 market value of claims (deposits plus capital) on the date 2 part of the claim, P_2, is $\frac{P_2}{1+\gamma}$. This limit is imposed by the banker's ability to threaten to quit just before date 2. If date 1 maturing deposits minus date 1 loan

payments received were to exceed $\frac{P_2}{1+\gamma}$, the bank would have no way to pay them all, and the bank would be closed due to insolvency (negative capital). If instead maturing date 1 deposits minus date 1 loan payments received were less than or equal to $\frac{P_2}{1+\gamma}$, then the bank would be able to issue new deposits and sufficient additional capital to survive.

At date 1, the bank could potentially pay up to $P_1 + \frac{P_2}{1+\gamma}$ to outside investors (depositors plus holders of capital) by collecting P_1 and issuing claims worth $\frac{P_2}{1+\gamma}$. Because the bank can threaten to quit just before date 1, the amount that the bank will be able to commit to pay to outsiders at date 1 is less than this. Suppose that before date 1, the bank has date 1 demand deposits of D_1, and the banker threatens to quit and not represent the capital holders this period to collect P_1 from the borrower. If not collecting the loan at date 1 breaks the relationship (eliminates the specific loan collection skill), a capital holder who does not reach an agreement to keep the banker would get zero at date 2 as well because the capital holder would be unable to hire the banker to collect the loan at date 2. Alternatively, if the relationship is maintained at date 2 even if the banker does not collect the loan at date 1, but the lost value from the bank not collecting the loan at date 1 implies that the bank is closed immediately due to low capital, then both the capital holder and the banker get zero unless they reach an agreement at date 1 (both capital and banker have an "outside option" to go it alone that is worth zero).[4] The total surplus available to the banker and the capital holders from reaching an agreement is $P_1 - D_1 + \frac{P_2}{1+\gamma}$: this is the value of collecting P_1, raising $\frac{P_2}{1+\gamma}$ with new deposits and capital and repaying the maturing deposits of D_1. Because capital and the banker divide the surplus equally, the value of capital before date 1 is $\frac{1}{2}\left(P_1 + \frac{P_2}{1+\gamma} - D_1\right)$ and the value of capital plus deposits is $\frac{1}{2}\left(P_1 + \frac{P_2}{1+\gamma} + D_1\right)$.

The bank must also meet its minimum capital requirement before date 1 (or it will be closed). To meet the capital requirement before date 1, the ratio of the value of capital to capital plus deposits must not exceed γ, that is, $\gamma \geq \frac{\frac{1}{2}\left(P_1+\frac{P_2}{1+\gamma}-D_1\right)}{\frac{1}{2}\left(P_1+\frac{P_2}{1+\gamma}+D_1\right)}$. In terms of D_1, this result implies that date 1 deposits must satisfy $D_1 \leq \frac{P_1(1-\gamma^2)+P_2(1-\gamma)}{(1+\gamma)^2}$ or the bank will not be able to meet its capital requirement before date 1. If date 1 deposits exceed this amount, the bank will be closed because it cannot recapitalize itself. Substituting for D_1, this result implies that the maximum market value of date 1 capital plus deposits ($\frac{1}{2}\left(P_1 + \frac{P_2}{1+\gamma} + D_1\right)$) is at most equal to $\frac{P_1}{1+\gamma} + \frac{P_2}{(1+\gamma)^2}$. More distant payments are less reflected in capital value because they give more bargaining power to

[4] For a very high level of initial capital ($D_1 \leq \frac{X_2(1-\gamma)}{(1+\gamma)^2}$), a bank meets the minimum before date 1, even if it collects nothing at date 1. This turns out to imply that the bank's capital level does not constrain its loan collection ability (see Appendix).

the banker. For $\gamma = \frac{1}{9}$, the capital requirement before date 1 is satisfied if and only if $D_1 \leq 0.8P_1 + 0.72P_2$ and the maximum market value of external liabilities (deposits plus capital) before date 1 is $0.9P_1 + 0.81P_2$.

3. ENDOGENOUS PAYMENTS AND BANK FORECLOSURE

The analysis of minimum capital requirements to this point has taken the payments from the borrower as given and determined whether the bank will remain open. The borrower's cash holdings on each date, the constraints imposed by minimum capital on the banker's ability to respond to default, and bank's control rights (i.e., the right to call the loan and foreclose absent a current default) are all important.

We examine the ex-post capital position of banks and the ex-post financial position of borrowers in order to examine the effects of capital on banks, borrowers, and depositors and holders of bank capital. The ex-post positions are presumably realizations of uncertain ex-ante prospects, but we look only ex-post. All analysis occurs on a date before date 1 because date 1 and 2 cash flows and liquidation values become known on a date before date 1, and the capital requirement must be met on that date. The market prices of the bank's capital will reflect the position of the borrower, any new capital or deposits issued, anticipations of future such issues, and the outcome of any negotiations between the borrower and banker.

If the bank has no liquidation rights over the borrower absent default, then obviously the borrower will pay no larger amounts than the contracted amounts, P_1 and P_2. An undercapitalized bank must close. But the borrower may be unable to make these payments if short of cash—for example, if the cash on date 1 is less than P_1. In addition, the borrower may choose not to pay over all of his or her cash, anticipating that the bank will accept less and not foreclose.

An Unconstrained Bank's Loan Collection Ability

A borrower's project produces a cash flow of $C_t > 0$ if it is not liquidated before date t (where $t = 1$ or $t = 2$) and if the borrower has initial cash $C_0 \geq 0$ that the lender cannot seize, but which the borrower may use to pay his or her loan. The relationship lender can obtain a liquidation value of X_t just before date t. Suppose that the capital requirements do not influence the relationship lender's behavior toward the borrower. We determine through backward induction how payments will be renegotiated over time if the borrower defaults. The borrower's effort and skills are needed to operate the borrower's firm. I assume that the borrower can credibly threaten not to produce that period's cash, at either date 1 or date 2, unless the bank makes a concession. Suppose at date 2 that the borrower defaults and refuses to make the pre-specified payment P_2

but instead makes an offer of a lower payment. Once the borrower defaults, the lender has the right to liquidate. If the bank rejects the offer and does not liquidate, no cash will be produced at all. In response, the relationship lender can accept the offer or reject it and liquidate the assets to obtain X_2. Thus, if P_2 exceeds X_2 the borrower will renegotiate. At date 2, the borrower will pay min $[P_2, X_2]$.

Now consider what happens at date 1. Suppose that the borrower at date 1 threatens not to produce that period's cash unless the bank makes a concession (offering a lower payment). If the borrower makes this threat and offers a lower payment, the lender can accept the offer. Alternatively, the lender can either reject it and liquidate immediately and get X_1, or reject it and hold on to the asset and get X_2 at date 2. In this last case, no date 1 cash is produced, but the lender gets X_2 at date 2. Thus, the lender will accept any offer to renegotiate that makes its payments amount to max$[X_1, X_2]$ over dates 1 and 2, where any payment left for date 2 should be enforceable, i.e., should be less than X_2. If the sum of promised payments $P_1 + P_2$ exceeds max$[X_1, X_2]$, they will be renegotiated down to this level. If the borrower is short of cash, and can commit to pay less than max$[X_1, X_2]$, the lender will liquidate. Because $X_1 < X_2$, an unconstrained bank can collect a loan worth up to X_2.

A Capital-Constrained Bank's Loan Collection Ability

When the bank lender must meet its capital requirement, it can constrain the bank's ability to follow the unconstrained loan negotiation policy. This is important for two reasons. A constrained negotiation policy will affect the outcomes of forced defaults by borrowers who have less cash than they owe. If the bank's capital constraint weakens its bargaining position, then the borrower will enter negotiations to get a reduction in the amount to be paid even if immediate default is avoidable.

I now consider negotiations that occur before date 1. I begin by assuming that the loan is in default, either because the borrower has missed a promised payment or because the bank has the right to call the loan and liquidate the collateral at any time. If the borrower threatens not to produce the cash C_1 before date 1, and makes an offer that the bank turns down, the bank can either

1. get X_1 by liquidating on or before date 1 and get nothing at date 2; this choice implies a value of the bank and bank capital that is equivalent to a collectable loan that pays nothing after date 1, pays $P_1 = X_1$ and $P_2 = 0$, or

2. get X_2 by liquidating at date 2 and get nothing at date 1 (because the borrower does not supply human capital); this choice implies a value of the bank and bank capital that is equivalent to a collectable loan that pays nothing before date 1, $P_1 = 0$ and $P_2 = X_2$.

If the bank would be closed before date 1 under the second option, it does not have the freedom to wait to reject a borrower's offer and to collect X_2 by date 2 liquidation. Thus, an undercapitalized bank may have a short horizon and be forced to ignore its ability to wait to collect a defaulted loan, potentially weakening its bargaining power over the borrower whenever immediate liquidation is less profitable than delayed liquidation ($X_1 < X_2$). The bank will have a short horizon if it can survive with immediate foreclosure or $P_1 + X_1$ and $P_2 = 0$, but not with an excused default with deferred liquidation or $P_1 = 0$ and $P_2 = X_2$.

A bank with enough capital so that it is free to reject an offer and wait to collect at date 2, or $D_1 \leq \frac{X_2(1-\gamma)}{(1+\gamma)^2} = 0.72$ will be called *well-capitalized*. A bank that is not free to reject an offer and wait to collect at date 2, or $D_1 > \frac{X_2(1-\gamma)}{(1+\gamma)^2} = 0.72$, will be called *undercapitalized.*

If deposits before date 1 are so high that the bank's loan collection skills at date 1 and date 2 are insufficient to collect enough to allow the bank to survive, or $D_1 > \max[X_1, \frac{X_2(1-\gamma)}{(1+\gamma)^2}] = 0.99$, then the bank is termed *severely undercapitalized.*

In addition to limiting a bank's ability to wait to foreclose after it rejects a borrower's offer of partial payment, low capital can limit the types of offers that the bank can choose to accept as an inducement to abstain from liquidation of the borrower. To meet the capital requirement on a date before date 1, I showed above that the borrower must offer collectable date 1 and 2 payments of P_1 and P_2 respectively, plus possibly an immediate payment of P_0 financed out of the borrower's initial cash (C_0), such that $\frac{P_1(1-\gamma^2)+P_2(1-\gamma)}{(1+\gamma)^2} \geq D_1 - P_0$. Otherwise the bank would have to close if it accepted the borrower's offer. If the bank can survive by one of its liquidation options, then an unacceptable offer will be followed by liquidation. If neither liquidation option allows the bank to survive, then the borrower will watch the bank fail if it makes a low offer.

An acceptable offer before date 1 to a well-capitalized bank must satisfy

$$\frac{P_1(1-\gamma^2) + P_2(1-\gamma)}{(1+\gamma)^2} \geq D_1 - P_0$$

$$P_0 + P_1 + P_2 \geq X_1,\ P_0 \leq C_0,\ P_1 \leq C_1,\ P_2 \leq X_2 \leq C_2,$$

and

$$P_0 + P_1 + P_2 \geq X_2.$$

Only the last constraint is binding because $X_2 > X_1$ and $X_2 \leq C_2$, and the well-capitalized bank will collect a total of X_2.

An acceptable offer to an undercapitalized bank must satisfy

$$\frac{P_1(1-\gamma^2)+P_2(1-\gamma)}{(1-\gamma)^2} \geq D_1 - P_0,$$

$$P_0 \leq C_0,\ P_1 \leq C_1,\ P_2 \leq X_2 \leq C_2,$$

and

$$P_0 + P_1 + P_2 \geq X_1.$$

The offer has two possible effects. When the first constraint is binding (which requires one of the cash constraints to bind), then the borrower needs to pay a total sum of payments exceeding X_1 and can be forced to offer very large payments. Offers of lower payments would lead the bank to foreclose to maintain its capital requirement. When the last constraint is binding, then the borrower can get away with paying only X_1, despite the bank's ability to liquidate for more at date 2. When there is no solution, then the borrower must face liquidation or the bank must fail.

The level of initial capital, a decreasing function of D_1, determines how the bank will respond to a default. Suppose that the borrower has defaulted on the original deal, and the bank has the right to foreclose. What offers can the bank accept, and how much can the bank force the borrower to pay? The example below will illustrate this point.

Example

Assume that the capital requirement is $\gamma = \frac{1}{9}$, that $X_1 = 0.99$, and that $X_2 = 1$. A well-capitalized bank (with $D_1 \leq \frac{X_2(1-\gamma)}{(1+\gamma)^2} = 0.72$) would receive and accept an offer of $P_0+P_1+P_2 = 1$ (for example, $P_2 = 1$ and $P_0 = P_1 = 0$) and would be able to collect the date 2 payment.

Consider an undercapitalized bank with $D_1 = 0.8 > \frac{X_2(1-\gamma)}{(1+\gamma)^2} = 0.72$. The borrower has defaulted and will make an offered set of payments before date 1. If the bank rejects the borrower's offer, it cannot wait until date 2 to foreclose, but it can survive by date 1 foreclosure because $D_1 \leq \frac{X_1}{1+\gamma} = 0.8$ (the bank is not severely undercapitalized).

The bank would like the largest total payment ($P_0 + P_1 + P_2$), but its undercapitalized position requires that any acceptable and collectable offer must satisfy $\frac{P_1(1-\gamma^2)+P_2(1-\gamma)}{(1+\gamma)^2} \geq D_1 - P_0$, which works out to $P_0 + 0.8P_1 + 0.72P_2 \geq D_1 = 0.8$. Because the liquidation value at date 2 is one, $X_2 = 1$, the borrower cannot commit to pay more than one at date 2 and $P_2 \leq 1$. To avoid foreclosure by the bank, the borrower must offer collectable payments such that $P_0 + 0.8P_1 + 0.72(1) \geq D_1 = 0.8$, or $P_0 + 8P_1 \geq 0.8$. If the borrower has less cash than this at dates 0 and 1, the bank must foreclose.

If the borrower has no date 0 cash, there can be no immediate payment ($P_0 \leq C_0 = 0$). The maximum collectable date 2 payment is $X_2 = 1$. To induce the bank not to liquidate, the borrower must make a collectable offer that satisfies $0.8P_1 + 0.72(1) \geq D_1 = 0.8$ (in addition to a nonbinding $P_1 + P_2 \geq X_1$), or $P_1 \geq 0.1$. If date 1 cash is too low ($C_1 < 0.1$), the bank will foreclose.

The undercapitalized bank discounts future payments to meet its capital requirement. Its limit on pledging its cash flows to outside investors makes it discount future payments heavily. In addition, because the bank discounts future payments but the borrower does not, the borrower will pay as rapidly as possible. The total amount paid, then, is $P_0+P_1+P_2 = C_0+C_1+\frac{D_1-C_0-0.8C_1}{0.72}$, assuming that a positive payment at date 2, P_2, is required (i.e., $C_0+0.8C_1 < 0.8$).

The bank's desperation either leads to liquidation or changes the amount that it forces a liquidity-constrained borrower to pay. Moreover, the bank's ability to extract payment from the borrower does not change monotonically in its capital and depends on the borrower's project characteristics (such as the interim cash flow it generates).

These possibilities are most easily seen by considering three subexamples, with differing amounts of date 1 cash possessed by the borrower. In one, the borrower is short on cash (has none until date 2); in the second, the borrower has a large amount of immediate cash; and in the third, the borrower has an intermediate amount of cash. I retain the parameter assumptions $\gamma = \frac{1}{9}$, that $X_1 = 0.99$, and that $X_2 = 1$.

A Borrower with No Date 0 or Date 1 Cash ($C_0 = C_1 = 0$)

If the borrower has no cash at date 1 or date 0, but will have cash at date 2, the banker would like to wait until date 2 to collect $X_2 = 1$. However, 0.72 is the most cash the bank can raise before date 1 against the date 2 loan collection. The bank can raise 0.99 by liquidating before date 1. The bank's decisions are as follows.

1. If the bank is *well capitalized* (has initial date 1 maturing deposits of 0.72 or less), the borrower will offer $P_2 = 1$. The bank will not liquidate, but will wait until date 2, collect one, and will be able to meet its date 1 capital requirement.

2. If the bank is *undercapitalized* (has deposits in excess of 0.72 to pay on date 1, but less than 0.99) the bank will (inefficiently) liquidate the borrower's collateral. It would not be able to meet its capital standard otherwise. By liquidating, the bank can raise 0.99, pay down deposits,

and meet the capital standard. So long as deposits are less than 0.99, the bank can avoid failure at date 1.

3. If the bank is *severely undercapitalized* (deposits exceed 0.99), the bank fails at date 1. The borrower faces liquidation after the bank fails because it can offer no cash to avoid it.

Bargaining with a Borrower with Lots of Cash

Suppose that the borrower has date 0 cash of $C_0 \geq X_1 = 0.99$. The borrower would like to pay down the loan as soon as possible when the bank charges very high rates to abstain from liquidation. For borrowers with high cash:

1. If the bank is *well capitalized* (has initial date 1 maturing deposits of 0.72 or less), the bank is free to wait until date 2 and collect one, and the borrower will pay $X_2 = 1$ in total.

2. If the bank is *undercapitalized* (has deposits in excess of 0.72 to pay on date 1, but less than 0.99), the bank cannot reject an offer and wait until date 2 to collect because it will violate its capital standard. The borrower will pay 0.99 immediately.

3. If the bank is *severely undercapitalized* (deposits exceed 0.99), the bank must fail at date 1. The borrower will be able to negotiate a settlement with the government receiver after the bank fails to pay $\frac{1}{2}X_1 = 0.4545$ because the receiver would otherwise need to hire the banker to collect the loan, and the banker would charge a rent of $\frac{1}{2}X_1 = 0.4545$.

The borrower with substantial cash benefits from the bank's desperation if the bank accepts a low payment in order to survive or if the bank fails and the borrower can make a partial payment to the receiver of the failed bank (whose collection skills are weaker).

This result does not just apply to borrowers with initial cash of $C_0 \geq 0.99$. A borrower who can pay C_0 immediately, as well as pay $X_1 - C_0$ at date 1 such that $0.8(X_1 - C_0) + C_0 \geq D_1$, will be able to benefit from the bank's desperation. For example, if $C_0 \geq 0.4$ and $C_1 \geq 0.59$, then $0.8(0.99 - C_0) + C_0 \geq D_1 = 0.8$ implies that the borrower can make an acceptable total payment of 0.99. If $C_0 < 0.4$, then the borrower's total payment must exceed 0.99 because higher date 1 or 2 payments that satisfy $C_0 + 0.8(P_1) + 0.72(P_2)$ will exceed $D_1 = 0.8$. If $C_0 < 0.4$, then the borrower has an intermediate amount of cash. This is the third (and most complicated) example.

An Intermediate Amount of Cash

If the borrower has enough date 0 cash to avoid liquidation, but not enough to induce the bank to accept a total payment of $X_1 = 0.99$, then the total amount that the borrower must pay, $P_0 + P_1 + P_2$, is decreasing in the borrower's cash holding.

Because the bank can raise at most 0.72 without liquidating, the *undercapitalized* bank will have constraints on its behavior at date 1. In particular, the banker's horizon is affected by its bargaining with borrowers. If the bank responds to default by waiting until date 2 to liquidate, then it will close. Any borrower who wants to avoid immediate liquidation needs to offer a positive date 1 payment. This necessity can force the borrower to pay more than the value of the bank's liquidation threat. A borrower with date 0 cash of exactly $D_1 - 0.72$, and no date 1 cash, would need to pay all the date 0 cash to the bank and also allow the bank to collect $X_2 = 1$ (the maximum that it can collect) at date 2. As borrowers have more cash, they can reduce their total payment, taking advantage of the undercapitalized bank's desperation. Borrowers with date 0 cash of less than $D_1 - 0.72$ meet the fate of the borrower with no date 1 cash: immediate liquidation.

This analysis implies the following characterization when the borrower has this intermediate amount of date 1 cash.

The bank needs to satisfy the constraints

$$P_0 + \frac{P_1(1-\gamma^2)+P_2(1-\gamma)}{(1+\gamma)^2} \geq D_1 \text{ and } P_0 + P_1 + P_2 = X_1.$$

1. If the bank is *well capitalized* (has initial date 1 maturing deposits of 0.72 or less), the bank will collect a total of $X_2 = 1$ from the borrower and will not liquidate.

2. If the bank is *undercapitalized* (has deposits in excess of 0.72 to pay on date 1, but less than 0.99), the binding constraint is $P_0 + \frac{P_1(1-\gamma^2)+P_2(1-\gamma)}{(1+\gamma)^2} \geq D_1$ and the borrower will want to pay as quickly as possible. The borrower will set $C_0 + 0.8(P_1) + 0.72(P_2) = D_1$, where $P_2 = \min\{0, (D_1 - C_0 - 0.8C_1)/0.72\}$ and $P_1 = \min\{C_1, (D_1 - C_0)/0.8\}$.
   With $D_1 = 0.8$, if $C_0 = 0$ and $P_1 = C_1 \in ((0.08)/0.8, 0.99) = (0.1, 0.99)$, then $P_2 = (0.8 - 0.8C_1)/0.72$ and the borrower will pay all of its date 1 cash to the bank, plus offer a positive payment to the bank at date 2 to deter the bank from liquidation. The total payment $P_0 + P_1 + P_2$ declines monotonically from 1.1 to 0.99 as cash C_1 increases from 0.1 to 0.19.
   If $C_1 = 0$, but $C_0 > 0$, then the total payment declines from 1.08 to 0.99 as C_0 increases from 0.08 to 0.31149.

3. If the bank is *severely undercapitalized*, $D_1 > X_1 = 0.99$, the bank fails. After the bank fails, the borrower is liquidated if $C_0 < \frac{1}{2}X_1 = 0.4545$ and otherwise pays 0.4545 to avoid liquidation.

It is worth noting that an undercapitalized bank facing a borrower with an intermediate amount of cash can force the borrower to make a very large payment—a payment as large as 1.1, which is in excess of the $X_2 = 1$ that a well-capitalized bank can collect.

4. POLICY RESPONSE TO UNDERCAPITALIZED BANKS WHEN FUTURE UNDERCAPITALIZATION LEADS TO CLOSURE

What is a government to do? The well-capitalized bank makes appropriate decisions, but it may collect less from borrowers with a moderate amount of current cash. The undercapitalized bank will squeeze cash-poor borrowers, break mutually beneficial relationships with very low cash borrowers, and collect less than the maximum amount that it can from liquid borrowers. Severely undercapitalized banks face immediate closure.

A government that cares about preserving the banking system itself might be very tempted to add at least enough capital to prevent immediate closure. But what is the effect of this action on the borrower, the corporate sector, employment, and growth? If the bank fails, then there will be bargaining such that the borrower can be forced to pay $\frac{1}{2}X_1$ because the government would be forced to hire the banker to collect the loan at date 1 if the borrower paid less than this amount. Returning to "Example" on page 86, the borrower must pay $\frac{1}{2}X_1 = 0.4545$ (it is $\frac{1}{2}X_1$ because the government will be forced to hire the banker to collect the loan at date 1) or face immediate liquidation. If the borrower has a very large amount of date 1 cash (at least $\frac{1}{2}X_1 = 0.4545$), then the borrower would benefit from the bank's failure because it has little future value in its relationship with the bank and can get rid of its debt burden more cheaply if the bank fails. However, this case requires the borrower to have current cash flows that are a very large fraction of its total long-run value. If the borrower has less cash, the borrower will be liquidated if the bank fails, but only one-half of the proceeds would go to depositors and the government deposit insurer. The corporate sector will be very anxious to have the bank recapitalized in this case if their cash is just below $\frac{1}{2}X_1 = 0.4545$. How much recapitalization they desire will depend on how much cash they have. If they have enough date 1 cash to frontload the payment to the bank, so that its total value and its pledged value are close to $X_1 = 0.99$, then a small recapitalization is desired. In this case, the borrower could avoid the liquidation threat by making date 1 payments and small date 2 promises to the bank. If the borrower has too little cash to do this, a large recapitalization is desired.

Once the bank has been given enough capital to be well capitalized, any additional capital will transfer rents to the banker and reduce the rate of return received by the government. Too small a recapitalization (from severely undercapitalized to undercapitalized) may be bad because it will not prevent inefficient foreclosure. This is especially true if the borrowers are short on cash. This is a bit outside the model, but it can be less expensive for a government that wants to avoid inefficient liquidation to give banks a smaller amount of capital and give the firms cash to pay the banks. This approach reduces the banker's rents and protects the human capital in firms; however, it also requires the government to know which firms are viable but short on cash. The latter seems unlikely, but is outside the model so cannot be confirmed here. Too large a recapitalization will lead not to inefficient loan decisions, but to inefficient operations in the bank, and it will increase the cost to the government.

Evergreening and Loss of Bargaining Power When Book Capital Is Inaccurate

Suppose that if a bank exercised its liquidation threat, its book capital would fall sufficiently to force immediate closure. The bank will never foreclose in this situation, which protects the borrower from foreclosure, but implies that the borrower will not have an incentive to pay the bank at all. If the borrower is the efficient user of the firm's capital, valuable human capital is protected, but further reductions in the real economic capital of the bank result. For borrowers with nonviable businesses that should be liquidated for efficiency, this effect delays efficient redeployment of capital and increases the losses to the banking system, due to lost bargaining power.

This case occurs when deposits exceed X_1, the amount that the bank can get from liquidation, but when regulatory capital is inflated by the overvaluation of the loan. Such banks would fit into the severely undercapitalized category in the examples.

In the model outlined above, where the borrower is viable and thus is the best user of the firm's capital, bank recapitalization sufficient to avoid evergreening can be a free lunch for the government. This result occurs if the borrower has sufficient cash to reach a negotiated settlement with the bank, worth at least X_2. If the bank evergreens and then fails, the borrower will end up paying a very small amount (one-half of what the bank could liquidate for, or one-half of X_1). By recapitalizing the bank sufficiently to have it negotiate a larger payment (equal to the full liquidation value), the government can save the deposit insurer money. The real decision is the same, but the borrower pays more. This saves the deposit insurer money.

Once enough capital has been advanced to allow a negotiated settlement, the analysis in the remainder of the article applies. The results imply that if the

borrower is short of date 1 cash, a small recapitalization that is just sufficient to avoid evergreening (to $D_1 = 0.99$ and leaving the bank undercapitalized) is a bad policy. An undercapitalized bank will liquidate inefficiently, and the borrower and society are worse off than if the bank had received no capital and continued to be afraid to liquidate. If the government provides this small amount of capital and borrowers are cash poor, the borrower will lobby the government for relief. It will ask for cash or ask the government to force the banks to convert some debt into equity, reducing the amount that the banks obtain from liquidation. After the Japanese government provided the initial recapitalization of banks in Japan, this position was taken by the Japan Federation of Economic Organizations (Keidanren) (see Rowley [1999]). Viable borrowers would be less afraid of a bank recapitalization if the bank were well capitalized ($D_1 < 0.72$).

The Intertemporal Problem with Repeated Government Recapitalization

Government recapitalization leads to a classic time consistency problem. If the deposit insurer cannot put capital into banks, but can only allow them to stay in business without recapitalization, then there is a limit on the concessions that can be extracted from deposit insurers over the short term. However, anticipations of regulators' closure behavior can give bankers perverse current incentives. If a period of persistent undercapitalization exists, then a government will wish to provide a subsidized recapitalization. If the future closure policy did not change, all parties in the economy could be better off (protecting human and physical capital). The government would have a bad influence if it generated a belief that recapitalizations were always forthcoming. That influence would totally eliminate liquidity creation by banks and lead to large future government expenditure on bank bailouts. It would be desirable to use political constraints to recapitalize banks only when called for by external conditions, and not because of banker rent-taking or incompetence. However, bankers will realize that this discrimination will be imperfect. The possibility of future recapitalization will lead to rents to banker human capital (overemployment, excessive costs, and resistance to change). It is therefore very appropriate that Japanese recapitalization has been accompanied both by a promise of commitment to future prompt corrective action and employment reduction and by improved portfolio disclosure and valuation. But the very logic that suggests that recapitalization can be ex-post desirable also suggests that the government may have a difficult time forcing banks to carry through with their commitments if they remain unprofitable.

5. BANKS THAT SHOULD NOT BE RECAPITALIZED

A Bank with No Relationship Lending

The financial health of a bank without lending relationships is of no consequence to the borrower. Such a bank can sell loans to meet the capital requirements, and the sale or retention of loans is of no consequence to the borrower. If the value of capital is negative, then the bank will not be able to recapitalize without subsidized capital; again, this is of no consequence to the borrower. The decision to liquidate or to continue lending is independent of the identity of the lender.

A Nonviable Borrower

A borrower is nonviable if the current management is not the best user of the firm's capital, and as a result the lender can collect more by foreclosure than by continuing to lend. If there is no lending relationship, then anyone can collect more from foreclosure, implying that independent of the capital position of a bank, there will be foreclosure after default. In this case, the only value of recapitalization is to avoid evergreening that prevents loans from being foreclosed, but such liquidation could be achieved by a government agency that foreclosed on the loans, perhaps by hiring bankers from the failed bank. There is no long-run value to retaining relationships to nonviable borrowers.

6. SUMMARY AND CONCLUSION

The analysis presented here suggests that for banks with viable lending relationships, it may be a good policy to recapitalize banks until they are well capitalized. Recapitalizing them only to the point where they are willing to write off loans (stop the evergreen policy) or to the undercapitalized point where they avoid failure only by liquidating the collateral of viable borrowers are both bad policies. These policies make sense only if some cash is provided to borrowers by the government or if the banks are forced to extend the viable loans in return for receiving the capital. But such multiple-level bailouts by the government would require more information and long-run commitment than a government possesses.

Providing too much capital to the banks will leave them with rents, which in the Japanese context take the forms of a too-large wage bill and continued inefficient operations. The government faces a difficult problem. Too little capital may be worse than none, and too much will be wasted. It is appropriate in this context that the capital injections to date have in return required labor force reductions and explicit management plans. However, nothing focuses a bank on rent reduction as much as the threat of impending closure.

Table 2 Details of Desirable and Undesirable Forms of Recapitalization

	Financially distressed bank with a relationship borrower	**Financially distressed bank without a relationship borrower**
Borrower has the best use of the collateral (and is thus viable)	Main case analyzed. Provide subsidized capital to well-capitalized level unless borrowers have substantial cash. Providing just enough capital to end fear of writing off loans due to book capital problems ("evergreen") is worse than providing no capital.	No reason to recapitalize. Will not liquidate inefficiently. Recapitalization just to the level to avoid fear of writing off loans due to book capital problems ("evergreen") has no effect.
Borrower does not have the best use of the collateral (and is thus not viable)	Undercapitalized bank will liquidate (efficiently) unless subject to the evergreen effect on book capital. Recapitalization just sufficient to avoid evergreen is a good policy. More capital has no beneficial effect.	No reason to recapitalize. Recapitalization just enough to avoid evergreen leads to efficient foreclosure. Equivalent to transferring loans to an outside collection agency.

The recent recapitalization in Japan has come in two stages, and it has been suggested that more stages might be forthcoming. Given the time-consistency problem, repeated recapitalization can cause problems. Guaranteed future recapitalization is equivalent to an all-capital bank. This guarantee leads to maximum rents and destroys liquidity creation.

Finally, the analysis has focused on banks with valuable relationships whose borrowers are still viable. Banks not in this category should be closed. A change in capital will not change a bank's incentive to inefficiently foreclose unless it has a relationship, so there is no extra efficiency gain from recapitalizing them. If the bank has a relationship, but the borrowers are not viable, then efficient allocation of capital requires that the borrowers' collateral be liquidated and redeployed. Absent accounting-based reluctance to foreclose, the banks would have every incentive to liquidate such borrowers, even if undercapitalized. If evergreening is the issue, recapitalizing the bank slightly could be sensible, but just for the purpose of closing it very soon thereafter. Alternatively, if the bank's extra efficiency in liquidating those loans is small, the best option will be to close it and transfer collection to a receiver (such

as the Japanese Resolution and Collection Corporation (RCC)). These results are summarized in Table 2.

This analysis is just a first step in the study of the optimal amount of recapitalization to provide to banks. There is much to add to make the results robust. However, I am not aware of any other theoretically based analysis of this topic, so this first step is an important beginning. It is clear that recapitalization by the government has time-consistency problems if it is expected to continue in the future. To my mind, this is not an argument against the current recapitalization. When (nearly) all the banks are underwater, it is desirable to recapitalize at least some of them. We need a framework to determine which ones are to be provided with subsidized capital, and how much to provide.

APPENDIX: BANK LOAN COLLECTION AT HIGH LEVELS OF CAPITAL

If the relationship-lending skill to collect the loan at date 2 is not lost, if the banker does not collect the loan in period 1 this period, and if the bank has enough capital that it would not be closed if it collected nothing at date $1 (D_1 \leq \frac{X_2}{1+\gamma})$, the payoff to capital would be greater than zero if the holders of capital rejected an offer from the banker to collect the loan at date 1. If the borrower defaults and the holders of capital do not reach an agreement with the banker to collect the loan at date 1, capital holders will be able to hire the banker at date 2 to collect X_2 at that time. This high level of date 1 capital holders will turn out to imply that the bank is well capitalized, by the definition in the article. The only difference in the analysis is that because capital holders have a positive outside option to reject the banker's offer to collect the loan at date 1, capital holders get a payment from bankers that will exceed $\frac{1}{2}(P_1 + \frac{P_2}{1+\gamma} - D_1)$. This difference has no effect on the banker's negotiation with the borrower: the bank can still collect the unconstrained amount, $\max[X_1, X_2]$.

REFERENCES

Aoki, Masahiko, Hugh Patrick, and Paul Sheard. 1994. *The Japanese Main Bank System: Its Relevance for Developing and Transforming Economies.* Oxford: Oxford University Press.

Bond, Philip. 1999. Lending with Joint Liability. Ph.D. dissertation, University of Chicago, Department of Economics.

Diamond, D. W. 2001. "Should Japanese Banks be Recapitalized?" Bank of Japan *Monetary and Economic Studies* 19: 1–19.

_____ and R. G. Rajan. 2000. "A Theory of Bank Capital." *Journal of Finance* 55 (December): 2431–65.

_____. 2001a. "Liquidity Risk Liquidity Creation and Financial Fragility: A Theory of Banking." *Journal of Political Economy* 109 (April).

_____. 2001b. "Banks, Short Term Debt, and Financial Crises: Theory, Policy Implications, and Applications." Carnegie Rochester Conference on Public Policy 54 (Summer).

_____. 2001c. "Banks and Liquidity." *American Economic Review, Papers and Proceedings* 91: 422–25.

Hogarth, Glenn, and Joe Thomas. 1999. "Will Bank Recapitalization Boost Domestic Demand in Japan?" Bank of England *Financial Stability Review* (June).

Hoshi, Takeo, and Hugh Patrick, eds. 2000. *Innovations in Financial Markets and Institutions.* Boston: Kluwer Academic.

_____, Anil Kashyap, and David Scharfstein. 1991. "Corporate Structure, Liquidity, and Investment: Evidence from Japanese Industrial Groups." *Quarterly Journal of Economics* 106 (February): 33–60.

Ito, Takatoshi, and Yuri Sasaki. 1998. "Impacts of the Basle Capital Standard on Japanese Bank's Behavior." Working paper. Hitotsubashi University and Takachiho University, August.

Nakaso, Hiroshi. 1999. "Recent Banking Sector Reforms in Japan." Federal Reserve Bank of New York *Economic Policy Review* (July): 1–7.

Rowley, Anthony. 1999. "Japan's urgent call to forgive." Singapore *Business Times*, (May 27): 20.