Changes in the Size Distribution of U.S. Banks: 1960–2005

Hubert P. Janicki and Edward Simpson Prescott

In 1960, there were nearly 13,000 independent banks. By 2005, the number had dropped in half, to about 6,500. In 1960, the ten largest banks held 21 percent of the banking industry's assets. By 2005, this share had grown to almost 60 percent. A great deal of these changes started during the deregulation of the 1980s and 1990s. (Figures 1 and 2 report the time paths for these two measures.)

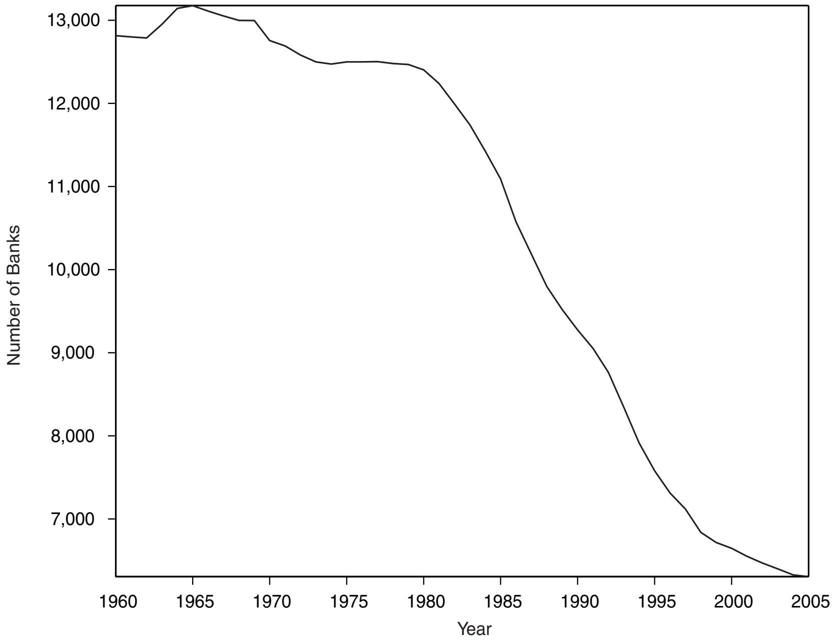
By any measure, these numbers represent a dramatic change in the bank size distribution over the 1960–2005 period. This article documents the extent of this change. It also documents the change in bank size dynamics, that is, the entry and exit of banks and the movement of banks through the size distribution. During this period, new banks formed, many more exited, either because of failure or merger, and many others changed in their size. For example, of the ten largest banks in 1960, only three were still among the top ten largest in 2005.[1]

We document these facts because they are an important step in developing a theory of bank size distribution. Although we do not provide one, such a theory would be valuable because it could be used to answer important questions such as: How costly were the pre-1980 limits on bank size? How

Hubert P. Janicki is a graduate student in economics at Arizona State University. The views expressed in this article are those of the authors and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System. The authors would like to thank Huberto Ennis, Yash Mehra, Andrea Waddle, and John Weinberg for helpful comments.

[1] The ten largest banks in 1960 were Bank of America, Chase Manhattan Bank, First National City Bank of New York, Chemical Bank New York Trust Company, Morgan Guaranty Trust Company, Manufacturer's Hanover Trust Company, Bank of California, Security First National Bank, Banker's Trust Company, and First National Bank of Chicago. At the end of 2005, only three of these banks still existed: Bank of America, Citigroup (formerly National City Bank of New York), and JP Morgan Chase (an amalgamation of many of the banks listed above).

Figure 1 Total Number of Independent Banks



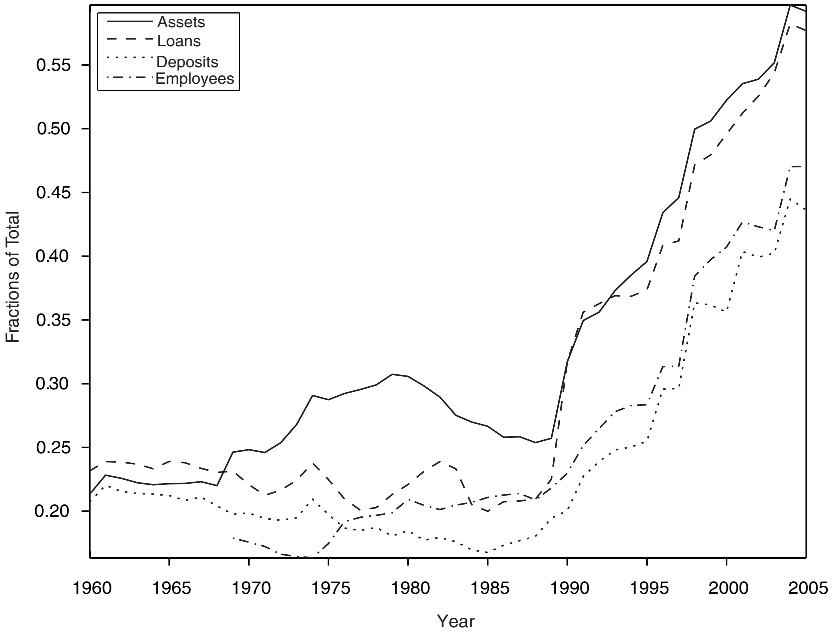
Notes: All banks and bank holding companies that are under a higher level holding company are treated as a single independent bank. Section 3 gives a more precise definition of an independent bank.

will the bank size distribution continue to evolve? And will there be more concentration? If so, should policy do anything about it?

Our analysis of the size distribution emphasizes fitting the data to the lognormal and Pareto distributions. These distributions are utilized because they are commonly used to describe skewed distributions and frequently have been used to describe firm size distribution. As we will demonstrate, the lognormal poorly fits the upper right tail of the size distribution. The Pareto distribution fits better with this part of the distribution, but the quality of the fit is much better before deregulation.

We examine bank size dynamics along several dimensions. First, we determine whether the data satisfies Gibrat's Law, that is, whether growth is independent of firm size. We find that Gibrat's Law is a good description of the data during the 1960s and 1970s, before deregulation, but not a good description afterward. After the 1970s, large banks grew faster than small banks, though more so in the 1980s and 1990s than they did during the

Figure 2 Market Shares of Ten Largest Banks



Notes: The definition of a bank is given in Section 3. Data on number of employees starts in 1969.

2000–2005 period. Second, we document that entry into banking was remarkably stable over the entire period. Entry is cyclical but averages about 1.5 percent of total operating banks. Finally, we calculate transition matrices, that is, the probability a bank will move from one size category to another, over each of the decades. Following Adelman (1958) and Simon and Bonini (1958), we use these transition probabilities and the entry data from 2000–2005 to forecast continued changes in the size distribution. The forecast predicts a continued decline in the total number of banks, but at a much slower rate than in the 1980s and 1990s, followed by a leveling off in the decline. It also predicts that there will still be a large number of small banks as well as a sizable number of mid-size banks. If the present trends continue, the transition in banking that began in the 1980s is slowing down and coming to an end.

1. LITERATURE

In many industries, the distribution of firm size is highly skewed to the right, that is, there are many small firms and a few large ones. One distribution that has this characteristic and is frequently used to describe firm size distribution is the lognormal distribution. A random variable is lognormally distributed if the logarithm of the random variable is normally distributed.[2]

Early studies of firm size distribution, namely Gibrat (1931), found that the lognormal distribution fit the empirical data fairly well. Gibrat (1931) also found evidence that firm growth was independent of firm size. This latter finding, often called Gibrat's Law or the Law of Proportionate Effect, was important because a statistical process that satisfies it would generate a lognormal distribution in the long run.

Later studies have found mixed support for Gibrat's findings. In particular, studies in the 1980s found that the proportional rate of growth of a firm conditional on survival decreases in size. Sutton (1997) is a good survey of these results.

Another category of distributions used to fit the size distribution is based on the power law. This category takes the form

$$f(x) = cx^{-\alpha},$$

where $x > 0$ and $c > 0$. In economics, the Pareto distribution is a power law distribution often used to describe highly skewed data.[3] It is similar to the lognormal, but with a thicker right tail.

Power law distributions have been used in the sciences to fit data in a wide variety of applications. Newman (2005) surveys various applications of the power law, including studies of word frequency, magnitude of earthquakes, diameter of moon craters, intensity of solar flares, and population of cities. One property observed in many applications is that the size of the r-th largest observation is inversely proportional to its rank.[4] It is observed so frequently that it is called Zipf's Law. Like the lognormal distribution, Zipf's Law can be generated with appealing assumptions on the dynamics. For example, Simon and Bonini (1958) study entry dynamics by assuming a constant probability of entry and show that the distribution follows a power law in the upper tail. See Gabaix (1999) for a detailed study of Zipf's Law.

[2] The probability density function of a lognormal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{x}e^{-(log(x)-\mu)^2/(2\sigma^2)},$$

where $x > 0$, μ and σ are the mean and standard deviation, respectively, of the natural log of x.

[3] The probability density function of a Pareto distribution is

$$f(x) = kx_{min}^{k}x^{-(k+1)}$$

where $x > x_{min}$, $x_{min} > 0$ and $k > 0$. The Pareto distribution is usually expressed as the probability that random variable X is greater than x, that is, $Prob(X \geq x) = x_{min}^{k}x^{-k}$, which is also a power law.

[4] Formally, $x_r = cr^{-\alpha}$, where observation x_r has rank r, c is a constant, and α is close to 1. We solve for r, normalize the equation by dividing by N (where the Nth ranked observation is x_{min}) and simplify to obtain the counter cumulative function: $r/N = 1 - F(x) = (x_{min}/x)^{1/\alpha}$. This is the Pareto distribution with coefficient $k = 1/\alpha$.

Whether Zipf's Law fits U.S. firm size data is a matter of some debate. Using 1997 U.S. census data, Axtell (2001) finds that it fits the firm size distribution. Using a different data set, however, Rossi-Hansberg and Wright (2006) find that it does not fit so well. They also find that establishment growth and exit rates decline with size.

The banking literature has long been interested in the size distribution of banks, partly because of the large degree to which laws and regulations limited bank size. Recent studies include Berger, Kashyap, and Scalise (1995), Ennis (2001), and Jones and Critchfield (2005).

While the banking literature has long noted that bank size distribution is skewed, it has not typically tried to fit this using the previously mentioned category of distributions. This absence has made it difficult to compare bank size distribution with that of other industries.

There is part of this literature, however, that tests for Gibrat's Law in the banking industry but for smaller samples than those used in this study. Alhadeff and Alhadeff (1964) analyze growth in assets of the 200 largest U.S. banks between 1930–1960 and find the largest banks grew more slowly than the banking system itself. They find, however, that the top banks that survived throughout the sample period grew faster than the system as a whole and attribute this to mergers among the largest banks. Rhoades and Yeats (1974) analyze growth among U.S. banks by deposits for 1960–1971 and find that the largest banks grew more slowly than the whole banking system. In a study of the 100 largest international banks from 1969 to 1977 by assets, Tschoegl (1983) finds that growth rates of banks are roughly independent of size, but that growth rates exhibit positive serial correlation. Saunders and Walter (1994) use international data for the 200 largest banks for the 1982–1987 period and reject Gibrat's Law, finding that the smaller banks grow faster than larger banks in terms of assets. More recently, Goddard, McKillop, and Wilson (2002) find evidence that during the 1990s in the United States, large credit unions grew faster than their smaller counterparts.

As we discussed, the firm growth results are important because they ultimately determine the distribution of firm sizes. One interesting strand in the literature calculates a Markov transition matrix of movement between size categories and then calculates stationary size distributions. Early examples include Hart and Prais (1956), Simon and Bonini (1958), and the study of the steel industry by Adelman (1958). The only study that we are aware of that applies this technique to banking is Robertson (2001).

More recently, the literature has attempted to generate firm size dynamics, and ultimately a size distribution, from models of maximizing behavior of firms. Sutton (1997) surveys several such models. One paper in this literature is the learning model of Jovanovic (1982), in which new firms receive productivity shocks, learn about them over time, and then decide whether to continue or exit. Another prominent example is Hopenhayn (1992).

2. LEGAL AND REGULATORY LIMITS ON BANK SIZE

Describing bank size distribution is particularly important because of the many legal and regulatory limits on bank size that existed through the 1970s and were removed during the 1980s and 1990s. As we will see, the removal of these barriers coincide with dramatic changes in size dynamics and the size distribution.

In 1960, banks could not branch across state lines and some states even forbade branching within a state.[5] The 1966 Douglas Amendment to the 1956 Bank Holding Company Act allowed interstate banking only with expressed authorization by participating states. However, no state allowed interstate banking at the time and the amendment was not even exercised until 1978 when Maine allowed out-of-state bank holding companies (BHCs) to operate within the state.[6] Over the next 12 years, many of the intrastate and interstate restrictions were removed by the states.[7]

The remaining interstate banking restrictions were removed by the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994. The act permitted bank holding companies to acquire banks in any state and, beginning in 1997, allowed interstate bank mergers. (See Kane [1996] for a summary of the act.) More recently, the Gramm-Leach-Bliley Act of 1999 allowed banks to engage in nonbanking financial activities such as insurance.

3. THE DATA

Many banks operate under a bank holding company structure. A bank holding company is a legal entity that i) directly or indirectly owns at least 25 percent of the bank's stock, ii) controls the election of a majority of a bank's directors, or iii) is deemed to exert controlling influence of bank policy by the Federal Reserve (Spong 2000). Many bank holding companies have multiple banks and even other holding companies under their control. Historically, this legal organization was used to avoid some of the restrictions on branching (Mengle 1990). In many cases, a bank holding company would operate many activities jointly. For this reason, we follow Berger, Kashyap, and Scalise (1995) and treat all banks and bank holding companies under a higher level holding

[5] The 1927 McFadden Act forbade interstate branching by federally chartered banks. Later, the Federal Reserve extended the ruling to include all state-chartered banks that are regulated by the Federal Reserve.

[6] There were some means around these restrictions. For example, the 1956 Bank Holding Company Act did not limit the location of *nonbank* subsidiaries of bank holding companies, so some banks had a cross-state network of *nonbranch offices* that would specialize in activities like lending. Also, some exemptions were allowed for the acquisition of insolvent banks from government deposit insurance funds (Kane 1996).

[7] Jayaratne and Strahan (1997) list when states removed restrictions to interstate banking and intrastate branching.

company as a single independent banking enterprise. For convenience, we will typically refer to each of these entities as a bank.

Data on banks are taken from the Reports on Condition and Income (the "Call Report") collected by federal bank regulators. We use fourth quarter data on all commercial banks in the United States. We look specifically at commercial banks and exclude savings banks, savings and loan associations, credit unions, investment banks, mutual funds, and credit card banks. Individual commercial banks that belong to a holding company are then grouped according to a unique bank holding company regulatory number, and their assets, deposits, loans, and employees are summed and replaced by one entry in our data set. Our data set, therefore, tracks bank holding companies and independent commercial banks not affiliated with a holding company. It does not distinguish between a merger and a failure. Both events are treated as an exit.

We use four measures of bank size. The first is commercial bank assets. For this variable, we have data from 1960–2005. Prior to 1969, we only have domestic holdings, but after 1969, we have foreign and domestic holdings. The next two size measures are domestic holdings of deposits and loans. For both of these variables, we have data from 1960–2005. The final size measure is the number of domestic employees.[8] For this last measure, we only have data for 1969–2005. Assets and loans are adjusted to include off-balance-sheet items starting in 1990. (See the Appendix for details.) All the variables are adjusted for the total aggregate size of that variable in each year. In particular, firm size data are converted into market share numbers for that year and then multiplied by the total quantity in the banking industry of that variable in 2004. The market share adjustment facilitates comparison across years, while the scaling by 2004 aggregate quantities gives a sense of the size in terms of recent quantities.

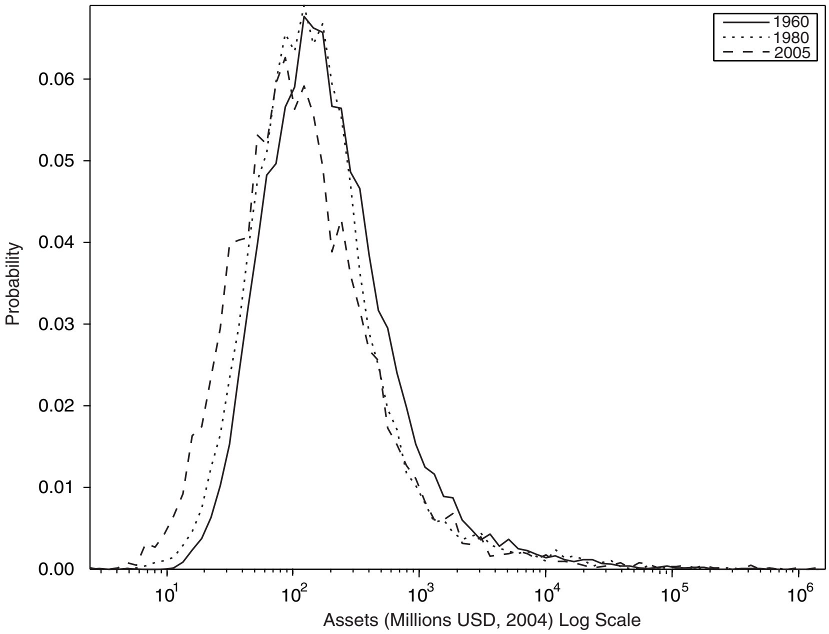
4. THE SIZE DISTRIBUTION

The size distribution of banks has always been skewed, but it has become more so since the 1960s. Figure 3 shows the distribution of assets for 1960, 1980, and 2005. Each year is normalized by the total assets in that year relative to 2004 so that the distributions are comparable over time. The distribution is plotted on a *log* scale. Because the size distribution is so skewed to the right, the log scale—or something similar—is needed to fit all the banks on the graph.

Figure 3 demonstrates that there are a large number of small banks and a few large banks. As is evident, there is a shift in the distribution to the left

[8] The Call Report counts employees in terms of full-time equivalents.

Figure 3 Change in Bank Size Distribution Over Time



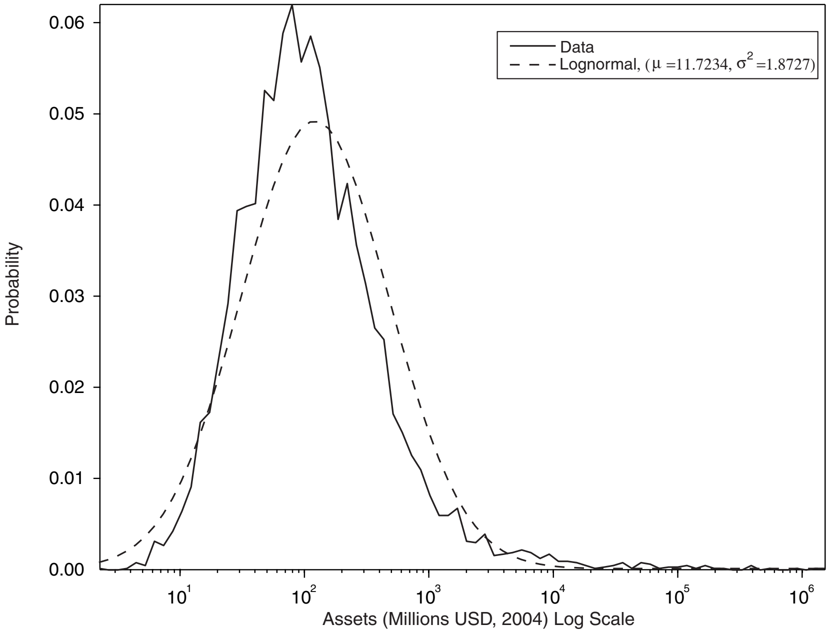
Notes: Each line is a probability distribution of bank size as measured by assets for a given year.

over time. Since we have scaled assets in each year to be the same scale, this change means that a higher fraction of the assets are being held by the small number of large banks, as indicated earlier in Figure 2.

Visually, the graphs suggest that the distribution might be accurately represented by a lognormal distribution. Figure 4 reports the actual distribution and an estimated lognormal distribution for assets in 2005, where the lognormal has parameters μ and σ^2 obtained from the 2005 data set. However, the distribution fails the Kolmogorov-Smirnov test for goodness-of-fit. This is true for almost all the years in the data set, as well as for the other size measures.

The estimated lognormal distribution does a particularly poor job of fitting the right tail of the distribution. This is hard to see in Figure 4 because of the small number of large banks. The fit at the right tail is better seen if we use a rank-frequency, or Zipf plot. For a power law distribution plotted on a log scale, this type of graph has the valuable property that the slope will be linear. For example, let $x_r = cr^{-\alpha}$, where r is the rank of a variable. Taking the

Figure 4 Size Distribution of Banks in 2005



Notes: The parameters μ and σ are the mean and standard deviation, respectively, of the natural log of assets. The lognormal distribution is calculated using these parameters.

logarithm of both sides, we obtain the equation,

$$ln(x_r) = ln(c) - \alpha ln(r).$$

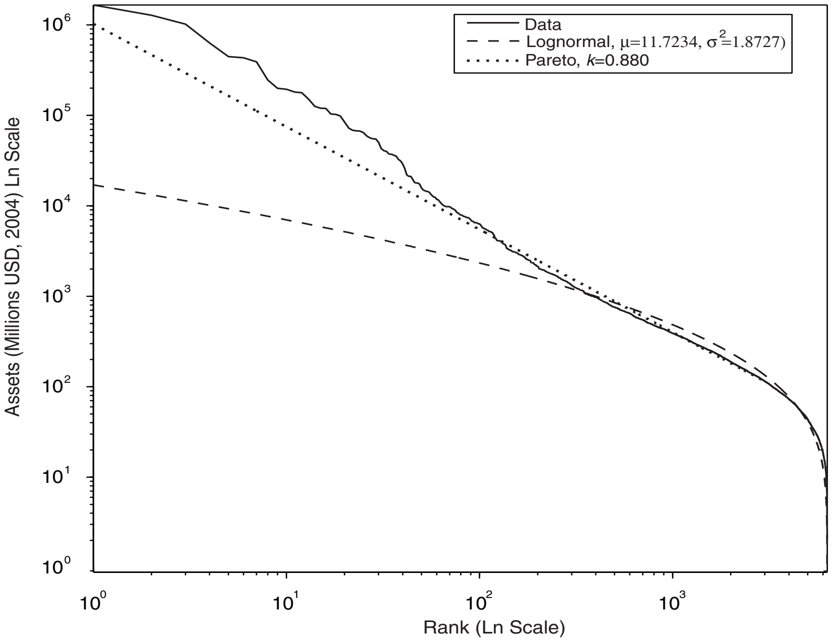
If $\alpha = 1$, then Zipf's Law holds.

Figure 5 is a Zipf plot for the bank size distribution in 2005.[9] It plots the data and the lognormal distribution using the sample mean and variance.

Figure 5 demonstrates that the lognormal distribution underestimates the density of the right tail (which is the left side in the Zipf plot). Indeed, if the plot is linear, we know that the right tail of the distribution of bank holding companies can be better approximated by the Pareto distribution. Note, however, that only the right tail of the distribution—which corresponds to the left side of the figure—appears to fit Zipf's Law. That is, the distribution of bank holding companies seems to be lognormally distributed with a Pareto-distributed tail.

[9] Zipf plots for different size measures look very similar.

Figure 5 Zipf Plot of Banks in 2005



Notes: The parameters μ and σ are the mean and standard deviation, respectively, of the natural log of assets. The lognormal distribution is calculated using these parameters. The estimate for the Pareto distribution is only for the 3,000 largest banks.

We can formally test whether the right tail of the distribution satisfies Zipf's Law. One common method is to estimate the slope by fitting a linear regression of the form,

$$ln(x_i) = \alpha_0 + \alpha_1 ln(r_i) + \epsilon_i, \tag{1}$$

where r is the rank of the bank. The coefficient α_1 is a power exponent in the Pareto distribution and if it is equal to –1, then Zipf's Law holds.

Ordinary least squares estimates of (1) will underestimate α_1 (see Gabaix and Ioannides 2004). For this reason, we use the maximum likelihood estimator in Newman (2005) to estimate the power coefficient in the Pareto distribution, or equivalently the slope α_1:

$$\alpha_1 = -n^{-1}[\sum_{i=1}^{n}(ln(x_i) - ln(x_{\min}))].$$

Table 1 Zipf's Law: Maximum Likelihood Estimates

	1960		1970		1980		1990		2005	
	$-\alpha_1$	SE	$-\alpha_1$	SE	$-\alpha_1$	SE	$-\alpha_1$	SE	$-\alpha_1$	SE
Assets	1.001	0.001	1.030	0.001	1.018	0.001	1.017	0.001	1.136	0.001
Deposits	0.998	0.001	1.016	0.001	0.983	0.001	0.982	0.001	1.092	0.001
Loans	1.028	0.001	1.051	0.001	1.007	0.001	1.082	0.001	1.187	0.001
Employees	–	–	1.022	0.065	1.031	0.073	1.005	0.084	1.052	0.096

Notes: The estimates are reported as $-\alpha_1$.

The maximum likelihood estimates for several different years are reported in Table 1.[10] We restrict the estimates to the tail by limiting the sample to the largest 3,000 banks for each year. Although they are not identical across years, they broadly support Zipf's Law in the upper tail, but the results are sensitive to the cutoff.

The worst fit in Table 1 is in 2005. This can be seen in Figure 5. The straight line in Figure 5 graphs the power distribution for the maximum likelihood estimate based on the 3,000 largest banks. The distribution is a straight line with slope –1.136. The slope is too high for Zipf's Law. Furthermore, the slope is high because the larger banks are larger than predicted by Zipf's Law, though this is less true for the largest.

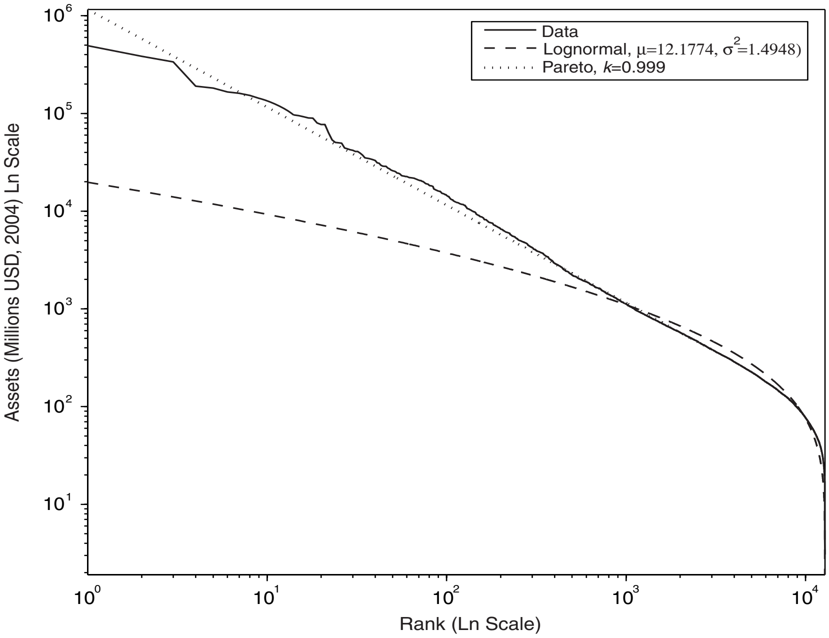
In contrast, the fit for assets in 1960 is excellent. Figure 6 is a Zipf plot for assets in 1960. The distribution is a straight line with slope –0.999, practically the same as in Zipf's Law.

The estimates for assets in 1970, 1980, and 1990 are similar, but this hides an important difference. The year 1970 is similar to 1960 in that many of the largest banks are smaller in size than predicted by the estimate. In 1980, this pattern changes to one where the largest banks are larger than predicted. The differences in the predictions for the largest banks are even larger in 1990 and look similar to that of 2005 (see Figure 5).

To summarize, only the right tail can reasonably be considered to be fitted by Zipf's Law, and the fit depends on the year. It does well in 1960, but starting in 1980 Zipf's Law predicts that the largest banks will be smaller than in the data. The lognormal distribution poorly fits the size of the largest banks but better fits the small banks.

[10] We calculate standard errors using the method outlined in Gabaix and Ioannides (2004).

Figure 6 Zipf Plot of Banks in 1960



Notes: The parameters μ and σ are the mean and standard deviation, respectively, of the natural log of assets. The lognormal distribution is calculated using these parameters. The estimate for the Pareto distribution is only for the 3,000 largest banks.

5. DOES GIBRAT'S LAW HOLD FOR BANKS?

Gibrat's Law states that firm size growth is independent of firm size. For our first test of this law, we fit the following linear equation,

$$ln(x_{it+1}) = \beta_0 + \beta_1 ln(x_{it}) + \epsilon_{it}, \qquad (2)$$

where x_{it} is the size measure (assets, employees, etc.) of bank i at time t. A coefficient value of $\beta_1 = 1$ means that growth is independent of size.

We estimated (2) over each decade and over 2000–2005 using ordinary least squares. We only considered banks in the sample that were around at the beginning and end of the estimation period. Those that exited were dropped from the sample.[11] The estimates are reported in Table 2.

[11] Sometimes the literature includes exiting banks and sometimes it does not. See Sutton (1997) for more information. To reduce survivorship bias, we also estimated equation (2) over each year and found similar results to those reported in Table 2.

Table 2 Gibrat's Law: Estimates

	1960–1969		**1970–1979**		**1980–1989**		**1990–1999**		**2000–2005**	
Assets										
β_0	0.143	(0.031)	0.574	(0.034)	0.043	(0.057)	–0.187	(0.074)	0.015	(0.061)
β_1	**0.987**	(0.003)	**0.953**	(0.003)	**1.011**	(0.005)	**1.024**	(0.006)	**0.998**	(0.005)
R^2	0.929		0.915		0.847		0.823		0.869	
Deposits										
β_0	0.435	(0.030)	1.004	(0.033)	0.131	(0.058)	0.654	(0.075)	0.379	(0.063)
β_1	**0.971**	(0.003)	**0.929**	(0.003)	**0.999**	(0.005)	**0.972**	(0.007)	**0.970**	(0.005)
R^2	0.923		0.902		0.826		0.787		0.847	
Loans										
β_0	0.545	(0.037)	1.368	(0.039)	0.000	(0.071)	0.635	(0.074)	0.444	(0.069)
β_1	**0.960**	(0.003)	**0.901**	(0.003)	**0.993**	(0.006)	**0.964**	(0.007)	**0.962**	(0.006)
R^2	0.889		0.870		0.767		0.782		0.812	
Employees										
β_0	–	–	–0.021	(0.012)	0.240	(0.016)	0.284	(0.022)	0.101	(0.017)
β_1	–	–	**1.006**	(0.003)	**1.001**	(0.004)	**1.008**	(0.006)	**0.994**	(0.004)
R^2	–		0.897		0.862		0.829		0.900	

Notes: The table provides ordinary least squares estimates of equation (2) for each sample period. Standard errors are in parentheses. Numbers in bold are the slope estimates of the effect of firm size on growth. An estimate close to one is consistent with Gibrat's Law.

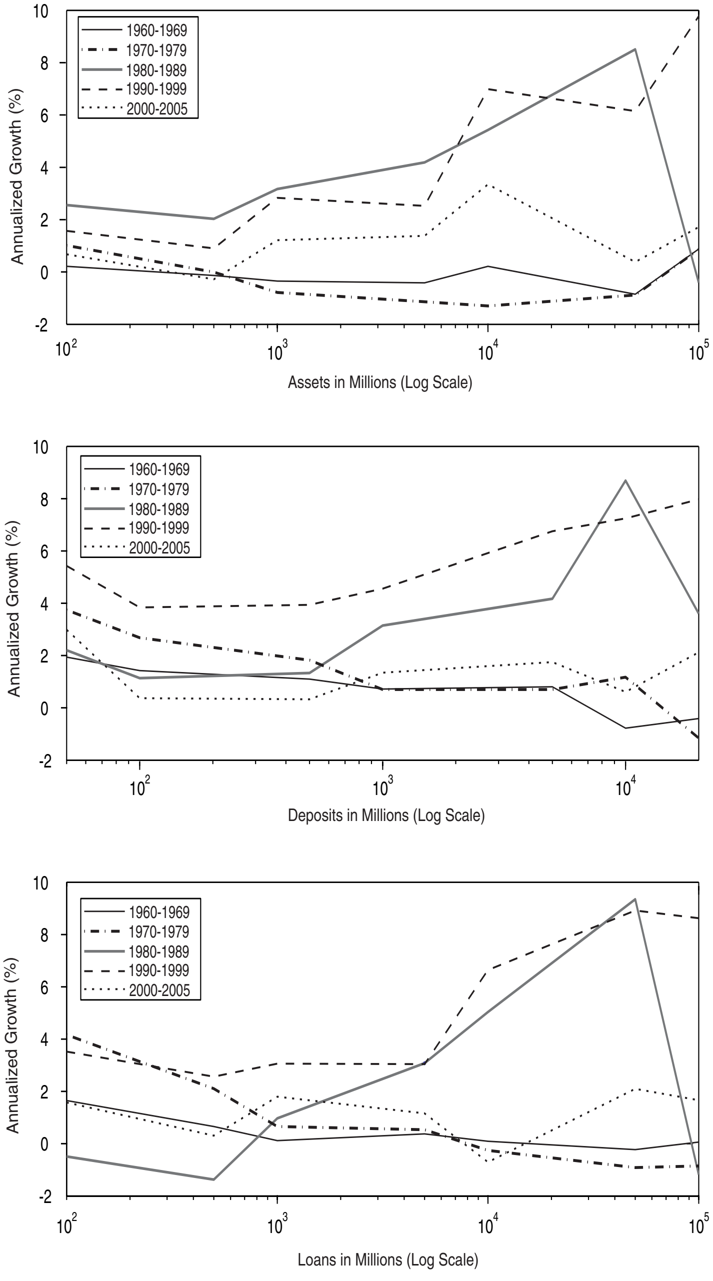
The estimates in Table 2 are close to one for all the decades and variables. While broadly supportive of Gibrat's Law, these estimates put a great deal of emphasis on small banks because they comprise most of the sample. For this reason, we broke the sample into different size categories and then calculated the annualized growth rates over the same periods for banks in each category. As before, we only considered banks that survived. Figure 7 reports the growth rates for assets, deposits, and loans.[12]

Just as in Table 2, growth measures do not vary much with size in the 1960s and 1970s.[13] In the 1980s and 1990s, however, the numbers in Figure 7 present a different picture than the estimates in Table 2. As demonstrated by the figure, the largest banks clearly grew faster than the small banks. This is also true for the 2000–2005 period, but the effect is less pronounced.

[12] A figure showing employee growth rates is excluded because the growth rates jump around significantly and do not show any clear patterns.

[13] We have also calculated annual growth rates over each period and then averaged them to get a similar figure. (Over the 1960–1969 period, for example, this meant calculating the 1960–1961, 1961–1962, etc., growth rates and then averaging them.) The results are similar.

Figure 7 Annualized Growth Rates by Size Categories



Notes: The size measure is broken into seven size categories and then average annualized growth rates are reported for each category.

Gibrat's model that yields the law of proportionate effect assumes that growth rates are not persistent over time. To check the validity of this assumption, we calculate the correlation of growth rates for surviving banks in each decade. We find that correlations of growth rates for all variables are very low. In the 1960s, it is 0.0610; in the 1970s, it is 0.0331; in the 1980s, it is 0.0691; in the 1990s, it is 0.0722; and after 2000, it is 0.1617. Correlation coefficients of growth in the remaining variables are of similar magnitude.

To conclude, growth rates appear to be independent of size in the 1960s and 1970s, but they are positively related to size in the 1980s and 1990s. In the 2000–2005 period, the growth rates are also higher for the largest banks, but less so than in the previous two decades. It appears that the 1960s and 1970s were relatively stationary periods, but the 1980s and 1990s were a long transition period, no doubt due to the legal, regulatory, and technological changes of the period. Finally, the 2000–2005 period appears to be the end of the transition, as the size dynamics seem to be returning slowly to the numbers of the 1960s and 1970s. Of course, this conclusion is tentative because trends calculated from five years of data can easily be transitory.

6. ENTRY AND EXIT

Despite the large number of banks that have exited the industry over the last 45 years, there has been a consistent flow of new bank entries. The number of entries and exits (including mergers) expressed as a fraction of the banking population is reported in Figure 8.

Visually, it is apparent that the flow of new banks is a relatively constant fraction of the banking population. To check this, we estimated a linear time trend of the number of entries as a fraction of the number of banks operating. Specifically we estimated the equation,

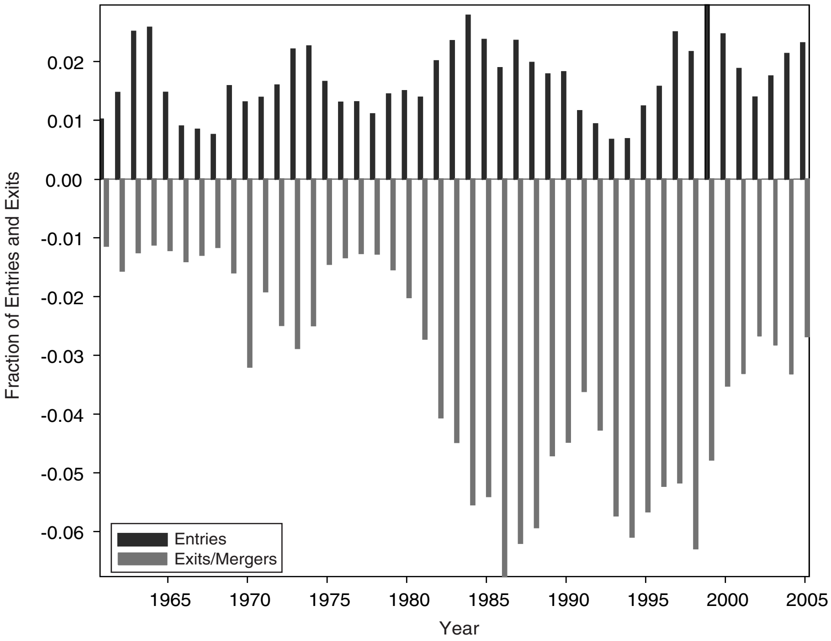
$$y_t = \gamma_0 + \gamma_1 t + \epsilon_t,$$

where y_t is the fraction of entries per year and t is the time trend. The ordinary least squares estimates are $\gamma_0 = 0.0145$ (0.0018), $\gamma_1 = 0.0001$ (0.0001) with R^2=0.0612. Standard errors are in parentheses. There is no time trend in the flow of new banks, though there is a significant amount of cyclical variation.

The fraction of banks that exit varied a great deal over time. There was a significant increase starting in the 1980s and, except for a short dip in the early 1990s, the high level continued through the late 1990s. No doubt much of this exit was due to mergers, particularly those that occurred in the 1990s, but our data does not allow us to distinguish between these two sources of exit. It is only in the last five years that the rate of exit seems to slow down.

It is striking that despite the huge number of bank exits starting in the 1980s, entry remained strong throughout the entire period. Interestingly, it is virtually uncorrelated with exit. For example, the correlation between exit

Figure 8 Fraction of Banks that Enter and Exit by Year



Notes: The chart reports the gross flow of banks that enter and exit expressed as a fraction of banks in each year.

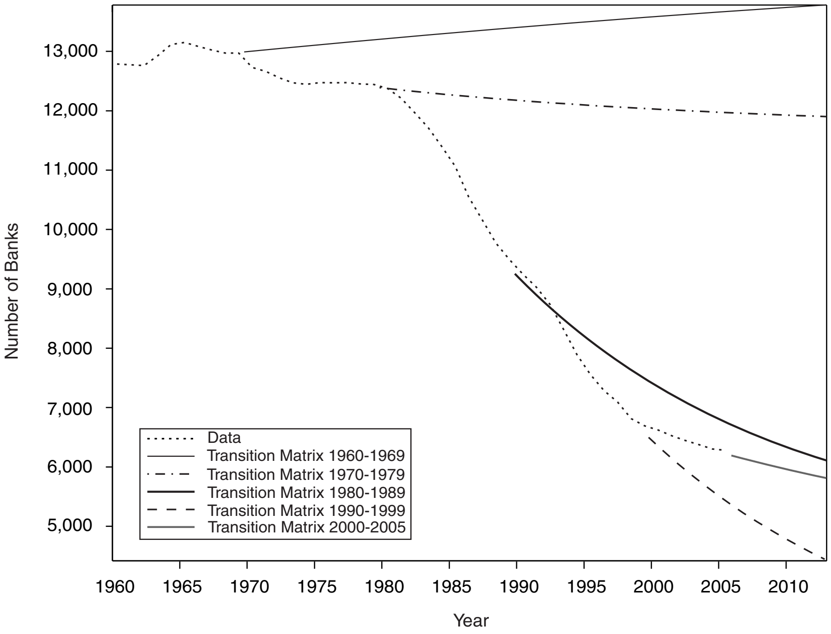
and entry for the 1985–2005 period is only –0.07. This evidence means that a theory of the bank size distribution needs to be able to generate robust entry, even when the total number of banks is declining.

7. THE DYNAMICS OF THE SIZE DISTRIBUTION

In this section, we use the information previously documented—the size distribution at a point of time, entry and exit rates, and bank dynamics—to make forecasts of what will happen to the size distribution. We also perform counterfactual experiments such as what would have happened to the bank size distribution if the dynamics had *not* changed. The purpose of these exercises is to develop a sense of how changes in size dynamics have mattered for the size distribution.

To make these calculations, we do not use growth rates at the individual bank level. Instead, we do something similar by constructing a simple Markov chain model following Adelman (1958). A Markov chain model splits the

Figure 9 Projected Number of Banks



Notes: Each projection is made by taking the transition matrix estimated over a given period and then calculating the total number of banks that would operate after the last year in that given period.

banks into a finite number of size categories. Probabilities of moving between size categories are summarized with a Markov, or transition, matrix. A Markov matrix is a square matrix, P, where element P_{ij} specifies the probability a bank that starts in size category i will move into size category j in the next period.

A Markov model allows for straightforward predictions about changes to a size distribution. For example, if the size distribution at time t is s_t, then the size distribution at time $t + n$ is

$$s_{t+n} = P^n s_t.$$

If a Markov model has the property that a bank starting in any category has a positive probability of moving to any other size category in a finite number of periods, then several useful theorems apply. First, there exists a

Table 3 Transition Matrix Size Categories (Scaled Dollars)

Size Categories	Assets
1	<100m
2	100m-500m
3	500m-1b
4	1b-5b
5	5b-10b
6	10b-50b
7	50b<

Notes: A list of size categories used in the Markov model. In all years, data are converted into market share numbers and then multiplied by the total quantity of assets in 2004.

stationary distribution, that is, there exists s such that $s = Ps$.[14] Second, the stationary distribution is unique and independent of the initial distribution of banks. Therefore, regardless of the initial distribution, if the transition matrix, P, is repeatedly applied to the distribution, then the size distribution will approach the unique stationary distribution.

We construct seven different size categories. These are listed in Table 3. All the data are scaled as we discussed earlier to make it possible to compare across years. We also include an eighth category that represents banks that are inactive. New banks come from this category and exiting banks move into it. We calculate the number of banks in each of the seven active categories. For the inactive banks, we assume that there is a large pool of 100,000 potential entries.

Our transition matrix is calculated by counting the fraction of banks that move from size category i to j each year over the specified time period. Entry rates are calculated so that a constant fraction of potential banks enter.

For our first exercise, we use the transition matrices estimated over several ranges of time periods to forecast the aggregate number of banks in the industry. We estimate several transition matrices and make predictions to 2013. The results are illustrated in Figure 9. For each period, we calculate the transition matrix and then forecast the change in the total number of banks as if the transition probabilities had not changed from that time period forward. This exercise is similar to one in Jones and Critchfield (2005), although our methodology is very different. An advantage of our Markov chain model is that we have information on the entire distribution at each point in Figure 9.

As is clear from the earlier analysis, as well as from Figure 9, the size dynamics changed significantly over this period. Nevertheless, the exercise

[14] Let s^j be the fraction of banks in size category j. The stationary distribution s is the solution to the set of equations $s = Ps$ and $\sum_j s_j = 1$.

Table 4 Stationary Distribution of Banks by Assets

Size	1960–1969		1970–1979		1980–1989		1990–1999		2000–2005	
Categories	**1969**	**SS**	**1979**	**SS**	**1989**	**SS**	**1999**	**SS**	**2005**	**SS**
1	0.333	0.357	0.379	0.439	0.379	0.471	0.440	0.523	0.480	0.540
2	0.504	0.516	0.492	0.485	0.488	0.435	0.436	0.377	0.399	0.363
3	0.085	0.076	0.064	0.044	0.067	0.049	0.061	0.051	0.060	0.050
4	0.060	0.042	0.046	0.021	0.043	0.027	0.044	0.037	0.043	0.033
5	0.008	0.004	0.007	0.003	0.008	0.005	0.006	0.004	0.008	0.006
6	0.009	0.005	0.010	0.005	0.010	0.007	0.007	0.005	0.006	0.005
7	0.002	0.000	0.002	0.003	0.005	0.007	0.005	0.002	0.005	0.003

Notes: Columns under a year list the actual size distribution in that year. Columns under "SS" list the stationary distribution as calculated from the estimated transition probabilities for that period.

is interesting because it illustrates how the dynamics matter for the size distribution. For example, the 1960s and 1970s look relatively stable. If these dynamics had not changed, there would not have been much change in the total number of banks. In the 1980s and 1990s, the dynamics predicted continued substantial declines in the number of banks. However, the transition probabilities changed significantly in the 2000–2005 period and as a consequence, the decline in the number of banks leveled off.

Table 4 reports stationary size distributions calculated from the estimated transition matrices for several periods in the data. We compare them with the actual distribution at the end of each period in order to illustrate whether the existing distribution was close to the stationary distribution. Interestingly, for every period, we find that the fraction of banks that are in the smallest category is higher in the stationary distribution than in the final year of the period. We also find that for every other period, the fraction of banks in each of the other size categories is less in the stationary distribution than in the final year (the only exceptions are during Category 2 [1960–1969] and Category 7 [1970–1979] and [1980–1989]).

Table 4 has several implications. First, even in the relatively stable decades of the 1960s and 1970s, the size distribution was not at a stationary point and under the estimated transition probabilities, the distribution would have continued to change. Second, there will continue to be large numbers of small banks, even if the fraction of assets they hold is not large. Third, we can see that the dynamics for the 1990s imply even more concentration than the dynamics from the 1980s. This was also suggested by Figure 9. Finally, there will continue to be a large number of mid-size banks. The most recent merger wave led some commentators to speculate that the bank size distribution would take a "barbell" shape with only small banks and large banks. The small banks would

Table 5 Distribution of Bank Assets

	2005		Stationary	
Size Categories	Fraction of Banks	Fraction of Assets	Fraction of Banks	Fraction of Assets
1	0.480	0.014	0.540	0.022
2	0.399	0.050	0.363	0.062
3	0.060	0.024	0.050	0.027
4	0.043	0.047	0.033	0.050
5	0.008	0.034	0.006	0.035
6	0.006	0.068	0.005	0.083
7	0.005	0.763	0.003	0.721

Notes: Fraction of assets for the stationary distribution was calculated by assuming that the mean assets in each size category are the same as in 2005.

survive because of their comparative advantage at small business lending and the big banks would take advantage of their scale economies. Based on the stationary distribution calculated from the 2000–2005 transition probabilities, there will continue to be more mid-size banks than large banks. Indeed, the 2000–2005 stationary distribution has a higher fraction of Category 5 banks than the stationary distributions of the 1960s and 1970s.

Table 5 reports the fraction of assets held by each size category in 2005 and in the stationary distribution based on the 2000–2005 data. Interestingly, in the stationary distribution, only the largest size category holds a smaller percentage of assets than it does in 2005. This striking finding demonstrates the importance of the transition probabilities for determining the size distribution.

Finally, we report in Table 6 the estimated transition matrix for the 2000–2005 period. The first row lists the probability of a new bank forming and starting in each size category. The first column lists the probability a bank of each size category exits (and for an inactive bank, stays inactive). For the given time period, all entering banks enter into the smallest size category. The first column shows the probability of exit, either by failure or merger, from the industry. Exit is most common in the largest category and represents mergers. We see that banks in any size category are most likely to remain in the same bin as denoted by the diagonal entries. Finally, the matrix shows that banks gradually change in size. Except for exits, banks almost always stay within one size category of the previous year.

8. CONCLUSION

In this paper, we documented the large changes in the size distribution of banks that occurred starting in the 1980s. We found that the lognormal distribution poorly fits the right tail of the size distribution. Zipf's Law fits better, but for some size measures in some periods, this distribution does not fit the largest banks that well.

We also documented some differences and similarities in the size dynamics over time. First, we found that new banks are a constant fraction of the total number of banks. Second, we also found that Gibrat's Law is a good approximation for the 1960s and 1970s, before deregulation, but does not describe the 1980s and 1990s. In these decades, the large banks grow the fastest. The last five years of data suggest that the dynamics are returning to the earlier, more stable period. Of course, five years of data are not enough to make a strong prediction.

We also performed a simple forecasting exercise, using the transition probabilities taken from different time periods. Again, the relative constancy of the number of banks in the 1960s and 1970s suggests that this period was relatively stable. The projected rapid decline in the number of banks using 1980s and 1990s transition probabilities is evidence of the rapid changes that occurred in the banking industry during that time. Finally, the 2000–2005 transition probabilities predict a leveling off in the number of banks.[15] If that trend continues, then we will be returning to a relatively stable period in banking, at least as measured by the number of banks. The size dynamics imply that the U.S. banking structure will continue to have large numbers of small banks and a decent number of mid-size banks.

As illustrated by the transition probability analysis, the size distribution depends, ultimately, on the size dynamics. Therefore, a theory of the changes in bank size distribution needs an explanation of why the size dynamics changed and by how much. The data demonstrate that these changes started in the 1980s as deregulation proceeded, so the natural place to start is with an understanding of how removals to growth and size limits change the growth rates of different size banks. A successful theory would also need to account for the robust entry over this period, despite the large number of banks that exited.

[15] Jones and Critchfield (2005) make a similar prediction, using a different forecasting method.

Table 6 Transition Probability Matrix for Bank Assets: 2000–2005

Size Categories	Inactive	1	2	3	4	5	6	7
Inactive	**0.999**	**0.001**	0.000	0.000	0.000	0.000	0.000	0.000
1	**0.024**	**0.927**	**0.049**	0.000	0.000	0.000	0.000	0.000
2	**0.030**	**0.043**	**0.908**	**0.018**	0.000	0.000	0.000	0.000
3	**0.040**	0.000	**0.076**	**0.827**	**0.058**	0.000	0.000	0.000
4	**0.050**	0.000	0.003	**0.048**	**0.879**	**0.020**	0.001	0.000
5	**0.028**	0.000	0.000	0.000	**0.060**	**0.827**	**0.086**	0.000
6	**0.053**	0.000	0.000	0.000	0.000	**0.068**	**0.849**	**0.030**
7	**0.061**	0.000	0.000	0.000	0.000	0.000	**0.020**	**0.920**

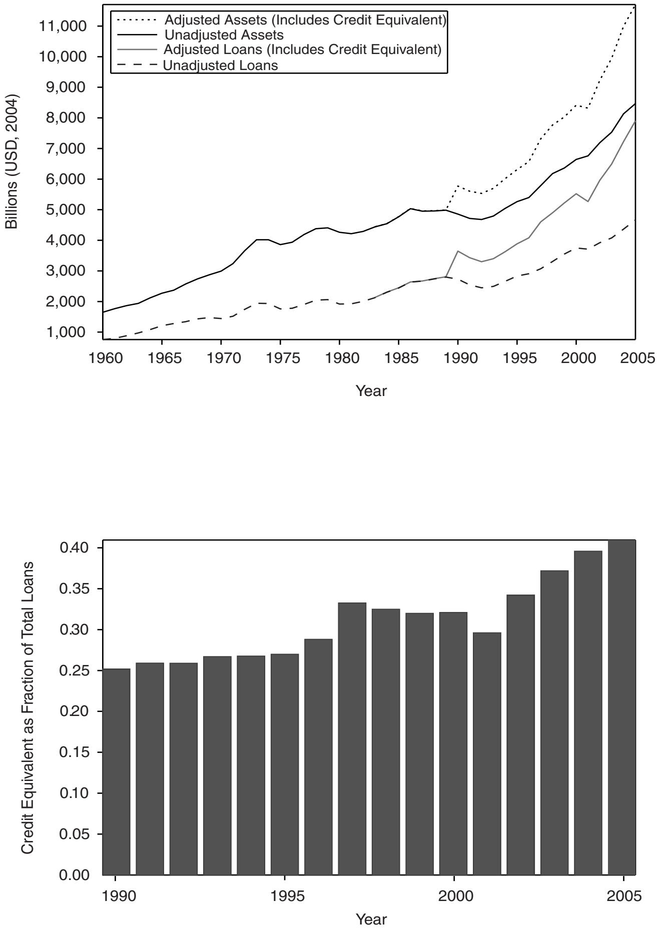
Notes: This matrix has the property that a bank starting in any size category will reach with positive probability any other size category in a finite number of periods. Probabilities that are significantly different than zero are highlighted in bold.

APPENDIX: OFF-BALANCE-SHEET ITEMS

Banks can make commitments that are not directly measured by a traditional balance sheet. For example, a loan commitment is a promise to make a loan under certain conditions. Traditionally, this kind of promise was not measured as an asset on a balance sheet. As documented by Boyd and Gertler (1994), providing this and other off-balance-sheet items have become an important service provided by banks, which means that traditional balance sheet numbers did not accurately report some of the implicit assets and liabilities of a bank.

We account for loan commitment and other off-balance-sheet items such as derivatives by converting them into *credit equivalents* and then adding them to on-balance-sheet assets and loans. This method is similar to the "Basel Credit Equivalents" series found in Boyd and Gertler (1994). Off-balance-sheet items are weighted by a credit conversion factor to create credit equivalents. We make these adjustments starting in 1989 because it is only from this year that we have the complete data to make them. Both panels of Figure 10 demonstrate the importance of the adjustment by plotting aggregate assets and loans with and without the adjustment, as well as by plotting credit equivalents as a share of total loans. A detailed list of off-balance-sheet items and credit equivalent weights is found in Table 7. These weights are used by federal regulators to determine credit equivalents for regulatory capital purposes.

Figure 10 Total Credit Equivalents in Assets and Loans



Notes: Top panel reports assets and loans both unadjusted and adjusted for off-balance-sheet items. Bottom panel reports the fraction of adjusted loans that are due to the credit equivalent adjustment. Credit equivalents are based on the weights used by regulators to determine regulatory capital requirements.

Table 7 Off-Balance-Sheet Items and Credit Equivalents

Item	Conversion Factor
Financial Standby Letters of Credit	1.00
Performance and Standby Letters of Credit	0.50
Commercial Standby Letters of Credit	0.20
Risk Participations in Bankers' Acceptances	1.00
Securities Lent	1.00
Retained Recourse on Small Business Obligations	1.00
Recourse and Direct Credit Substitutes	1.00
Other Financial Assets Sold with Recourse	1.00
Other Off-Balance-Sheet Liabilities	1.00
Unused Loan Commitments (maturity >1 year)	0.50
Derivatives	–

Notes: Conversion factors used by regulators for determining credit equivalents of off-balance-sheet items. The source is FFIEC 041 Schedule RC-R retrieved from: www.ffiec.gov/forms041.htm (accessed on November 10, 2005).

REFERENCES

Adelman, Irma G. 1958. "A Stochastic Analysis of the Size Distribution of Firms." *Journal of the American Statistical Association* 53: 893–904.

Alhadeff, David, and Charlotte Alhadeff. 1964. "Growth of Large Banks, 1930–1960." *Review of Economics and Statistics* 46: 356–63.

Axtell, Robert L. 2001. "Zipf Distribution of U.S. Firm Sizes." *Science* 293: 1818–20.

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.

Boyd, John H., and Mark Gertler. 1994. "Are Banks Dead? Or Are the Reports Greatly Exaggerated?" Federal Reserve Bank of Minneapolis *Quarterly Review* 18 (Summer): 2–23.

Ennis, Huberto M. 2001. "On the Size Distribution of Banks." Federal Reserve Bank of Richmond *Economic Quarterly* 87 (Fall): 1–25.

Gabaix, X. 1999. "Zipf's Law for Cities: An Explanation." *Quarterly Journal of Economics* 114: 739–67.

Gabaix, X., and Yannis Ioannides. 2004. "The Evolution of City Size Distributions." In J. Vernon Henderson and Jacques-Francois Thisse (eds.) *Handbook of Regional and Urban Economics IV*. North-Holland.

Gibrat, Robert. 1931. *Les Inégalités Économiques*. Paris, France: Librairie du Recueil Sirey.

Goddard, John A., Donald G. McKillop, and John O. S. Wilson. 2002. "The Growth of U.S. Credit Unions." *Journal of Banking and Finance* 26: 2327–56.

Hart, P. E., and S. J. Prais. 1956. "The Analysis of Business Concentration: A Statistical Approach." *Journal of the Royal Statistical Association* 119: 150–91.

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

Jayaratne, Jith, and Philip E. Strahan. 1997. "The Benefits of Branching Deregulation." Federal Reserve Bank of New York *Economic Policy Review* 3 (December): 13–29.

Jones, Kenneth D., and Tim Critchfield. 2005. "Consolidation in the U.S. Banking Industry: Is the 'Long, Strange Trip' About to End?" FDIC *Banking Review* 17: 31–61.

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

Kane, Edward J. 1996. "De Jure Interstate Banking: Why Only Now?" *Journal of Money, Credit and Banking* 28: 141–61.

Mengle, David L. 1990. "The Case for Interstate Branch Banking." Federal Reserve Bank of Richmond *Economic Review* 76 (November/December): 3–17.

Newman, M. E. J. 2005. "Power Laws, Pareto Distributions and Zipf's Law." *Contemporary Physics* 46: 323–51.

Rhoades, S. A., and A. J. Yeats. 1974. "Growth, Consolidation and Mergers in Banking." *Journal of Finance* 29: 1397–1405.

Robertson, Douglas D. 2001. "A Markov View of Bank Consolidation: 1960–2000." OCC Economic and Policy Analysis Working Paper 2001-4.

Rossi-Hansberg, Esteban, and Mark L. J. Wright. 2006. "Establishment Size Dynamics in the Aggregate Economy." Mimeo, Princeton University.

Saunders, Anthony, and Ingo Walter. 1994. *Universal Banking in the United States*. New York, NY: Oxford University Press.

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

Spong, Kenneth. 2000. *Banking Regulation*. 5th ed. Kansas City, MO: Federal Reserve Bank of Kansas City.

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

Tschoegl, Adrian E. 1983. “Size, Growth, and Transnationality Among the World’s Largest Banks.” *Journal of Business* 56: 187–201.

Bond Price Premiums

Alexander L. Wolman

This article provides a detailed introduction to consumption-based bond pricing theory, a special case of the consumption-based asset pricing theory associated with Robert Lucas (1978). To help make the theory more accessible to novices, we organize the article around the two famous interest rate decompositions associated with Irving Fisher. These complementary decompositions relate real or nominal long-term interest rates to expected future short-term interest rates (the expectations theory of the term structure), and relate short- or long-term nominal interest rates to the ex ante real interest rate and the expected inflation rate (the Fisher equation). According to consumption-based theory, the Fisherian relationships hold exactly only under certain restrictive conditions. We show what those conditions are, and we show that generalizations of the Fisherian relationships hold quite broadly in the consumption-based model.

The pure Fisherian relationships are shown to hold only as special cases of the relationship between individual preferences, future economic activity, and the returns on assets. Notable sufficient conditions for the pure expectations hypothesis are that households be neutral to risk and the price level behave like a random walk; the pure Fisher equation requires only risk neutrality.[1] In turn, long-term nominal bond prices may lie above or below the values dictated by the pure expectations hypothesis and the pure Fisher relationship—forward premiums and inflation-risk premiums may be positive or negative.

Interpreting bond prices of various maturities is an important challenge for the Federal Reserve. Nominal bond prices contain information about the public's expectations of inflation and of future short-term rates. And they contain information about the levels of short-term and long-term real interest rates. All these variables can be valuable signals to the Federal Reserve of the

For helpful comments, we would like to thank Kartik Athreya, Brian Minton, Ned Prescott, and Roy Webb. This article does not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Risk aversion lies at the heart of much of asset pricing theory. For example, it is what leads us to assume that riskier assets will have a higher average return than safer assets.

appropriateness of its policy.[2] However, extracting these signals requires an understanding of the potential limitations of the pure expectations hypothesis and the pure Fisher relationship.[3]

The article proceeds as follows. In Section 1 we provide a brief historical overview of the two interest rate decompositions. Section 2 lays out a modeling framework for thinking about bond price determination, and derives the basic bond pricing equations from which all else will follow. Section 3 derives the generalized expectations theory of the term structure and Section 4 derives the generalized Fisher equation. Section 5 combines the results of the previous two sections for a general discussion of the yield differential between short- and long-term bonds. Sections 2–5 provide a textbook treatment of bond pricing relationships.[4] Section 6 provides a selective review of applied research based on bond pricing theory. Section 7 concludes the article.

Although the usual statements of the expectations hypothesis and the Fisher equation are made in terms of interest rates, most of our derivations use zero-coupon bond prices. This is for analytical simplicity; working with bond prices is slightly easier, especially when the bonds are zero-coupon bonds. And given an expression for the price of a bond, one can always work out the corresponding interest rate.

1. BRIEF HISTORY OF INTEREST RATE DECOMPOSITIONS

The expectations hypothesis of the term structure and the Fisher equation both made early appearances in Irving Fisher's *Appreciation and Interest* (1896).[5] Chapter 2 of that work is devoted to a discussion of the equation, or "effect," that would later bear the author's name. The Fisher equation is typically thought of as relating "real" and "nominal" interest to the expected rate of inflation, but Fisher's analysis in *Appreciation and Interest* is more general. He relates the interest rates between two standards (for example real vs. nominal, or dollars vs. yen) to the relative rate of appreciation of the standards, as

$$1 + j = (1 + a)(1 + i), \tag{1}$$

[2] See Bernanke and Woodford (1997), however, on the risks involved in the Federal Reserve basing its policy actions solely on such data.

[3] The empirical limitations of the pure expectations hypothesis have been well documented, for example, by Campbell and Shiller (1991). There has been less emphasis on violations of the Fisher relationship. Sarte (1998) presents evidence that the violations are small using a standard form of preferences. Kim and Wright's (2005) results suggest larger violations, using a different approach outlined in Section 6.

[4] Textbook treatments are available, see for example, Sargent (1987) and Cochrane (2001). However, they tend to provide fewer details, concentrating instead on the method by which one can price any asset in the consumption-based framework.

[5] See Humphrey (1983) for the intellectual history of the Fisher equation before Fisher. Humphrey shows that the relationship was well understood before Fisher.

where i is the rate of interest in the appreciating standard, a is the rate of appreciation, and j is the rate of interest in the depreciating standard. In Fisher's words,

> The rate of interest in the (relatively) depreciating standard is equal to the sum of three terms, viz., the rate of interest in the appreciating standard, the rate of appreciation itself, and the product of these two elements. (p. 9)

In our context, j is the nominal rate, i is the real rate, and a is the expected inflation rate.

In Chapter 5 and to some extent in Chapters 3 and 4, one can find the essence of the expectations hypothesis of the term structure. Most notably perhaps, on pages 28 and 29, Fisher writes,

> A government bond, for instance, is a promise to pay a specific series of future sums, the price of the bond is the present value of this series and the "interest realized by the investor" as computed by actuaries is nothing more or less than the "average" rate of interest in the sense above defined.

By " 'average' rate of interest in the sense above defined," Fisher means what we now understand to be the expected future path of short-term rates.

John Hicks (1939) and F. A. Lutz (1940) elaborated on Fisher's version of the expectations hypothesis in the 1930s and 1940s. Their versions of these interest rate decompositions continued to be based on reasoning regarding how returns among different assets should be related. Later, the development of consumption-based asset pricing theory (Lucas 1978) gave a formal foundation to Fisher's reasoning, while making clear that restrictive assumptions were needed for the Fisherian relationships to hold exactly. The discipline provided by consumption-based theory and the rise of rational expectations and dynamic equilibrium modeling in macroeconomics also led economists and finance theorists to de-emphasize certain elements of Fisher's theories regarding interest rates. For example, with respect to the expectations hypothesis, early versions were "usually understood to imply . . . that interest rates on long-term securities will move less, on the average, than rates on short-term securities" (Wood 1964). It is now well understood that whether this will be true depends on the behavior of monetary policy and on the real shocks hitting the economy (Watson 1999). And with respect to the Fisher equation, prior to the consumption-based theory, researchers often emphasized not just the decomposition into real rates and expected inflation, but also the extent to which the real rate was invariant to changes in expected inflation (Mundell 1963). It is now understood that one cannot make general statements about this invariance, even though a version of the Fisher equation holds under general conditions.

2. MODELING FRAMEWORK

We use the modern theory of consumption-based asset pricing, first developed by Robert Lucas (1978), to study bond prices. For our purposes, the crucial elements in this theory are as follows. There is a representative consumer who has an infinite planning horizon, has a standard utility function (exhibiting risk aversion) over consumption each period, and discounts the utility from future consumption at a constant rate.[6] The consumer has a budget constraint which states that the sum of income from sales of real and financial assets (including income from maturing bonds), and income from other sources must not be exceeded by the sum of spending on current consumption, on purchases of real and financial assets, and on any other uses. With this framework, it is possible to price any asset. To do this, we use conditions describing individuals' optimal behavior.

Preferences and Budget Constraint

The consumer's preferences in period t are given by

$$v_t = E_t \sum_{j=0}^{\infty} \beta^j u\left(c_{t+j}\right), \tag{2}$$

where $u\left(c_{t+j}\right)$ is a utility function that is increasing and strictly concave in consumption (c), and the discount factor $\beta \in (0, 1)$. Before specifying the budget constraint, it will be helpful to provide more detail about the set of financial assets that play a role in our analysis. They are

1. n-period real discount bonds; if issued in period t, they pay off one unit of consumption with certainty in period $t + n$. Their price in period t in terms of goods is $q_t^{(n)}$, and the quantity that the consumer purchases in period t is $b_t^{(n)}$.

2. n-period nominal discount bonds; if issued in period t, they pay off a dollar with certainty in period $t + n$. Their dollar price in period t is denoted $Q_t^{(n)}$, and the quantity that the consumer purchases in period t is denoted $B_t^{(n)}$. Notice that if the dollar-denominated price of consumption at date $t + n$ is very high, a nominal bond provides little

[6] The representative consumer idea can be taken literally or can be viewed as a shortcut for the assumption that whatever individual-level heterogeneity does exist has been insured away by the existence of a complete set of Arrow-Debreu securities (state-contingent claims).

consumption. Therefore, an asset that yields a dollar with certainty is still a "risky" asset.

3. One-period nominal forward contracts, s periods ahead; these contracts represent a commitment in period t to purchase a one-period nominal discount bond in period $t+s$ at the pre-specified dollar price $Q^f_{t,s}$. The quantity that the consumer commits in period t to purchase in period $t+s$ is $B^f_{t,s}$.

4. One-period real forward contracts, s periods ahead; these contracts represent a commitment in period t to purchase a one-period real discount bond in period $t+s$ at the pre-specified price in terms of goods $q^f_{t,s}$. The quantity that the consumer agrees in period t to purchase in period $t+s$ is $b^f_{t,s}$.

With this set of assets, the consumer's flow budget constraint in period t is

$$\begin{aligned} &\textstyle\sum_{n=1}^{\infty} Q_t^{(n-1)} B_{t-1}^{(n)} + \sum_{s=1}^{t-1} B^f_{t-s-1,s} + \\ &P_t \textstyle\sum_{n=1}^{\infty} q_t^{(n-1)} b_{t-1}^{(n)} + P_t \sum_{s=1}^{t-1} b^f_{t-s-1,s} + W_t \\ &= P_t c_t + \textstyle\sum_{n=1}^{\infty} Q_t^{(n)} B_t^{(n)} + \sum_{s=1}^{t} Q^f_{t-s,s} B^f_{t-s,s} + \\ &P_t \textstyle\sum_{n=1}^{\infty} q_t^{(n)} b_t^{(n)} + P_t \sum_{s=1}^{t} q^f_{t-s,s} b^f_{t-s,s} + Z_t, \end{aligned} \tag{3}$$

where P_t is the price level—the dollar price of the consumption good, Z_t is purchases of other assets, and W_t is labor income and income from all other sources. The left-hand side of (3) represents income and the right-hand side represents spending.

There are several things to note with regard to the budget constraint. First, in period t, for any $j > k > 0$, a j-period bond issued in period $(t-k)$ is identical to a $(j-k)$-period bond issued in period t, because they have the same maturity and the same payoff at maturity.[7] Thus, in the budget constraint we include only the latter on the right-hand side. Second, a discount bond that matures in period t can be thought of as having a price of one dollar (for a nominal bond) or one good (for a real bond) in period t. Thus, on the left-hand side of the budget constraint we have imposed $Q_t^{(0)} = q_t^{(0)} = 1$. Third, the prices of nominal bonds are written in terms of dollars and the prices of real bonds are written in terms of goods; with the budget constraint written in nominal terms, this means that prices of real bonds must be multiplied by P_t. Finally, it is important to be clear about the forward contracts that appear in the budget constraint. On the left-hand side, the terms $\sum_{s=1}^{t-1} B^f_{t-s-1,s}$ and

[7] For example, set $j = 10$, $k = 3$ and $t = 20$. In period 20, a ten-period bond issued in period 17 is identical to a seven-period bond issued in period 20.

$P_t \sum_{s=1}^{t-1} b^f_{t-s-1,s}$ represent income from maturing one-period bonds that were purchased under forward contracts entered into in periods earlier than $t - 1$. On the right-hand side, the terms $\sum_{s=1}^{t} Q^f_{t-s,s} B^f_{t-s,s}$ and $P_t \sum_{s=1}^{t} q^f_{t-s,s} b^f_{t-s,s}$ represent purchases of one-period bonds under forward contracts entered into in periods earlier than t. For example, in period t, the consumer purchases a quantity $B^f_{t-2,2}$ of one-period bonds at price $Q^f_{t-2,2}$ in accordance with a forward contract entered into in period $t-2$. Similarly, the consumer purchases a quantity $B^f_{t-3,3}$ of one-period bonds at a price $Q^f_{t-3,3}$ in accordance with a forward contract entered into in period $t - 3$. Forward contracts entered into in period t do not appear in the period t budget constraint because they do not affect income or spending in period t; they do show up in future budget constraints.

As mentioned earlier, by limiting our attention to zero-coupon bonds, it is natural to focus on bond prices rather than interest rates. However, one can easily recover interest rates from bond prices. Let $R_t^{(n)}$ denote the gross nominal yield on a bond that sells in period t for price $Q_t^{(n)}$ and pays one dollar in period $t + n$. On a standardized per-period basis, the yield satisfies

$$R_t^{(n)} = \left(1/Q_t^{(n)}\right)^{1/n}; \tag{4}$$

$R_t^{(n)}$ is the constant per-period interest rate that is implied by the price $Q_t^{(n)}$. Likewise, for an n-period real bond we have

$$r_t^{(n)} = \left(1/q_t^{(n)}\right)^{1/n}, \tag{5}$$

and for one-period nominal and real forward contracts entered into in period $t - s$ for execution in period t, we have

$$R^f_{t-s,s} = 1/Q^f_{t-s,s}, \tag{6}$$

and

$$r^f_{t-s,s} = 1/q^f_{t-s,s}. \tag{7}$$

Individual Optimality Conditions

The consumer chooses consumption and holdings of each asset to maximize expected utility subject to the sequence of flow budget constraints. One way to carry out this maximization is to form a Lagrangian from the utility function and the sequence of budget constraints, and then use first-order conditions for consumption and each asset. The Lagrangian is

$$L_t = E_t \sum_{j=0}^{\infty} \beta^j \left[u\left(c_{t+j}\right) + \right. \tag{8}$$

$$\Lambda_{t+j}\cdot\left\{\sum_{n=1}^{\infty}Q_{t+j}^{(n-1)}B_{t+j-1}^{(n)}+\sum_{s=1}^{t+1}B_{t+j-s-1,s}^{f}+\right.$$
$$\left.P_{t+j}\sum_{n=1}^{\infty}q_{t+j}^{(n-1)}b_{t+j-1}^{(n)}+P_{t+j}\sum_{s=1}^{t+1}b_{t+j-s-1,s}^{f}+W_{t+j}\right\}$$
$$-\Lambda_{t+j}\cdot\left\{P_{t+j}c_{t+j}+\sum_{n=1}^{\infty}Q_{t+j}^{(n)}B_{t+j}^{(n)}+\sum_{s=1}^{t}Q_{t+j-s,s}^{f}B_{t+j-s,s}^{f}+\right.$$
$$\left.P_{t+j}\sum_{n=1}^{\infty}q_{t+j}^{(n)}b_{t+j}^{(n)}+P_{t+j}\sum_{s=1}^{t}q_{t+j-s,s}^{f}b_{t+j-s,s}^{f}+Z_{t+j}\right\}].$$

The first-order condition for consumption in period t is

$$u'(c_t)=P_t\Lambda_t. \tag{9}$$

The multiplier Λ_t is the marginal utility of nominal income.

Nominal and Real Bonds

The first-order conditions for n-period nominal bonds are

$$\Lambda_t Q_t^{(n)}=\beta E_t\left[\Lambda_{t+1}Q_{t+1}^{(n-1)}\right],\ n=1,2,... \tag{10}$$

where we have used the fact that an n-period bond in period t becomes an $n-1$ period bond in period $t+1$. This expression implies that the price in period t of an n-period discount bond is the ratio of the present value of expected marginal utility in period $t+n$ to marginal utility in period t :

$$Q_t^{(n)}=\beta^n E_t\left[\Lambda_{t+n}/\Lambda_t\right],\ n=1,2,... \tag{11}$$

To show this, first write (10) for period $t+1$:

$$\Lambda_{t+1}Q_{t+1}^{(n)}=\beta E_{t+1}\left[\Lambda_{t+2}Q_{t+2}^{(n-1)}\right],\ n=1,2,... \tag{12}$$

and substitute the result into (10), dividing both sides by Λ_t and using the law of iterated expectations:

$$Q_t^{(n)}=\beta^2 E_t\left[(\Lambda_{t+2}/\Lambda_t)\,Q_{t+2}^{(n-1)}\right],\ n=1,2,... \tag{13}$$

If $n=1$ then we have (11), because $Q_{t+2}^{(0)}=1$. If $n>1$ then repeat the process, substituting for $\Lambda_{t+2}Q_{t+2}^{(n-1)}$ using (10), etc. Intuitively, $\Lambda_t Q_t^{(n)}$ is the utility cost of a bond in period t, and $\beta^n E_t\left[\Lambda_{t+n}\right]$ is the expected utility benefit from the payoff at maturity, discounted back to the present. When the agent has optimized over bond holdings these two values are identical.[8] Holding

[8] The object on the right-hand side of (11) is referred to as the intertemporal marginal rate of substitution.

constant the current marginal utility of a dollar (Λ_t), a higher bond price $Q_t^{(n)}$ will correspond to a higher payoff in utility terms, that is a higher $u'(c_{t+n})$ or a lower P_{t+n}.

For n-period real bonds the derivations are analogous, though it will be useful to define the marginal utility of consumption as $\lambda_t \equiv P_t \Lambda_t$. The price in period t of an n-period real discount bond is the ratio of the present value of expected marginal utility in period $t + n$ to marginal utility in period t:

$$q_t^{(n)} = \beta^n E_t \left[\lambda_{t+n}/\lambda_t\right],\ n = 1, 2, ... \tag{14}$$

In contrast to the price of a nominal bond, which depends on the joint properties of the marginal utility of consumption ($\lambda_t = u'(c_t)$) and the price level, the price of a real bond depends only on the marginal utility of consumption.

Forward Contracts

As for forward contracts, the first-order condition for s-period-ahead one-period nominal forward contracts, committed to in period t is

$$E_t \Lambda_{t+s} Q_{t,s}^f = \beta E_t \Lambda_{t+s+1}.$$

However, because the forward price is known in period t, we can bring the price outside the expectation operator:

$$Q_{t,s}^f E_t \Lambda_{t+s} = \beta E_t \Lambda_{t+s+1}. \tag{15}$$

Likewise for real forward contracts, we have

$$q_{t,s}^f E_t \lambda_{t+s} = \beta E_t \lambda_{t+s+1}. \tag{16}$$

From (11) and (15), the price of an n-period nominal bond is identical to the product of the prices of a sequence of one-period forward contracts,

$$Q_t^{(n)} = Q_t^{(1)} Q_{t,1}^f Q_{t,2}^f \cdots Q_{t,n-1}^f. \tag{17}$$

An n-period bond and a sequence of forward contracts can each be used to provide a certain return n-periods ahead. To get a dollar in $t + n$ using the n-period bond, one needs to spend $Q_t^{(n)}$ today, whereas to get a dollar in $t+n$ using the sequence of forward contracts, one needs to spend $Q_t^{(1)} Q_{t,1}^f Q_{t,2}^f \cdots Q_{t,n-1}^f$ today. This is easily illustrated in the two-period case: to get a dollar in $t + 2$ using forward contracts, one needs to have $Q_{t,1}^f$ in period $t + 1$— for this is the forward price of a bond which will deliver a dollar in period $t + 2$. In turn, receiving $Q_{t,1}^f$ in period $t + 1$ means spending $Q_t^{(1)} Q_{t,1}^f$ in period t on one-period bonds—the price of a bond that delivers a dollar in $t + 1$ is $Q_t^{(1)}$, and one needs to purchase $Q_{t,1}^f$ of these bonds. True arbitrage would be possible if $Q_t^{(1)} Q_{t,1}^f$ were not equal to $Q_t^{(2)}$. The same reasoning holds for a long-term real bond and a sequence of real forward contracts, so we have

$$q_t^{(n)} = q_t^{(1)} q_{t,1}^f q_{t,2}^f \cdots q_{t,n-1}^f. \tag{18}$$

Note that from (17) or (18), the ratio in period t of the price of an n-period bond to the price of an $n-1$ period bond is equal to the forward price of one-period bond in period $t+n-1$, as of period t.

The optimality conditions (11) – (16) and the relationships between prices of long bonds and forward contracts (17) and (18) serve as the basis for the generalized Fisher relationship and generalized expectations theory.

3. EXPECTATIONS THEORY OF THE TERM STRUCTURE

The standard version of the expectations theory of the term structure states that long-term interest rates are equal to an average of expected future short-term interest rates. We will derive a generalization of this theory, focusing on bond prices instead of rates, and we will see that only under certain conditions does the pure expectations theory hold. Our derivation exploits the fact that a long bond is equivalent to a sequence of forward contracts.

From (17), the price of an n-period bond is the product of the prices of n short-term forward contracts. Under the pure expectations hypothesis, the price of an n-period bond is also equal to the product of the expected prices of future short-term bonds, which we will denote by $PEH^{(n)}$, for pure expectations hypothesis:

$$PEH_t^{(n)} = Q_t^{(1)} E_t\left(Q_{t+1}^{(1)}\right) E_t\left(Q_{t+2}^{(1)}\right) \cdots E_t\left(Q_{t+n-1}^{(1)}\right). \tag{19}$$

In a way that we will make more precise shortly, the pure expectations hypothesis holds if covariances involving future bond prices and future marginal utility are zero. It follows from the previous equation and (17) that the deviation of the price of an n-period nominal bond from $PEH^{(n)}$ is the product of ratios of forward prices to expected future spot prices:

$$\frac{Q_t^{(n)}}{PEH_t^{(n)}} = \frac{Q_{t,1}^f}{E_t\left(Q_{t+1}^{(1)}\right)} \frac{Q_{t,2}^f}{E_t\left(Q_{t+2}^{(1)}\right)} \cdots \frac{Q_{t,n-1}^f}{E_t\left(Q_{t+n-1}^{(1)}\right)}, \tag{20}$$

or, in shorthand,

$$\frac{Q_t^{(n)}}{PEH_t^{(n)}} = F_{t,1}^{(1)} \cdots F_{t,n-1}^{(1)}, \tag{21}$$

where we call $F_{t,j}^{(1)}$ the j-period-ahead forward premium,

$$F_{t,j}^{(1)} \equiv \frac{Q_{t,1}^f}{E_t\left(Q_{t+1}^{(1)}\right)}. \tag{22}$$

In terms of marginal utilities, using (11) and (15), the forward premium is

$$F_{t,j}^{(1)} = \frac{\frac{E_t \Lambda_{t+j+1}}{E_t \Lambda_{t+j}}}{E_t\left(\frac{\Lambda_{t+j+1}}{\Lambda_{t+j}}\right)}, \tag{23}$$

and it is straightforward to show that the forward premium is pinned down by the autocovariance properties of marginal utility, or equivalently by the covariance between the future short-term bond price and future marginal utility:

$$
\begin{aligned}
F_{t,j}^{(1)} &= \frac{E_t\left(\frac{\Lambda_{t+j+1}}{\Lambda_{t+j}}\Lambda_{t+j}\right)}{E_t\left(\frac{\Lambda_{t+j+1}}{\Lambda_{t+j}}\right)E_t\Lambda_{t+j}} && (24)\\
&= 1+\operatorname{cov}_t\left(\frac{\left(\Lambda_{t+j+1}/\Lambda_{t+j}\right)}{E_t\left(\Lambda_{t+j+1}/\Lambda_{t+j}\right)}, \frac{\Lambda_{t+j}}{E_t\Lambda_{t+j}}\right)\\
&= 1+\operatorname{cov}_t\left(\frac{Q_{t+j}^{(1)}}{E_tQ_{t+j}^{(1)}}, \frac{\Lambda_{t+j}}{E_t\Lambda_{t+j}}\right), && (25)
\end{aligned}
$$

with the last equality following from the law of iterated expectations.[9]

The deviation of the long bond price from the pure expectations hypothesis is thus accounted for by the product of the individual forward premiums ($F_{t,j}^{(1)}$),

$$
Q_t^{(n)} = PEH_t^{(n)} \times \left(F_{t,1}^{(1)} \cdots F_{t,n-1}^{(1)}\right). \qquad (26)
$$

If each of the individual covariances that determine the forward premiums are zero, then the pure expectations hypothesis holds. In turn, forward premiums will be zero if the level of future nominal marginal utility is uncorrelated with its subsequent growth rate. This will be the case, for example, if nominal marginal utility is constant, or if it follows a random walk. Note that for nominal bonds, risk neutrality is insufficient to drive all forward premiums to zero; if the future price level is correlated with its subsequent growth rate, there will be a forward premium, even if investors are risk-neutral.[10] For the case of risk aversion, the behavior of the marginal utility of consumption is crucial for determining the forward premium as well as the inflation-risk premium derived below. In standard models along the lines of Lucas (1978), the marginal utility of consumption is a simple function of consumption itself. Alternatively, one can consider more complicated specifications of $u(c)$, or be entirely agnostic on the specification of $u(c)$. These approaches are discussed in Section 6.

Why do the conditional covariances between future marginal utility and the subsequent growth rate of marginal utility affect the price of a long-term bond relative to the product of expected future short-term bond prices? Focus on one term ($F_{t,j}^{(1)}$), which is the price premium for a j-period-ahead forward

[9] From the law of iterated expectations we know, for example, that $E_t(\Lambda_{t+j+1}) = E_t(\Lambda_{t+j}E_{t+j}\left(\Lambda_{t+j+1}/\Lambda_{t+j}\right)) = E_t(\Lambda_{t+j}Q_{t+j}^{(1)})$.

[10] Under risk-neutrality, expected real returns are equated across assets (forward and spot, for example). Because the real return is the nominal return *divided by* the gross inflation rate, expected nominal returns are not necessarily equated.

contract relative to the expected j-period-ahead spot price of a one-period bond. If the growth rate of the marginal utility of a dollar (price of a one-period bond) is expected to covary positively with the level of the marginal utility of a dollar in $t + j$, then you pay a premium at $t + j$ to lock in at t the contract that gives you a dollar at $t + j + 1$. In this situation, you tend to value a dollar highly for consumption in $t + j$ precisely when buying a bond requires you to forego a lot of consumption. Thus, the expected spot market looks expensive, which means that the forward price must be high as well.

Note that the j-period-ahead forward premium can be positive or negative, and thus the overall term premium can be positive or negative. It is common to think of long *rates* as incorporating a positive term premium (meaning that $\prod F^1_{t,j} < 0$), but if future marginal utility of a dollar is positively correlated with the expected growth rate of marginal utility, then forward premiums and the term premium in rates will be negative.

Of course, we can also think about the forward-spot relationship from a no-arbitrage perspective: agents must be indifferent between committing to buy a one-period bond in period $t + j$ (the forward contract) and expecting to buy a one-period bond in the spot market at $t + j$:

$$Q^f_{t,j} E_t \Lambda_{t+j} = E_t \Lambda_{t+j} Q^{(1)}_{t+j}.$$

Expanding the right-hand side,

$$Q^f_{t,j} E_t \Lambda_{t+j} = E_t \Lambda_{t+j} E_t Q^{(1)}_{t+j} + \text{cov}_t \left(\Lambda_{t+j}, Q^{(1)}_{t+j} \right). \qquad (27)$$

Dividing both sides by $E_t \Lambda_{t+j} E_t Q^{(1)}_{t+j}$ replicates the expression for $F^1_{t,j}$ above.

4. FISHER RELATIONSHIP

The pure Fisher relationship states that nominal interest rates are equal to real rates plus expected inflation. We will derive a generalization of this expression, focusing again on bond prices instead of interest rates. The derivation follows directly from manipulating the pricing equation for a nominal bond.

From the fact that the real and nominal multipliers are related by $\frac{\lambda_t}{P_t} = \Lambda_t$, we can use (11) to write the price of a nominal bond ($\beta^n E_t \left[\Lambda_{t+n}/\Lambda_t\right]$) as

$$Q^{(n)}_t = \beta^n E_t \left[\frac{\lambda_{t+n}}{\lambda_t} \frac{P_t}{P_{t+n}} \right], \qquad (28)$$

or

$$Q^{(n)}_t = \beta^n \left\{ E_t \left[\frac{\lambda_{t+n}}{\lambda_t} \right] E_t \left[\frac{P_t}{P_{t+n}} \right] + \text{cov}_t \left[\frac{\lambda_{t+n}}{\lambda_t}, \frac{P_t}{P_{t+n}} \right] \right\}. \qquad (29)$$

This last expression can be used to decompose the price of a long-term nominal bond into the price of a long-term real bond ($q^{(n)}_t$ from [14]), the expectation of the inverse of inflation ($E_t \left(P_t/P_{t+n}\right)$), and a term we will call $\Theta^{(n)}_t$, which can be thought of as the inflation-risk premium:

$$\begin{aligned} Q_t^{(n)} &= \left\{ q_t^{(n)} E_t \left(P_t / P_{t+n} \right) \right\} \left(1 + \Theta_t^{(n)} \right) \\ \Theta_t^{(n)} &\equiv \beta^n \mathrm{cov}_t \left[\frac{\lambda_{t+n}/\lambda_t}{E_t \left(\lambda_{t+n}/\lambda_t \right)}, \frac{P_t/P_{t+n}}{E_t \left(P_t/P_{t+n} \right)} \right]. \end{aligned} \tag{30}$$

Alternatively, converting to interest rates instead of bond prices, we have

$$R_t^{(n)} = r_t^{(n)} \left[\frac{1}{E_t \left(P_t/P_{t+n} \right) \left(1 + \Theta_t^{(n)} \right)} \right]^{1/n}. \tag{31}$$

If there is zero conditional covariance between the normalized growth rate of marginal utility and the normalized inverse of inflation, then the inflation-risk premium is zero, and we recover the pure Fisher equation. In general, though, the price of a long-term nominal bond exceeds the "Fisher price" when the covariance between the growth of real marginal utility and the inverse of inflation is positive. What is the intuition behind this covariance effect? The bond pays off one dollar. If the covariance is positive, then the consumption value of a dollar is high (low) in those states where the marginal utility of consumption is high (low). In other words, the bond pays off well in terms of consumption when you value consumption highly, so it is worth more than the Fisher price.

Note that like the forward premium, the inflation-risk premium can be positive or negative. We usually think of the inflation-risk premium in rates as being positive, which would correspond to the price premium Θ_t being negative. However, this depends entirely on whether the inverse of the future price level is negatively correlated with the future marginal utility of consumption, conditional on time-t information.

5. YIELD DIFFERENTIAL AND HOLDING-PERIOD PREMIUM

Above we provided two decompositions of the price of an n-period nominal bond. The first expressed the bond price in terms of the price of a real bond, the expected inverse of the change in the price level, and an inflation-risk premium. The second decomposition expressed the bond price in terms of the expected product of the prices of future short-term bonds and the product of individual forward premiums at each maturity. We now use these decompositions to study the yield differential between bonds of any two maturities. In addition, in this section we provide an intuitive explanation of the holding-period premium, the expected differential between the return on a long-term bond sold before it matures and the return on a shorter-term bond sold at maturity.

Yield Differential

Recall that the yield is the inverse of the price, and that we standardize yields so that they are reported on a gross per-period basis:

$$R_t^{(n)} = (Q_t^{(n)})^{-1/n}. \tag{32}$$

We then can express the bond yield using the two decompositions as

$$R_t^{(n)} = \left[\left\{q_t^{(n)} E_t \left(P_t/P_{t+n}\right)\right\}\left(1 + \Theta_t^{(n)}\right)\right]^{-1/n} \tag{33}$$

or

$$R_t^{(n)} = \left[PEH_t^n \times \left(F_{t,1}^{(1)} \cdots F_{t,n-1}^{(1)}\right)\right]^{-1/n}. \tag{34}$$

Since these expressions hold for any n, the ratio of the yield on an n-period bond to the yield on a one-period bond can be written using the Fisher decomposition as

$$\frac{R_t^{(n)}}{R_t^{(1)}} = \frac{\left\{q_t^{(1)} E_t \left(P_t/P_{t+1}\right)\right\}\left(1 + \Theta_t^{(1)}\right)}{\left[\left\{q_t^{(n)} E_t \left(P_t/P_{t+n}\right)\right\}\left(1 + \Theta_t^{(n)}\right)\right]^{1/n}} \tag{35}$$

$$= \frac{r_t^{(n)}}{r_t^{(1)}} \frac{E_t \left(P_t/P_{t+1}\right)}{\left(E_t \left(P_t/P_{t+n}\right)\right)^{1/n}} \frac{1 + \Theta_t^{(1)}}{\left(1 + \Theta_t^{(n)}\right)^{1/n}}. \tag{36}$$

The nominal yield curve slopes upward if some combination of the following is true: (i) long-term real rates exceed short-term real rates, (ii) the value of a dollar is expected to increase at a higher rate in the short term than over the long term, and (iii) the short-term inflation-risk premium in bond prices exceeds the long-term inflation-risk premium in bond prices.

Alternatively, we can use the perspective of the expectations hypothesis to write the yield differential as

$$\frac{R_t^{(n)}}{R_t^{(1)}} = \frac{Q_t^{(1)}}{\left[PEH_t^{(n)} \times \left(F_{t,1}^{(1)} \cdots F_{t,n-1}^{(1)}\right)\right]^{1/n}}. \tag{37}$$

For $n = 1$, note that $PEH = Q_t^{(1)}$, and, by convention, $F_{t,0}^{(1)} = 1$. Long-term rates exceed short-term rates if short-term bond prices are expected to fall or if forward-price premiums are negative.

Holding-Period Premium

There are many ways that one can transport money or goods from period t to period $t + j$. Until now, we have emphasized the comparison between buying a j-period bond and buying a sequence of one-period bonds. Another option, however, is to purchase an i-period bond, where $i > j$, and sell the bond in period $t + j$ for the price $Q_{t+j}^{(i-j)}$, which is uncertain as of period t. Of course, the bond pricing relationships derived previously must imply that the consumer is indifferent between these two strategies. That is,

$$E_t\left[\Lambda_{t+j}\right] = \frac{Q_t^{(j)}}{Q_t^{(i)}} E_t\left[\Lambda_{t+j} Q_{t+j}^{(i-j)}\right]. \tag{38}$$

The left-hand side is the expected payoff in period $t+j$ in utility terms to buying one j-period bond in period t for price $Q_t^{(j)}$; note that the only uncertainty with respect to this strategy involves the marginal utility of a dollar in period $t+j$. The right-hand side is the expected return in period $t+j$ in utility terms to spending the same amount, $Q_t^{(j)}$, on an i-period bond in period t and selling the bond in period $t+j$. With this strategy, there is uncertainty both about the marginal utility of a dollar in period $t+j$ and about the price at which once can sell the bond in period $t+j$.

An intuitively appealing property similar to the pure expectations hypothesis is that the expected dollar return to these two strategies should be the same. By now, it is probably clear that while the expected utility return must be the same (as reflected in [38]), the expected dollar return will generally be different for the two strategies. The dollar return to buying the j-period bond and holding it to maturity is certain and given by $1/Q_t^{(j)}$. The expected dollar return to buying the i-period bond and selling it in period $t+j$ is given by

$$\frac{E_t\left[Q_{t+j}^{(i-j)}\right]}{Q_t^{(i)}}. \tag{39}$$

So the expected "premium" for holding an i-period bond for j periods is

$$H_t^{(i,j)} = \frac{Q_t^{(j)} E_t\left[Q_{t+j}^{(i-j)}\right]}{Q_t^{(i)}}. \tag{40}$$

From (38) we can write this premium as

$$H_t^{(i,j)} = \frac{E_t\left[\Lambda_{t+j}\right] E_t\left[Q_{t+j}^{(i-j)}\right]}{E_t\left[\Lambda_{t+j} Q_{t+j}^{(i-j)}\right]} \tag{41}$$

or

$$H_t^{(i,j)} = \frac{1}{1 + \operatorname{cov}_t\left(\frac{Q_{t+j}^{(i-j)}}{E_t Q_{t+j}^{(i-j)}}, \frac{\Lambda_{t+j}}{E_t \Lambda_{t+j}}\right)} \tag{42}$$

$$= \frac{1}{1 + \operatorname{cov}_t\left(\frac{E_{t+j}\left(\Lambda_{t+i}/\Lambda_{t+j}\right)}{E_t\left(\Lambda_{t+i}/\Lambda_{t+j}\right)}, \frac{\Lambda_{t+j}}{E_t \Lambda_{t+j}}\right)}. \tag{43}$$

The holding-period premium is driven by the same uncertainty as the forward premium, except over a possibly longer horizon. If future marginal utility is positively conditionally correlated with the future price of an $(i - j)$-period bond, then the i-period bond will tend to generate capital gains when they are highly valued, so individuals will not require a high expected return. That is, the relative expected return $H_t^{(i,j)}$ will be low when $\mathrm{cov}_t\left(\frac{\Lambda_{t+j}}{E_t\Lambda_{t+j}}, \frac{Q_{t+j}^{(i-j)}}{E_t Q_{t+j}^{(i-j)}}\right)$ is positive.

Of course, we can also use our earlier derivations to express the holding-period premium in ways related to the pure expectations hypothesis and the Fisher equation. Using the definition of the pure expectations hypothesis, we have from (26)

$$H_t^{(i,j)} = \frac{E_t\left[PEH_{t+j}^{(i-j)} \times \left(F_{t+j,1}^{(1)} \cdots F_{t+j,i-j-1}^{(1)}\right)\right]}{E_t\left(Q_{t+j}^{(1)}\right) E_t\left(Q_{t+j+1}^{(1)}\right) \cdots E_t\left(Q_{t+i-1}^{(1)}\right) \times \left(F_{t,j}^{(1)} \cdots F_{t,i-1}^{(1)}\right)}. \tag{44}$$

Very loosely speaking, this expression relates the holding-period premium to the conditional covariance between expected future short prices and expected future forward premiums. Analogously, using the Fisher equation, we have from (30)

$$H_t^{(i,j)} = \frac{E_t\left(P_t/P_{t+j}\right)\left(1+\Theta_t^{(j)}\right) E_t\left[\left\{q_{t+j}^{(i-j)}\left(P_{t+j}/P_{t+i}\right)\right\}\left(1+\Theta_{t+j}^{(i-j)}\right)\right]}{q_{t,j}^{f} q_{t,j+1}^{f} \cdots q_{t,i-1}^{f} E_t\left(P_t/P_{t+i}\right)\left(1+\Theta_t^{(i)}\right)}. \tag{45}$$

Again, loosely speaking, this expression relates the holding-period premium to the conditional covariance between the future inflation-risk premium, and the pure Fisher component of the future $(i - j)$-period bond price.

6. APPLICATIONS OF THE THEORY

The derivations above provide a textbook-like guide to bond price decompositions from the perspective of consumption-based asset pricing theory. As we stated at the outset, these decompositions can be a useful input into the formulation of monetary policy, contributing to an understanding of the term structure of real and nominal interest rates and expected inflation.[11] But any contribution to our understanding of these variables requires taking the theory to the data, and this, in turn, requires making some assumptions about

[11] We emphasize the usefulness for monetary policy, but there are other applications of the theory. Many areas of economics emphasize the behavior of real interest rates, and financial market practitioners use the kind of theories outlined here to aid in the pricing of interest rate derivatives.

the unobservable variables Λ or λ, the marginal utility of nominal or real consumption. According to the pure expectations hypothesis and the pure Fisher equation, marginal utility is extraneous: armed with an estimate of expected inflation, we can directly estimate the real rate from data on nominal rates; likewise, armed with data on the term structure of nominal rates, we can directly calculate the expected path of future short-term rates. Absent risk neutrality, however, marginal utility takes center stage, for it is the covariance of the marginal rate of substitution (marginal utility growth) with the evolution of the price level that pins down the inflation-risk premium (see [30]); and it is the autocovariance properties of marginal utility that determine the forward premium (see [24]).

Researchers applying theory to data on bond prices have gone in two directions concerning the degree of structure they impose on the marginal utility of consumption. The "pure" consumption-based approach takes a stand on the form of $u(c)$ in (2) and uses data on consumption and inflation to estimate the parameters of $u(c)$ and the stochastic processes for consumption and inflation. The combination of estimated preference parameters and stochastic processes then comprise a model of bond prices; most importantly, the specification of $u(c)$ together with data on c make marginal utility "observable." Campbell (1986) is a particularly accessible example of this strand of the literature, albeit an example that includes only real bonds; he uses a simple specification of $u(c)$ and the stochastic process for consumption and derives closed form expressions for interest rates of all maturities. Campbell's paper is primarily pedagogical.[12]

Two recent papers that use more complicated forms of preferences, include nominal bonds and make a serious attempt to match data are Wachter (2006) and Piazzesi and Schneider (2006).[13] Wachter uses a habit-persistence specification similar to the one Campbell and Cochrane (1999) apply to equity pricing. She argues that the model "accounts for many features of the nominal term structure of interest rates." Most importantly, from our perspective, the model-based forward premiums that Wachter computes help to account for the empirical disparity between long-term rates and the corresponding average of expected future short-term rates—that is, the violation of the pure expectations hypothesis. However, there is still a noticeable divergence between the actual time series for short-term rates and the path implied by Wachter's model. Piazzesi and Schneider use the recursive utility preference specification of Epstein and Zin (1989) and Weil (1989). They emphasize the inflation-risk premium in long-term bonds that arises when inflation brings bad news

[12] See also Ireland (1996) and Sarte (1998).

[13] Both Wachter (2006) and Piazzesi and Schneider (2006) use preference specifications that are not encompassed by (2). However, for our purposes, we can view their approaches as involving complicated specifications of $u(c)$.

about future consumption growth. Although their model is broadly consistent with the behavior of the term structure, the short-term rates implied by their model are substantially less volatile than the data. Regarding the approach taken by Wachter (2006) and by Piazzesi and Schneider (2006), Campbell (2006) writes, "The literature on consumption-based bond pricing is surprisingly small, given the vast literature given to consumption-based models of equity markets." We can thus expect much more work of this sort in the coming years.

Ravenna and Seppälä (2006) is one recent example of studying the term structure of interest rates in a consumption-based model that endogenizes consumption—the papers mentioned in the previous two paragraphs treat consumption (and inflation) as exogenous. Ravenna and Seppälä embed the asset pricing apparatus in a New Keynesian business cycle model. They argue that their model accounts for the cyclical properties of interest rates and the rejections of the pure expectations hypothesis, but they do not provide time series comparisons of data and model-generated interest rates. Given the high degree of structure required, matching the data with this approach is a daunting task.

The second major empirical application of bond pricing theory is known as the "no-arbitrage" or "arbitrage-free" approach. With this approach, one avoids making any parametric assumptions about the form of $u(c)$. What the no-arbitrage approach does carry over from the theory laid out previously is the idea that there exists a marginal rate of substitution that prices all bonds; that is, (11) holds for some strictly positive random variable $m_t \equiv \Lambda_t / \Lambda_{t-1}$. A good introduction to this approach is Backus, Foresi, and Telmer (1998), and a recent example is Kim and Wright (2005). Those authors and many others assume that the time-series behavior of the yield curve is driven by a small number of latent factors. The arbitrage-free approach has the advantage of being able to fit observed time series on bond prices quite well, thereby opening the door to discussing relatively small changes in the term structure of real or nominal rates. For example, Kim and Wright provide time series plots of the forward rate along with the estimated expected short rate and the estimated term premium. However, this approach has limitations from the perspective of macroeconomics in that it does not provide a framework for studying the joint determination of bond prices and macroeconomic outcomes. Indeed, Duffee (2006) writes, "some readers . . . call this a nihilistic model of term premia." He views the approach in a positive light, though, as "an intermediate step in the direction of a correctly specified economic model of premia, not an end in itself."

7. CONCLUSION

Nominal and real interest rates are often viewed from the perspectives of the intuitively appealing Fisher relationship and pure expectations hypothesis. Modern asset pricing theory implies that those relationships should not be expected to hold exactly if investors are risk-averse. We have used that theory to describe how the deviations from the Fisher relationship and the pure expectations hypothesis depend on particular covariances. In the process, we have meant to provide an introduction to the consumption-based modeling of bond prices. From the standpoint of macroeconomics and monetary policy, the value of this approach is that it allows researchers to interpret the behavior of the term structure of real and nominal bond prices in ways that relate to macroeconomic activity and monetary policy.

REFERENCES

Backus, David, Silverio Foresi, and Christopher Telmer. 1998. "Discrete-Time Models of Bond Pricing." Available at: http://pages.stern.nyu.edu/˜dbackus/tuck.ps (accessed September 13, 2006).

Bernanke, Ben S., and Michael Woodford. 1997. "Inflation Forecasts and Monetary Policy." *Journal of Money, Credit and Banking* 29 (4–2): 653–84.

Campbell, John Y. 1986. "Bond and Stock Returns in a Simple Exchange Model."*Quarterly Journal of Economics* 101 (4): 785–804.

Campbell, John Y. 2006. Discussion of Monika Piazzesi and Martin Schneider, "Equilibrium Yield Curves." Available at: www.nber.org/books/macro21/campbell7-24-06comment.pdf. (accessed September 13, 2006).

Campbell, John Y., and John Cochrane. 1999. "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior." *Journal of Political Economy* 107 (2): 205–51.

Campbell, John Y., and Robert Shiller. 1991. "Yield Spreads and Interest Rate Movements: A Bird's Eye View." *Review of Economic Studies* 58 (3): 495–514.

Cochrane, John. 2001. *Asset Pricing*. Princeton, NJ: Princeton University Press.

Duffee, Gregory R. 2006. "Are Variations in Term Premia Related to the Macroeconomy?" Available at: http://faculty.haas.berkeley.edu/duffee/duffee_premia_macro.pdf (accessed September 13, 2006).

Epstein, Larry G., and Stanley E. Zin. 1989. "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework." *Econometrica* 57 (4): 937–69.

Fisher, Irving. [1896] 1997. *Appreciation and Interest*. Reprinted in vol. 1 of *The Works of Irving Fisher*, ed. William J. Barber. London: Pickering and Chatto.

Hicks, John R. 1939. *Value and Capital*. Oxford, UK: Oxford University Press.

Humphrey, Thomas M. 1983. "The Early History of the Real/Nominal Interest Rate Relationship." Federal Reserve Bank of Richmond *Economic Review* 69 (3): 2–10.

Ireland, Peter. 1996. "Long-Term Interest Rates and Inflation: A Fisherian Approach." Federal Reserve Bank of Richmond *Economic Quarterly* 82 (1): 21–35.

Kim, Don H., and Jonathan H. Wright. 2005. "An Arbitrage-Free Three-Factor Term Structure Model and the Recent Behavior of Long-Term Yields and Distant-Horizon Forward Rates." *Federal Reserve Board Finance and Economics Discussion Series* 2005-33 (August).

Lucas, Robert E., Jr. 1978. "Asset Prices in an Exchange Economy." *Econometrica* 46 (6): 1429–45.

Lutz, Friedrich A. 1940. "The Strucure of Interest Rates." *Quarterly Journal of Economics* 55 (1): 36–63.

Mundell, Robert. 1963. "Inflation and Real Interest." *Journal of Political Economy* 71 (3): 280–3.

Piazzesi, Monika, and Martin Schneider. 2006. "Equilibrium Yield Curves." *NBER Macroeconomics Annual* Vol. 21. Available at: http://www.nber.org/books/macro21/piazzesi-schneider7-22-06.pdf. (accessed September 13, 2006).

Ravenna, Federico, and Juha Seppälä. "Policy and Rejections of the Expectations Hypothesis." Available at: http://ic.ucsc.edu/%7Efravenna/home/mopoeh4.pdf (accessed September 13, 2006).

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

Sarte, Pierre-Daniel G. 1998. “Fisher’s Equation and the Inflation Risk Premium in a Simple Endowment Economy.” Federal Reserve Bank of Richmond *Economic Quarterly* 84 (4): 53–72.

Wachter, Jessica A. 2006. “A Consumption-Based Model of the Term Structure of Interest Rates.” *Journal of Financial Economics* 79 (2): 365–99.

Watson, Mark W. 1999. “Explaining the Increased Variability in Long-Term Interest Rates.” Federal Reserve Bank of Richmond *Economic Quarterly* 85 (4): 71–96.

Weil, Philippe. 1990. “Nonexpected Utility in Macroeconomics.” *Quarterly Journal of Economics* 105 (1): 29–42.

Wood, John H. 1964. “The Expectations Hypothesis, The Yield Curve and Monetary Policy.” *Quarterly Journal of Economics* 78 (3): 457–70.

Stark Optimal Fiscal Policies and Sovereign Lending

Pierre-Daniel G. Sarte

Capital income taxes are a salient feature in the taxation schemes of many modern countries, though countries make distinctions between capital gains and income earned by capital. Most countries include rents received on capital in the income calculation for each taxpayer. When the rates are averaged over the 1965–1996 period, average capital income tax rates can be moderately high in industrially advanced nations, ranging from 24.1 percent in France to 54.1 percent in the United Kingdom. When averaged across the first six years of the 1990s, average capital income tax rates for the same countries are 25 percent and 47.7 percent, respectively. Over the same periods, the United States had average capital income taxes of 40.1 percent (1965–1996) and 39.7 percent (1990–1996) (see Domeij and Heathcote 2004).

Following the work of Judd (1985), and Chamley (1986), much of the literature on optimal taxation has argued that it is not efficient to tax capital in the long run. As shown in Atkeson, Chari, and Kehoe (1999), this policy prescription is relatively robust in the sense that it holds whether agents are heterogenous or identical, the economy's growth rate is endogenous or exogenous, and the economy is open or closed.[1] At the same time, however, the notion that long-run capital taxation is inefficient arises in settings where

We wish to thank Andreas Hornstein, Juan Carlos Hatchondo, Alex Wolman, and Nashat Moin for their comments. The views expressed in this article do not necessarily represent those of the Federal Reserve Bank of Richmond, the Board of Governors of the Federal Reserve System, or the Federal Reserve System. Any errors are my own.

[1] Exceptions to this recommendation include Correia (1996), and Jones, Manuelli, and Rossi (1997), who show that the optimal long-run tax on capital differs from zero when other factors of production are either untaxed or not taxed optimally. As pointed out in Erosa and Gervais (2001), capital income taxes in an overlapping generations environment are not just distortionary, they involve some redistribution among agents. Hence, optimal steady state capital income taxes need not be zero in such a framework, as shown in the early work of Atkinson and Sandmo (1980), and later Garriga (1999), as well as in Erosa and Gervais (2002) with age-dependent taxes. Finally,

optimal policies are often extreme at shorter horizons. For instance, when the capital stock is sufficiently large and no restrictions are placed on the capital income tax, it is optimal for the government to raise all revenues through a single capital levy at date 0 and never again tax either capital or labor. For that reason, Chamley (1986) imposes a 100 percent exogenous upper bound on the capital income tax, which Chari, Christiano, and Kehoe (1994) show can be motivated by assuming that households have the option of holding onto their capital, subject to depreciation, rather than renting it to firms. With an upper bound imposed on capital income tax rates, optimal fiscal policy continues to be surprisingly stark, with the optimal capital tax rate set at confiscatory levels for a finite number of periods, after which the tax takes on an intermediate value between 0 and 100 percent for one period, and is zero, thereafter. This article points out that government lending to households, which is seldom observed in practice, plays a crucial role in generating extreme optimal fiscal policies. Absent a domestic debt instrument, more moderate capital and labor tax rates emerge as optimal, although the capital tax rate does converge to zero, asymptotically.

Without an upper bound imposed on capital income taxes, it is easy to see why a single capital levy at date 0 is optimal in the environment studied by Chamley (1986). Since the capital stock is fixed at date 0, the initial capital levy amounts to a single lump sum tax and no distortions are ever imposed on resource allocations over time. The economy, therefore, achieves a first-best optimum. That said, the presence of a debt instrument plays a key role in the implementation of a single initial capital levy. In particular, such an instrument allows the government to front-load all taxes in the initial period (equal to the net present discounted value of future government expenses), lend the proceeds to the private sector, and finance government expenditures from interest revenue on the loans. Thus, Chamley's model never generates government debt but generates government surpluses which are lent back to households.

Although Chamley's single initial capital levy allows for first-best allocations to be achieved, such a taxation scheme has evidently little to do with the kinds of policies one observes in practice. In particular, almost all countries continue to rely on distortional tax systems to finance their public expenditures. This article highlights the fact that environments that limit ex ante government lending are more apt to generate nonzero optimal distortional taxes at all dates. In particular, we provide an analysis of Chamley's taxation problem without a policy instrument that allows for sovereign lending. In that case, the government cannot carry the proceeds from a large initial tax to future

see Aiyagari (1995) for an environment with idiosyncratic uninsurable shocks where optimal capital income taxes are not zero in the long run.

periods, and the single capital levy prescribed in Chamley (1986) cannot be implemented.[2]

For the purposes of this article, we think of restrictions on sovereign lending as a (rudimentary or ad hoc) way of getting rid of extreme taxation policies in the short run.[3] More generally, gaining a better understanding of institutional or agency constraints that endogenously limit the kinds of contracts the government can write with households seems central in generating optimal fiscal policies that more closely resemble those observed in practice. It may be helpful, for instance, to consider in greater depth the kinds of frictions that may limit or impede government lending. At first glance, such frictions are not necessarily obvious. In particular, any potential commitment problems on the part of households (to repay their loans) should easily be overcome by suitable punishment such as the garnishment of wages. There have also been times in U.S. history, (such as during World War II), where the government directly owned privately operated capital.[4]

Formally, the taxation problem we study, introduced by Kydland and Prescott (1980), discusses the time inconsistency of optimal policy. While matters related to time inconsistency lie beyond the scope of this article, we use the insights of Kydland and Prescott (1980), as well as more recent work by Marcet and Marimon (1999), to present the solution to this problem in terms of stationary linear difference equations that can be solved using standard numerical methods. While Kydland and Prescott (1980) show how the taxation problem they consider can be written (and solved) as a recursive dynamic program, the article does not ultimately present properties of the solution either in the short run or long run.

What do optimal tax rates look like when one exogenously prohibits sovereign lending? Numerical simulations carried out in this article indicate that capital income tax rates are never set either at 100 percent or zero at any point during the transition to the long-run equilibrium. Furthermore, we find that labor income is subsidized in the first few periods. This feature of optimal fiscal policy drives up labor supply and allows household consumption to be everywhere above its long-run equilibrium along its transition path. Finally, we provide straightforward proof of the optimality of zero long-run capital

[2] In this case, what matters is a government's ability to bequeath revenue-generating assets to its successor that, potentially, render the use of future taxes unnecessary. Thus, optimal confiscatory short-run capital taxes would behave as stated in Chamley (1986) in environments in which the government can lend directly to firms. Alternatively, one can also imagine optimal allocations emerging in an environment in which the government directly owned the capital stock and had no disadvantage in operating production directly.

[3] With this restriction in place, an exogenously imposed upper bound on capital income tax rates is no longer necessary. Moreover, the upper limit of 100 percent imposed by Chamley (1986) is not helpful in creating moderate optimal tax rates since this limit turns out to be binding in the short run.

[4] See McGrattan and Ohanian (1999).

taxes that does not rely on the primal approach used in Chamley (1986) and summarized in Ljungqvist and Sargent (2000, chapter 12).

This article is organized as follows. Section 1 provides a brief summary of findings in the literature on optimal fiscal policy in the presence of domestic lending and borrowing. In Section 2, we describe the economic environment under consideration. Section 3 presents the Ramsey problem associated with the analysis of optimal fiscal policy. Salient properties of the long-run Ramsey equilibrium are discussed in Section 4. Section 5 gives a numerical characterization of the transitional dynamics of optimal tax rates. Section 6 offers concluding remarks.

1. A BRIEF DESCRIPTION OF STARK FISCAL POLICIES WITH SOVEREIGN LENDING

This section describes the kinds of extreme optimal fiscal policies that have been described in the optimal taxation literature. Beginning with Chamley (1986), we have already seen in the introduction that a single capital levy at date 0, if feasible, allows the economy to achieve first-best allocations. The problem, of course, is that the associated capital income tax rate might well exceed 100 percent, in which case one might interpret the tax as not only applying to capital income but more directly to the capital stock. Whether this policy is feasible ultimately depends on if the initial capital stock is large enough to finance the net present discounted value of future government expenditures. If so, the necessary revenue is raised entirely through the levy at date 0, and lent back to households with the proceeds from the loans used to finance the stream of government expenditures over time. In that sense, a need for strictly positive distortional taxes never arises.

When capital income tax rates are restricted to be at most 100 percent, Chari, Christiano, and Kehoe (1994) show that it is optimal to set tax rates at their upper limit for a finite number of periods, after which the capital tax rate takes on an intermediate value and is zero, thereafter. The intuition underlying this result relates directly to the distortional nature of capital income tax rates. In particular, having the capital tax rate positive in some period $t > 0$ distorts savings decisions, and thus, private capital allocations, in all prior periods. Hence, front-loading capital income taxes, by having the associated tax rates set at their upper limit from date 0 to some finite date $\bar{t} > 0$, distorts the least number of investment periods. This intuition is only partially complete in that household preferences also play an important role in determining the horizon over which capital income is initially taxed. When preferences are separable in consumption and leisure, it is not optimal to tax capital after the initial period, although labor taxes may be positive at all dates (see Chari, Christiano, and Kehoe 1994). Xie (1997) shows that when preferences are logarithmic in consumption less leisure, it is optimal never to tax labor while capital income tax rates (Chari, Christiano, and Kehoe 1994) hit their upper bound for a finite number of periods and are zero, thereafter.

All of the above policies have in common a radical character and a lack of resemblance to more moderate capital tax rates in practice (i.e., capital tax rates that are neither set at confiscatory rates nor zero). However, a key part underlying the mechanics of these policies relates to the fact that the government is able to build large negative debt holdings by having the capital income tax rate hit its upper limit over some initial period of time, date 0 to date $\bar{t} > 0$. In Xie (1997), it is apparent that once these negative debt holdings are large enough to finance the remaining net present discounted value of government expenditures, then no distortional taxes need ever be set again. In essence, date $\bar{t}$ is then analogous to date 0 in Chamley (1986).

The following sections examine the problem of optimal taxation initially posed by Chamley (1986), but without the policy instrument that allows the government to accumulate large negative debt holdings. Absent this instrument, numerical simulations suggest that it is possible to have more moderate taxes on capital and labor emerge as optimal at every date, without any bounds necessarily imposed on either capital or labor income tax rates. Since the restriction on sovereign lending takes away the usefulness of building a large negative debt position on the part of the government, setting capital tax rates at confiscatory levels is no longer warranted. More importantly, this observation suggests further consideration of the role of institutional or agency constraints that prevent the government from confiscating capital income for an extended period of time, frictions that limit sovereign lending in practice, and how these constraints help shape optimal fiscal policy more generally.

2. ECONOMIC ENVIRONMENT

Consider an economy populated by infinitely many households whose preferences are given by

$$\mathcal{U} = \sum_{t=0}^{\infty} \beta^t \left[\frac{c_t^{1-\sigma} - 1}{1-\sigma} - \nu \frac{n_t^{1+\frac{1}{\gamma}}}{1+\frac{1}{\gamma}} \right], \sigma > 0, \gamma > 0, \tag{1}$$

where c_t and n_t denote household consumption and labor effort at date t, respectively, and $\beta \in (0, 1)$ is a subjective discount rate.

A single consumption good, y_t, is produced using the technology

$$y_t = k_t^{\alpha} n_t^{1-a}, 0 < \alpha < 1, \tag{2}$$

where k_t denotes the date t stock of private capital. Capital can be accumulated over time and evolves according to

$$k_{t+1} = i_t + (1-\delta) k_t, \tag{3}$$

where $\delta \in (0, 1)$ denotes the depreciation rate and i_t represents household investment. Production can be used for either private or government consumption, or to increase the capital stock,

$$c_t + i_t + g_t = k_t^{\alpha} n_{t,}^{1-\alpha} \tag{4}$$

where $\{g_t\}_{t=0}^{\infty}$ is an exogenously given sequence of public expenditures.

As in Chamley (1986), the government finances its purchases using time-varying linear taxes on labor income and capital income. We denote these tax rates by τ_t^n and τ_t^k, respectively. At each date, the government's budget constraint is given by

$$\tau_t^k r_t k_t + \tau_t^n w_t n_t = g_t, \tag{5}$$

where r_t and w_t are the market rates of return to capital and labor. The left- and right-hand sides of (5) represent sources and uses of government revenue, respectively.

There exists a large number of homogenous small size firms that act competitively. Taking the sequences of prices $\{r_t\}_{t=0}^{\infty}$ and $\{w_t\}_{t=0}^{\infty}$ as given, each firm maximizes profits and solves

$$\max_{k_t,\, n_t} \; k_t^{\alpha} n_t^{1-\alpha} - r_t k_t - w_t n_t. \tag{6}$$

The implied first-order conditions equate prices to their corresponding marginal products, $r_t = \alpha k_t^{\alpha-1} n_t^{1-\alpha} = \alpha \frac{y_t}{k_t}$ and $w_t = (1-\alpha) k_t^{\alpha} n_t^{-\alpha} = (1-\alpha)\frac{y_t}{n_t}$.

At each date, households decide how much to consume and save in the form of private capital investment, as well as how much labor effort to provide. Taking the sequences of government expenditures, $\{g_t\}_{t=0}^{\infty}$, and tax rates, $\{\tau_t^n, \tau_t^k\}_{t=0}^{\infty}$, as given, these households maximize lifetime utility subject to their budget constraint,

$$\max_{c_t,\, n_t,\, k_{t+1}} \sum_{t=0}^{\infty} \beta^t \left[\frac{c_t^{1-\sigma} - 1}{1-\sigma} - \nu \frac{n_t^{1+\frac{1}{\gamma}}}{1+\frac{1}{\gamma}} \right] \tag{P$^{\text{H}}$}$$

subject to

$$\begin{aligned} c_t + k_{t+1} &= (1-\tau_t^k) r_t k_t + (1-\tau_t^n) w_t n_t + (1-\delta) k_t, \qquad (7) \\ k_0 &> 0 \text{ given.} \end{aligned}$$

The first-order necessary conditions implied by problem (7) yield a static equation describing households' optimal labor-leisure choice,

$$\nu n_t^{\frac{1}{\gamma}} = c_t^{-\sigma} (1-\tau_t^n) w_t, \tag{8}$$

as well as a standard Euler equation describing optimal consumption allocations over time,

$$c_t^{-\sigma} = \beta c_{t+1}^{-\sigma} \left[(1-\tau_{t+1}^k) r_{t+1} + 1 - \delta \right]. \tag{9}$$

The constraints (5) and (7), together with the optimality conditions (8) and (9) and the expression for prices given above, describe our economy's decentralized allocations over time.

3. THE RAMSEY PROBLEM

Having described the decentralized behavior of households and firms, we now tackle the problem of choosing policy optimally. Thus, consider a benevolent government that, at date 0, is concerned with choosing a sequence of tax rates that maximize household welfare given the exogenous sequence of government spending. In choosing policy, this government takes as given the behavior of households and firms. We further assume that, at date 0, the government can credibly commit to any sequence of policy actions. The problem faced by this benevolent planner is to maximize (1) subject to the constraints (5) and (7), and households' optimality conditions (8) and (9), where prices are given by marginal products.[5]

We can address the policy problem described at the start of this section by solving the following Lagrangian,

$$\max_{c_t,\, n_t,\, \tau_t^k,\, \tau_t^n,\, k_{t+1}} \mathcal{L} = \sum_{t=0}^{\infty} \beta^t \left[\frac{c_t^{1-\sigma} - 1}{1-\sigma} - \nu \frac{n_t^{1+\frac{1}{\gamma}}}{1+\frac{1}{\gamma}} \right] \qquad (\mathrm{P^R})$$

$$+ \sum_{t=0}^{\infty} \beta^t \mu_{1t} \left[\beta c_{t+1}^{-\sigma} \left[(1-\tau_{t+1}^k) r_{t+1} + 1 - \delta \right] - c_t^{-\sigma} \right]$$

$$+ \sum_{t=0}^{\infty} \beta^t \mu_{2t} \left[\tau_t^k r_t k_t + \tau_t^n w_t n_t - g_t \right]$$

$$+ \sum_{t=0}^{\infty} \beta^t \mu_{3t} \left[(1-\tau_t^k) r_t k_t + (1-\tau_t^n) w_t n_t + (1-\delta) k_t - c_t - k_{t+1} \right]$$

$$+ \sum_{t=0}^{\infty} \beta^t \mu_{4t} \left[c_t^{-\sigma} (1-\tau_t^n) w_t - \nu n_t^{\frac{1}{\gamma}} \right],$$

where the Lagrange multipliers μ_{jt}, $j = 1, ..., 4$, are all nonnegative at the optimum.

The first-order necessary conditions associated with problem (10) that are related to the optimal choices of c_t, n_t, and τ_t^k are as follows:

[5] It is tempting at this point to simply solve a Lagrangian corresponding to the policy problem we have just described. The exact way in which to write this Lagrangian, however, is not immediately clear. To apply Lagrangian methods to this constrained maximization problem, and in particular, to interpret the Lagrange multipliers associated with constraints (5), (7), (8), and (9) as nonnegative, one must first write these constraints as inequalities that define convex sets. See the Appendix for details.

$$\begin{gathered} c_t : c_t^{-\sigma} - \sigma\mu_{1t-1}c_t^{-\sigma-1}[(1-\tau_t^k)r_t + 1 - \delta] + \sigma\mu_{1t}c_t^{-\sigma-1} \\ -\mu_{3t} - \sigma\mu_{4t}c_t^{-\sigma-1}(1-\tau_t^n)w_t = 0, t > 0, \end{gathered} \tag{10}$$

with

$$c_0^{-\sigma} - \sigma\mu_{10}c_0^{-\sigma-1} - \mu_{30} - \sigma\mu_{40}c_0^{-\sigma-1}(1-\tau_0^n)w_0 = 0 \text{ at } t = 0, \tag{11}$$

$$\begin{aligned} n_t \quad : \quad & -\nu n_t^{\frac{1}{\gamma}} + \mu_{2t}\left[\tau_t^k k_t \frac{\partial r_t}{\partial n_t} + \tau_t^n (w_t + n_t \frac{\partial w_t}{\partial n_t})\right] \\ & +\mu_{3t}\left[(1-\tau_t^k)k_t \frac{\partial r_t}{\partial n_t} + (1-\tau_t^n)(w_t + n_t \frac{\partial w_t}{\partial n_t})\right] \\ & +\mu_{4t}\left[c_t^{-\sigma}(1-\tau_t^n)\frac{\partial w_t}{\partial n_t} - \frac{\nu}{\gamma}n_t^{\frac{1-\gamma}{\gamma}}\right] + \mu_{1t-1}c_t^{-\sigma}(1-\tau_t^k)\frac{\partial r_t}{\partial n_t} \\ = \quad & 0, \ t > 0, \end{aligned} \tag{12}$$

with

$$\begin{aligned} & -\nu n_0^{\frac{1}{\gamma}} + \mu_{20}\left[\tau_0^k k_0 \frac{\partial r_0}{\partial n_0} + \tau_0^n (w_0 + n_0 \frac{\partial w_0}{\partial n_0})\right] \\ & +\mu_{30}\left[(1-\tau_0^k)k_0 \frac{\partial r_0}{\partial n_0} + (1-\tau_0^n)(w_0 + n_0 \frac{\partial w_0}{\partial n_0})\right] \\ & +\mu_{40}\left[c_0^{-\sigma}(1-\tau_0^n)\frac{\partial w_0}{\partial n_0} - \frac{\nu}{\gamma}n_0^{\frac{1-\gamma}{\gamma}}\right] \\ = \quad & 0, \text{ at } t = 0, \end{aligned} \tag{13}$$

and

$$\tau_t^k : -\mu_{1t-1}c_t^{-\sigma} + (\mu_{2t} - \mu_{3t})k_t = 0, \ t > 0, \tag{14}$$

with

$$\mu_{20} - \mu_{30} = 0, \text{ at } t = 0. \tag{15}$$

The fact that the above first-order conditions differ at $t = 0$ and $t > 0$ suggests an incentive to take advantage of initial conditions in the first period only, with the promise never to do so in the future. It is exactly in this sense that the optimal policy is not time consistent. Once date 0 has passed, a planner at date $t > 0$ who re-optimizes would want to start with choices for consumption, labor effort, and capital taxes that differ from what was chosen for that date at time 0.

It should be clear that the incentives identified by Chamley (1986) continue to be present in our model economy. Consider that the difference between equations (14) and (15), which governs the optimal choice of τ_t at dates $t = 0$ and $t > 0$, and involves an additional term in (14),

$$-\mu_{1t-1}u_c(c_t) < 0. \tag{16}$$

This term originates from the Euler constraint in problem (10), $\beta u_c(c_t)\,[(1-\tau_t)r_t+1-\delta] = u_c(c_{t-1})$, and corresponds to the reduction in the after-tax real return to investment made at date $t-1$ which is created by an increase in the tax rate at time t. Consequently, in committing to a tax rate in a given period $t > 0$, the government takes into account the implied substitution effect on investment decisions undertaken in the preceding period. Of course, at date $t = 0$, no such distortion exists since history commences on that date with a predetermined capital stock, k_0. In choosing τ_0, therefore, the government is free to ignore its effects on previous investment decisions that can be thought of as "sunk"; and there exists some incentive for the optimal sequence of tax rates to begin with a high tax in period 0 relative to all other dates.

A central insight in Kydland and Prescott (1980) is that despite the time inconsistency problem we have just mentioned, it is actually possible to collapse equations (10) through (15) into a set of stationary difference equations $\forall t \geq 0$. This requires interpreting the lagged Lagrange multiplier μ_{1t-1} as a predetermined variable with initial condition $\mu_{1t-1} = 0$ at $t = 0$.

The remaining first-order conditions associated with problem (10) determining the optimal choice of labor income taxes and private capital are, respectively,

$$\tau^n_t : (\mu_{2t} - \mu_{3t})n_t - \mu_{4t}c_t^{-\sigma} = 0,\ t \geq 0, \tag{17}$$

and

$$\begin{aligned} k_{t+1} : \mu_{1t}\beta c_{t+1}^{-\sigma}(1-\tau^k_{t+1})\frac{\partial r_{t+1}}{\partial k_{t+1}} + \beta\mu_{2t+1} \\ \left[\tau^k_{t+1}(r_{t+1} + k_{t+1}\frac{\partial r_{t+1}}{\partial k_{t+1}}) + \tau^n_{t+1}n_{t+1}\frac{\partial w_{t+1}}{\partial k_{t+1}}\right] - \mu_{3t} + \beta\mu_{3t+1} \\ \left[(1-\tau^k_{t+1})[r_{t+1} + k_{t+1}\frac{\partial r_{t+1}}{\partial k_{t+1}}] + (1-\tau^n_{t+1})n_{t+1}\frac{\partial w_{t+1}}{\partial k_{t+1}} + 1 - \delta\right] \\ +\beta\mu_{4t+1}c_{t+1}^{-\sigma}(1-\tau^n_{t+1})\frac{\partial w_{t+1}}{\partial k_{t+1}} = 0,\ t \geq 0. \end{aligned} \tag{18}$$

4. THE STATIONARY RAMSEY EQUILIBRIUM

With the optimality conditions (10) through (18) in hand, we first turn to long-run properties of optimal taxes and revisit the notion that it is not efficient to tax capital in the long run. We do so, however, without any reference to the primal approach that is standard in the literature, but rely instead on the simple first-order conditions we have just derived. To this end, we define the long-run equilibrium of the Ramsey problem as follows:

Definition: *A stationary Ramsey equilibrium is a ninetuple* $(c, n, k, \tau^n, \tau^k, \mu_1, \mu_2, \mu_3, \mu_4)$ *that solves the government budget constraint (5), households' budget constraint (7), the optimality condition for labor effort (8), and the Euler equation (9), as well as the first-order conditions associated with problem (10), equations (10), (12), (14), (17), and (18), all without time subscripts.*

It is straightforward to show that in a stationary Ramsey equilibrium, equations (14), (17), and (18) imply that

$$\tau^k \mu_2 \beta r - \mu_3 \left[1 - \beta \left((1 - \tau^k) r + 1 - \delta\right)\right] = 0. \qquad (18)$$

From the Euler equation in the stationary equilibrium, it follows that $1 - \beta[\alpha(1 - \tau^k)\frac{y}{k} + 1 - \delta] = 0$. Hence equation (18) above reduces to

$$\tau^k \mu_2 \beta r = 0. \qquad (19)$$

Now, we have that either $\tau^k > 0$ or $\tau^k = 0$. Suppose first that $\tau^k > 0$. Then, it must be the case that $\mu_2 = 0$. From equation (14), this would mean that

$$\mu_1 = -\mu_3 k c^{\sigma},$$

which implies that $\mu_1 = \mu_3 = 0$ since μ_1 and μ_3 are both nonnegative. However, in that case, all Lagrange multipliers are zero in the steady state and $c^{-\sigma} = 0$ from equation (10), which cannot be a solution because household utility would be unbounded. Hence, $\tau^k > 0$ cannot be a solution, and therefore, $\tau^k = 0$. As in Chamley (1986), it is optimal not to tax capital in the long run. From the budget constraint, this implies that the steady state tax on labor is essentially determined by the extent of government expenditures. For instance, if government spending was a constant fraction, ϕ of output in the long run, we would have the optimal tax on labor income in the long run to be simply $\tau^n = \frac{\phi}{1-\alpha}$.

The notion that it is optimal to set capital income tax rates to zero in the long run is independent of whether government lending takes place. This is relatively easy to see within our framework when no upper bound is imposed on the capital income tax rate. In that case, the government budget constraint (5) reads as

$$\tau_t^k r_t k_t + \tau_t^n w_t n_t + b_{t+1} = g_t + (1 + r_t^b) b_t, \qquad (20)$$

where b_t denotes one-period government bonds that are perfectly substitutable with capital, and r_t^b is the return on bonds from period $t-1$ to t. Government lending takes place when $b_t < 0$. Moreover, the household budget constraint becomes

$$c_t + k_{t+1} + b_{t+1} = (1 - \tau_t^k) r_t k_t + (1 - \tau_t^n) w_t n_t + (1 - \delta) k_t + (1 + r_t^b) b_t. \quad (21)$$

Substituting these modified constraints in problem (10), the planner now also has to decide how much sovereign lending will take place. A simple arbitrage equation (obtained from the modified household problem) dictates that in the decentralized equilibrium, $1 + r_t^b = (1 - \tau_{t+1}^k)r_{t+1} + 1 - \delta$. Hence, the first-order condition associated with the optimal choice of b_{t+1} in the Ramsey problem is

$$(\mu_{2t} - \mu_{3t}) - \beta[(1 - \tau_{t+1}^k)r_{t+1} + 1 - \delta](\mu_{2t+1} - \mu_{3t+1}) = 0 \ \forall t \geq 0. \quad (22)$$

It is now easy for us to show that capital income taxes are zero in the long run. In fact, with no upper bound imposed on the capital income rate, $\tau_t^k = 0 \ \forall t > 0$. To see this, observe that equations (15) and (22) imply that $\mu_{2t} - \mu_{3t} = 0$ $\forall t \geq 0$. It follows from (14) that $\mu_{1t-1} = 0 \ \forall t \geq 0$ and from (17) that $\mu_{4t} = 0$ $\forall t \geq 0$. Substituting these results into equation (18) gives

$$c_t^{-\sigma} = \beta c_{t+1}^{-\sigma} \left[r_{t+1} + 1 - \delta \right] \ \forall t \geq 0,$$

which is simply the household's Euler equation (9) when $\tau_t^k = 0 \ \forall t > 0$. When an upper bound is imposed on the capital income tax rate, $\tau_t^k \leq 1$ $\forall t \geq 0$, it is still the case that $\tau_t^k = 0 \ \forall t > 0$ when preferences are separable in consumption and leisure, and that $\lim_{t \to \infty} \tau_t^k = 0$, otherwise. Proof of the latter results is more difficult to see using our Lagrangian formulation, but is nicely presented in Erosa and Gervais (2001).

5. TRANSITIONS TO THE STEADY STATE

Even in the absence of an instrument that allows government lending to households, we saw in the previous section that the optimal fiscal policy with commitment prescribes zero capital taxes in the long run. In the short and medium run, however, capital income tax rates are not as extreme as predicted in a model with a debt instrument. Compared to the environment studied by Chari, Christiano, and Kehoe (1994) for instance, where capital income tax rates are set at their upper bound up to some date $\bar{t}$ and are zero thereafter, capital income tax rates in our framework approach confiscatory rates only in the initial period and then decline monotonically over time. Labor income tax rates are also moderate at every point along the transition.

To illustrate these points, we carry out a numerical simulation of our economy when fiscal policy is determined optimally. The parameters we use are standard and selected along the lines of other studies in quantitative general equilibrium theory. A time period represents a quarter and we assume a 6.5 percent annual real interest rate, $\beta = 0.984$, and a 10 percent capital depreciation rate, $\delta = 0.025$. We set the intertemporal elasticity of substitution, $1/\sigma$, to 1/2 and the Frisch elasticity of labor supply, γ, to 1.25. The share of private capital in output in the United States is approximately 33 percent so we assign a value of $\alpha = 1/3$. Finally, we fix the share of government expenditures in output at 0.20.

To compute the transitional dynamics associated with optimal capital and labor income tax rates, we replace the optimality conditions in Section 3 with log-linear approximations around the stationary Ramsey equilibrium. The solution paths for the state and co-state variables are then computed using techniques described in Blanchard and Kahn (1980) or in King, Plosser, Rebelo (1988). The resulting system of linearized equations possesses a continuum of solutions, but only one of these is consistent with the transversality condition associated with the household problem.

Figure 1 depicts transitions to the stationary Ramsey equilibrium when the initial capital stock is set at its long-run level. In other words, Figure 1 shows transitions to the steady state when restarting the problem. In Panel A, we can see that capital income tax rates start near confiscatory rates in the initial period but quickly fall within 10 quarters to a more moderate range, at less than 35 percent.[6] Thus, the notion that capital income tax rates are optimally higher in the initial periods remain, but these rates are within a moderate range for the greater part of the transition. More specifically, in contrast to Chari, Christiano, and Kehoe (1994), capital income tax rates are never set at either 100 percent or zero at any point during the transition.[7] Interestingly, the optimal fiscal policy suggests subsidizing labor income in the first few periods, after which labor income taxes monotonically rise to their steady state. Because labor income represents $2/3$ of total output in our calibrated economy, and because government expenditures account for 20 percent of output, the labor income tax rate approaches 30 percent asymptotically as capital income tax rates approach zero.

The initial subsidization of labor income generates an increase in labor input, shown in Figure 1, Panel D, along the transition to the steady state. As a result, household consumption is everywhere above its long-run level on its way to the steady state. Moreover, the optimal fiscal policy is such that households are able to front-load consumption.

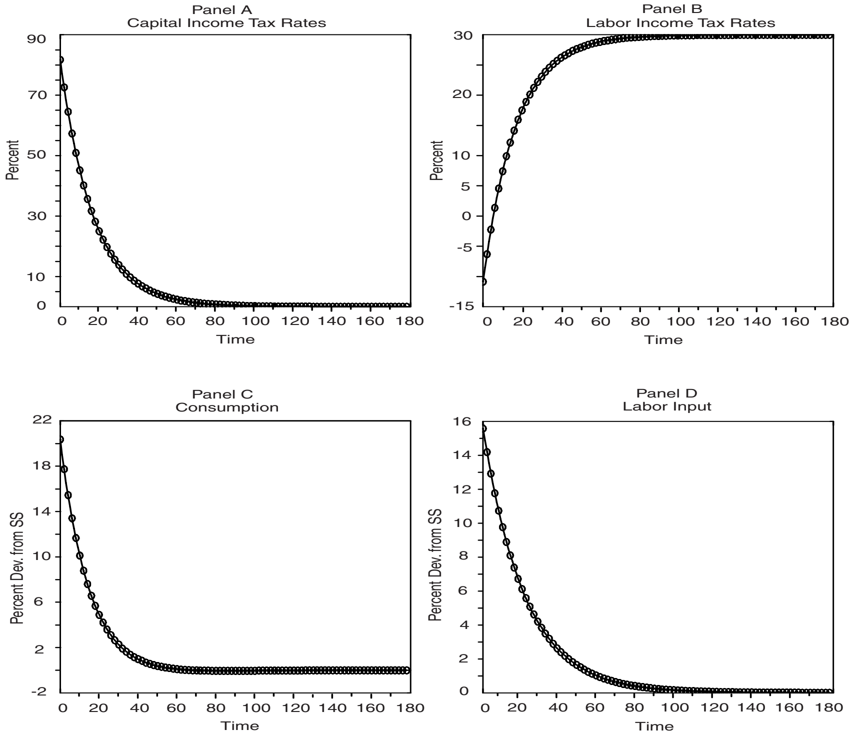
6. CONCLUSION

In this article, we highlighted that environments in which ex ante government lending is limited are more apt to generate optimal distortional taxes at all dates that do not share the stark character of those typically presented in the literature. Absent an instrument that allows households to borrow from the government, the government cannot carry the proceeds from a large initial tax to future periods, and the single capital levy prescribed in Chamley (1986) cannot be implemented.

[6] The captial stock is fixed in period 0. It then decreases slowly while capital income tax rates are relatively high and converges back to its long-run level.

[7] Chari, Christiano, and Kehoe (1994) present an actual numerical solution to the problem without linearizing. The linearization in our case involves an approximation, but the fact that optimal taxes do not involve corner solutions in our framework does not depend on the approximation.

Figure 1 Transitions to the Steady State, $k_0 k_{ss}$



For an economy whose capital stock is initially below its long-run level, we have shown that capital income tax rates are never set at either 100 percent or zero at any point during the transition to the long-run equilibrium. Furthermore, our analysis has highlighted that labor income is subsidized in the first few periods. This feature of optimal fiscal policy gave rise to increased labor supply and allowed household consumption to be everywhere above its long-run equilibrium along its transition path. As in Chamley (1986), however, even without a debt instrument, our analysis continued to prescribe zero-capital income tax rates in the long run.

We interpret our findings to suggest that a better understanding of institutional or agency constraints that prevent the government from confiscating capital income for an extended period of time, as well as commitment problems associated with household borrowing, may be central in explaining the character of observed fiscal policies.

APPENDIX

Marcet and Marimon (1999) point out that equalities associated with feasibility constraints can generally (and in this case) be replaced with weak inequalities, with an appeal to nonsatiated preferences thereby guaranteeing that the feasibility constraints rewritten as such will also be satisfied with equality. However, an analogous argument for the equality constraints (8) and (9) is less obvious. Because of the equality signs in (8) and (9), the set of allocations satisfying these equations is not convex.

Consider the Euler equation (9). Marcet and Marimon (1999) show that it is possible to rewrite this constraint as a weak inequality in such a way that, in the optimum of the new problem, this weak inequality is satisfied as a strict equality. One can be sure, therefore, that the optimum subject to the weak inequality constraint is the same as that subject to the strict equality (9), and that one is actually solving the problem of interest.

We now provide a brief description of the arguments presented in Marcet and Marimon (1999) but refer the reader to the paper for the formal proofs. The question is whether to write the inequality associated with (9) as $\leq$ or $\geq$. Consider the case $\geq$ first, in which $c_t^{-\sigma} \geq \beta c_{t+1}^{-\sigma}\left[(1-\tau_{t+1}^k)r_{t+1}+1-\delta\right]$. The authors show that writing the inequality constraint in this way actually makes the first-best allocation feasible, so that the solution would be the unconstrained optimum, which is not the same as the Ramsey equilibrium. Hence, this option does not yield a solution equivalent to the solution under equation (9).

Next, consider the case where the inequality constraint is written as $\leq$ so that $c_t^{-\sigma} \leq \beta c_{t+1}^{-\sigma}\left[(1-\tau_{t+1}^k)r_{t+1}+1-\delta\right]$. Writing the inequality in this way reproduces the household's first-order condition if the household faced the constraint $k_t \leq \overline{k}_t$, where $\overline{k}_t$ is an upper bound imposed on households' capital position. In other words, the modified Euler equation corresponds to a setting where the policy instruments available to the planner now include the ability to set an upper limit on capital accumulation, $\overline{k}_t$. In that case, Marcet and Marimon (1999) then show that the planner will actually choose allocations where the constraint $k_t \leq \overline{k}_t$ does not bind, since the equilibrium with distortional taxes is associated with too little capital relative to the full optimum. This implies that the government will act so that $c_t^{-\sigma} \leq \beta c_{t+1}^{-\sigma}\left[(1-\tau_{t+1}^k)r_{t+1}+1-\delta\right]$ is satisfied with equality, and the optimum is then the same as the Ramsey equilibrium. Similar arguments can be made regarding constraint (8).

REFERENCES

Aiyagari, S. 1995. “Optimal Capital Income Taxation with Incomplete Markets, Borrowing Constraints, and Constant Discounting.” *Journal of Political Economy* 103 (6): 1158–75.

Atkeson, A., V. V. Chari, and P. J. Kehoe. 1999. “Taxing Capital Income: A Bad Idea.” Federal Reserve Bank of Minneapolis *Quarterly Review* 23 (3): 3–17.

Atkinson, A., and A. Sandmo. 1980. “Welfare Implications of the Taxation of Savings.” *The Economic Journal*. 90 (359): 529–49.

Blanchard, O., and C. M. Kahn. 1980. “The Solution of Linear Difference Models under Rational Expectations.” *Econometrica* 48 (5): 1305–12.

Chamley, C. 1986. “Optimal Taxation of Capital Income in General Equilibrium with Infinite Lives.” *Econometrica* 54 (3): 607–22.

Chari, V. V., L. Christiano, and P. J. Kehoe. 1994. “Optimal Fiscal Policy in a Business Cycle Model.” *Journal of Political Economy* 102 (4): 617–52.

Correia, I. H. 1996. “Should Capital Income be Taxed in the Steady State?” *Journal of Public Economics* 60 (1): 147–51.

Domeij, D., and J. Heathcote. 2004. “Factor Taxation with Heterogenous Agents.” *International Economic Review* 45 (2): 523–54.

Erosa, A., and M. Gervais. 2001. “Optimal Taxation in Infinitely-Lived Agent and Overlapping Generations Models: A Review.” Federal Reserve Bank of Richmond *Economic Quarterly* (87) 2: 23–44.

Erosa, A., and M. Gervais. 2002. “Optimal Taxation in Life-Cycle Economies.” *Journal of Economic Theory* 105 (2): 338–69.

Garriga, C. 1999. “Optimal Fiscal Policy in Overlapping Generations Model.” Mimeo, Universitat de Barcelona.

Jones, L. E., R. E. Manuelli, and P. Rossi. 1997. “On the Optimal Taxation of Capital Income.” *Journal of Economic Theory* 73 (1): 93–117.

Judd, K. L. 1985. “Redistributive Taxation in a Simple Perfect Foresight Model.” *Journal of Public Economics* 28: 59–83.

King, R. G., C. Plosser, and S. Rebelo. 1988. “Production, Growth and Business Cycles: The Basic Neoclassical Model.” *Journal Monetary Economics* 21 (2/3): 195–232.

Kydland, F. E., and E. C. Prescott. 1980. “Dynamic Optimal Taxation, Rational Expectations and Optimal Control.” *Journal of Economic Dynamics and Control* 2 (1): 79–91.

Ljungqvist, L., and T. J. Sargent. 2000. *Recursive Macroeconomic Theory*. Cambridge, MA: MIT Press.

Marcet, A., and R. Marimon. 1999. “Recursive Contracts.” Mimeo, Universitat Pompeu Fabra.

McGrattan, E., and L. Ohanian. 1999. “Does Neoclassical Theory Account for the Effects of Big Fiscal Shocks? Evidence from World War II.” Federal Reserve Bank of Minneapolis Research Department Staff Report 315.

Xie, D. 1997. “On Time Inconsistency: A Technical Issue in Stackelberg Differential Games.” *Journal of Economic Theory* 76 (2): 412–30.

Not Your Father's Credit Union

John R. Walter

Proponents of the credit union industry have viewed credit unions as especially well-suited to making small-value, uncollateralized consumer loans since the time the first United States credit union was formed in 1909. At that time, other financial institutions such as banks and savings and loans were focused on alternative types of lending, specifically lending to businesses as well as to consumers with collateral. For consumers who lacked real estate or other simple-to-value collateral—in other words for *unsecured borrowers*—borrowing options were limited, at least for loans from depositories.

The high fixed costs of extending a small-value loan, and the significant risk of making an unsecured consumer loan to a borrower about whom the lender had little information on creditworthiness, meant that a lofty interest rate was necessary in order for the lender to cover costs. Such rates often exceeded usury ceilings, maximum interest rates set by state laws. Still, many borrowers were willing to pay high rates, even if it meant illegal borrowing. A 1911 estimate indicated that in larger towns and cities, one in five workers borrowed from illegal lenders (Calder 1999, 118).

Credit unions developed in this environment, allowing individuals to borrow at fairly low interest rates (Whitney 1922, 40–54). When consumers could band together into groups based upon a *common bond*—i.e., a shared characteristic such as working for the same company—they could substitute their knowledge of one another's creditworthiness for collateral. Credit unions offered creditworthy borrowers the opportunity to differentiate themselves from less creditworthy individuals in an era before the advent of nationwide credit bureaus. Credit union members were differentiated from unknown

The author benefited greatly from comments from Kevin Bryan, Huberto Ennis, Robert Hetzel, and Ned Prescott. Able research assistance was provided by Kevin Bryan and Nashat Moin. The views expressed herein are not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

borrowers because of in-depth information among the group of members about the character and economic prospects of one another. In 1932, the average size of a U.S. credit union was only 187 members, so intimate knowledge among the membership was feasible in the early decades of the industry (U.S. Department of Health, Education, and Welfare 1965, 1).

With this knowledge in hand, the credit union loan committee could make a low-risk and, therefore, low-interest loan to a credit union member. Borrowing members received low interest rate loans, and the owners of the credit union—also members—were repaid with interest, which then was returned to members in the form of interest (also called dividends) on deposits (also called shares) in the credit union. Since members knew one another, and failure to repay a loan harmed fellow members and damaged the defaulting member's reputation in the group, there was social pressure to repay, thus reducing the likelihood of default, and consequently providing another explanation for low interest rates on loans from credit unions.

The advent of nationwide credit bureaus, which provided in-depth information on the creditworthiness of most consumers, and the growth of credit card lending as well as loan products that allowed borrowers to inexpensively employ real estate collateral for small loans—for example, home equity lines—the necessity of credit unions as producers of common bond-driven information on creditworthiness declined. In order to continue to prosper and grow, which many did successfully, credit unions were required to evolve from their original structure. To do so, the industry shifted toward mortgage and credit card lending and, to a smaller degree, toward business lending. The industry was also granted liberalized membership rules by its regulator and later by statute.

As the credit union industry evolved and many credit unions became more bank-like in their product offerings, the banking industry expressed growing concern over a perceived credit union advantage. Specifically, credit unions are exempt from federal income taxes, while banks pay such taxes. Credit union proponents have responded that the tax exemption remains appropriate because credit unions continue to be nonprofit, member-owned financial institutions, and because they face restrictions not placed on banks. The significance of the controversy was illustrated in November 2005 when congressional hearings were held to examine the credit union tax-exemption and its justifications in the face of changes to the credit union industry (U.S. House of Representatives 2005).

How do banks and credit unions differ, and what is the history behind these differences? How have these differences changed over time and why? Observers of the credit union tax exemption controversy might ask such questions. This article provides a history of the evolution of credit unions and the market forces behind the evolution, without taking a stand on the merits of either side of the tax debate.

1. EARLY HISTORY OF THE CREDIT UNION INDUSTRY

The first credit union in the United States, St. Mary's Cooperative Credit Association, was chartered in Manchester, New Hampshire, in 1909 (National Credit Union Administration 2006a). This and other similar credit unions developed to fill a niche in the loan market that was largely unfilled by existing banks and savings institutions. The niche that they filled was that of unsecured, small dollar, consumer loans—for example, for the payment of a medical bill or to purchase a home appliance.

Lending institutions along the lines of United States credit unions first developed in Europe. Cooperative banks or people's banks appeared in Germany in the 1850s (Moody and Fite 1984, 1–10). These spread to a number of other countries, and then to Canada, at the turn of the 20th century, before coming to the United States.

Initial Leaders of the Credit Union Industry

A number of individuals are credited with the introduction and spread of credit unions in the United States. Three of the most important are Alphonse Desjardins, Pierre Jay, and Edward A. Filene.

Alphonse Desjardins, a Canadian journalist and politician, organized one of the first cooperative banks in Canada, *La Caisse Populaire de Levis*. *Caisse Populaire* loosely translates to people's credit society or people's bank (Moody and Fite 1984, 12–15). Caisse, located in the city of Levis, Quebec, opened for business in January 1901, and was modeled after European cooperative banks which Desjardins had been studying for years. Caisses grew rapidly in Canada, thanks to Desjardin's advocacy, numbering 150 by 1914 (Moody and Fite 1984, 17).

In late 1908, Desjardins helped organize the first credit union in the United States, at the request of the priest of St. Mary's Parish in Manchester. He also consulted with Pierre Jay and Edward Filene as they pushed forward the spread of credit unions in the United States following the formation of St. Mary's.[1]

Pierre Jay, Banking Commissioner of Massachusetts, became familiar with the cooperative banking concept soon after 1900 (Moody and Fite 1984, 22–23). As banking commissioner, he recognized a lack of lenders willing to make small-value loans to consumers. His interest led him to contact Alphonse Desjardins. In July 1908, Jay traveled to Ottawa to meet with Desjardins and learn first-hand about les caisses populaires in Canada (Flannery 1974, 14). In April 1909, Jay persuaded the Massachusetts legislature to pass legislation that allowed cooperative banks, named credit unions, to be chartered in that

[1] While St. Mary's was created in late 1908, it did not receive a charter to operate from the legislature of New Hampshire until early 1909 (Whitney 1922, 16).

state. Then, as now, most states prohibited the gathering of deposits by firms lacking a charter that specifically grants it this power.

Edward Filene was a Boston merchant (of Filene's Basement fame) who traveled widely and witnessed successful agricultural cooperative lenders in India. These lenders gathered small deposits from Indian farmers, members of the lending cooperative, and made small-value loans to their members (Moody and Fite 1984, 21). Because of his business experience in the United States, Edward Filene, like Jay, perceived a dearth of small-value lenders and wished to bring the cooperative bank to the United States.

While Desjardins and Jay were responsible for the first steps in establishing credit unions, Filene created momentum to produce growth in the industry over the coming several decades. His efforts included lobbying legislators to pass credit union legislation, speaking frequently on the benefits of cooperative lending, and funding organizations to promote the expansion of credit unions.

Credit union growth was fairly slow between 1909, when the first U.S. credit union opened, and 1920 (Moody and Fite 1984, 32–72). Growth of credit unions began to pick up in the mid-1920s, as an increasing number of states passed credit union laws (Moody and Fite 1984, 73). By 1925, credit union laws had passed in 26 states. In 1925, there were 176 credit unions; this count had grown to 868 in 1929 (U.S. Bureau of the Census 1975, 1,049) and 2,500 by 1934, when the Federal Credit Union Act created federal credit union charters (U.S. Treasury Department 1997, 16). According to Moody and Fite (1984, 55–108), much of the legislation that made this growth possible was the result of Edward Filene's lobbying efforts or the result of efforts financed by him.

Beyond Credit Unions

Credit unions were not the only response to limited borrowing opportunities for consumers. The Russell Sage Foundation was formed, based on funding provided by Russell Sage's widow (Sage was a railroad baron who died in 1906). The foundation soon began advocating for the liberalization of usury laws for small-value loans. Low usury ceilings in a number of states meant that small-value consumer loans could not be made profitably. Out of this advocacy grew a movement to encourage all states to adopt uniform small-value loan regulations that would provide for fairly high usury ceilings on small consumer loans, while at the same time implementing consumer protection rules. Approximately two-thirds of the states passed the uniform laws by the time the Foundation ended its push for such laws in the 1940s (Carruthers, Guinnane, and Lee 2005).

In addition to the effort of the Russell Sage Foundation, Morris Plan banks were formed around the country based on a bank founded in Norfolk, Virginia. The original Morris Plan bank was chartered by Virginia in 1910 at the request of Arthur J. Morris. It gathered deposits from and made small-value unsecured installment loans to consumers. Morris lowered losses and, therefore, was able to charge fairly low interest rates, by requiring several cosigners for all consumer loans. Further, he sidestepped usury ceilings to some degree by charging loan fees and deducting interest payments from the loan when it was initially made (Giles 1951, 81–86). Ultimately, Morris Plan banks were called industrial banks, which remain today in the form of industrial loan corporations, one of the few bank types that can be owned by commercial firms (Walter 2006).[2]

2. COMMON BOND LENDING

Given existing lenders' focus on other types of lending, the early 20th century consumer lending market was ripe for the development of financial institutions that could act as low-cost providers of small-value, unsecured consumer loans. Commercial banks, mutual savings banks, and savings and loans focused their lending efforts in other directions.

Pierre Jay noted the absence of providers of small-value, *unsecured* loans in Massachusetts, and foresaw the operating characteristics of credit unions:

> What are known in Massachusetts as "cooperative banks," are excellent examples of the success of cooperation between savers and borrowers in an important but limited field. Their members are almost entirely salary and wage earners, who have joined together for systematic saving and for the owning of their homes. But inasmuch as all of their loans are secured, either on real estate or on the cash value of their shares, there is no need for selecting the members, no scrutiny is made of the object of loans made on the cash value of shares, and but little interest is taken by the members in the management of the institution. Furthermore they are of no service to those who have neither real estate nor shares as security, many of whom undoubtedly have legitimate need for loans. (quoted in Moody and Fite 1984, 24–25)

[2] The early 20th century saw not only the development of the credit union industry, but also the growth of other cooperative business organizations, i.e., organizations in which customers are also owners or members. For example, agricultural cooperatives, which began to spread in the United States in the 1870s, grew rapidly between 1890 and 1920. The number of agricultural cooperatives peaked, at 14,000 in the 1920s (Frederick 1997). Consumer cooperatives gained popularity in the 1920s, and cooperatives also formed to provide rural electric and telephone service, encouraged by the Rural Electrification Administration, which passed in 1935 (Frederick 1997).

Relationships As a Substitute for Collateral and Credit Ratings

Credit unions were small, with, as noted earlier, on average only 187 members in 1932. Further, they were formed based on a common bond, where members knew one another. In other words, members of a common bond likely knew, at least better than other potential lenders, the economic prospects of the borrowers who were fellow members.

Alphonse Desjardins noted the benefit of common bond lending when discussing exceedingly low loan losses at La Caisse Populaire. He claimed that losses were low because, "The main security is the fact that the association is working within a small area and that everybody knows each other" (Moody and Fite 1984, 16). Pierre Jay had an eye toward the members of a common bond, monitoring one another while discussing the lack of such monitoring when loans are secured by real estate or deposits (shares), when he said, "there is no need for selecting members, no scrutiny is made of the object of loans."

At the time, nationwide consumer credit rating agencies did not exist, though there were credit rating agencies in some of the larger cities. Because of the lack of these agencies, it was difficult for lenders to differentiate between creditworthy and non-creditworthy consumers. The common bond acted as a substitute, allowing its members to differentiate good credit risks from poor credit risks.

Monitoring Within the Common Bond

Not only were members of small credit unions, or the credit committee, likely to know something of the financial circumstances of their fellow members, they had incentive to deny loans to poor credit risks and to monitor borrowers. Federal deposit insurance coverage was not granted to credit unions until 1970. A few states had state insurance schemes for credit unions, but in general, depositors were subject to losses when borrowers failed to repay (Flannery 1974, 26). Depositors faced the danger that if losses were large enough they might lose principal on their deposits.

The view that common bond lending contained this monitoring advantage was expressed by Alphonse Desjardins when he noted that "a second security [in addition to members' knowledge of one another's creditworthiness] is that everybody is interested by being a shareholder" (Moody and Fite 1984, 16).

In effect, common bond lending used social pressure from associates instead of the threat of confiscation of collateral, to ensure loan repayment. Since these associates suffered loss due to a member's default, the social pressure to repay was likely significant. The borrower knows that if he does not repay, he is costing his associates lost interest earnings and perhaps lost principal.

A 1922 Department of Labor study of early credit unions reported that loans were typically made without collateral, but with careful scrutiny of the character of the borrower by the credit committee—typically the loan decision-making body in credit unions (Whitney 1922, 33). Small loans were made on the signature of the borrower only, while loans over $50 required the endorsement of two other credit union members. According to this 1922 study, typical interest rates charged by credit unions ranged from 6 to 12 percent per annum (Whitney 1922, 40–54).[3]

Still, the incentive of any individual member to keep an eye on another borrower is somewhat muted. The cost of one member's efforts to monitor another member's financial condition is borne only by that member, but the benefit if loans are properly denied, or remedial action is taken, is shared by all members in the form of higher returns on their credit union deposits.[4]

Reputational Incentive to Repay

While members' incentive to monitor one another's loans in order to protect credit union earnings may be muted, there is another reason to expect borrowers within a common bond group to repay. Members of a credit union are likely to have many valuable interactions with one another, outside of interactions through the credit union. Failure to repay a loan is likely to damage the reputation of the defaulting borrower and undercut these interactions.[5]

For instance, members of a credit union who share the common bond of working for the same company may worry that defaulting on a credit union loan will decrease their opportunities for advancement. Their co-workers, including supervisors, would be likely to learn of the default, undercutting the defaulting worker's reputation.

Further, a credit union member who shares a close common bond with fellow members might be concerned that his default could damage his reputation and limit social interactions with members. For example, the credit union borrower might believe that members will view a default as a sign of dishonesty and, therefore, be less willing to include him in confidences.

[3] The author of the 1922 study notes that some credit union loans were made on collateral such as deposits in the credit union, stocks, or real estate collateral (Whitney 1922, 40).

[4] Group lending, whereby loans are made to members of a group and all members of the group are jointly liable for each other's debts, can provide strong incentive to members to monitor and discipline one another to ensure repayment. Group lending, which, like common bond lending, uses social connections to encourage repayment, has received wide attention as a means of inexpensively lending to individuals for whom little formal information on creditworthiness is available. Group lending and a number of interesting examples as practiced in developing countries are discussed and explored in Prescott (1997).

[5] Besley and Coate (1995) develop a model which is used to explore the ways "social connectedness" can encourage debt repayment.

Members of a common bond will be aware of the potential damage to one's reputation. As a result, they will be less likely to default because of the potential damage to other interactions with fellow members of a common bond.

3. FACTORS DIFFERENTIATING CREDIT UNIONS FROM OTHER DEPOSITORIES

Credit unions share many features common to all depository institutions: they gather savings and transaction deposits, make various types of loans, and employ branch networks, ATMs, telephone service departments, and Web sites to provide services to their customers. They are insured by an agency of the federal government, receive a charter to operate, and are closely regulated. Beyond these similarities, credit unions are distinct. They differ from most depositories in that they are mutually owned, rather than stockholder-owned. Another distinction is their greater concentration on unsecured lending, though the growth of credit card lenders has meant that credit unions are less distinct here. While other depositories take deposits from and make loans to anyone meeting creditworthiness requirements, credit unions face restrictions limiting whom they may accept as customers. Last, credit unions have the unique advantage of being exempt from federal taxes.

Mutual Ownership

State credit union laws and the Federal Credit Union Act require that credit unions be cooperative or mutually owned organizations. This implies that a credit union's members, meaning its depositors, are also its owners. As a result, depositors have the right to vote for the members of the board of directors of the credit union, who, in turn, hire managers who run the day-to-day operations of the credit union.

Credit unions do not raise equity by selling shares in the equity markets, as do nonmutual corporations. Instead, credit unions raise equity by retaining earnings. Credit union depositors, as owners, are due not only the repayment of their deposits and any interest owed to them, but also the credit union's equity, meaning funds that might remain after all liabilities are repaid if the credit union is dissolved. Consequently, credit union depositors are the residual claimants to the credit union's assets, like corporate shareholders, explaining why credit union deposits are often called “shares.” Depositors in most other depository institutions are due only the repayment of deposits plus interest.[6]

Uncollateralized Consumer Lending

Beyond their ownership, credit unions were originally set apart from other depositories by their emphasis on small-value, nonmortgage loans to individuals and households, meaning uncollateralized loans for household expenses and the purchase of consumer durables. For example, in the early 1920s, the average size loan made by credit unions in Massachusetts and New York—two states in which numerous credit unions operated—was $272 ($3,092 in current dollar terms) in Massachusetts and $169 ($1,720) in New York (Whitney 1922, 55).

With the exception of some securities investments, credit union assets were devoted almost completely to loans to individuals. The earliest data on the types of loans credit unions made comes from a 1948 survey of federally chartered credit unions conducted by the Bureau of Federal Credit Unions. The survey indicated that the largest categories were loans for the purchase of automobiles to consolidate loans for current living expenses and medical expenses (U.S. Bureau of Federal Credit Unions 1948, 4).

In contrast to credit unions, savings and loans and savings banks focused on mortgage lending and commercial banks on business lending. In the case of mutual savings banks, as of 1910, mortgages and investments in securities accounted for 89 percent of assets (U.S. Bureau of the Census 1975, 1,046). Figures on types of lending by savings and loans is reported starting with 1929, when mortgages accounted for 75 percent of all assets (U.S. Bureau of the Census 1975, 1,047). Data on types of loans and assets for commercial banks are reported beginning with 1939. During that year, 91 percent of bank assets were accounted for by loans to businesses, securities investments, cash holdings, and mortgage lending (Board of Governors 1959, 34–35). In each of these cases, uncollateralized loans to individuals were no more than 25 percent of assets.

Until 1977, federal credit unions were largely prohibited from making real estate loans, though some states allowed such lending by state-chartered credit

[6] Another group of depository institutions, mutual savings banks, which traditionally focused on home mortgage lending, are like credit unions, also owned by depositors.

The advantages and disadvantages of mutual ownership relative to stockholder ownership have been studied widely, especially in the insurance industry, where both types of ownership are prevalent. Generally these studies find that the choice is driven by conflicts of interest between owners and managers on one hand, and owners and customers on the other. Examples of such studies are Mayers and Smith (2002), Mayers and Smith (1986), and Fama and Jensen (1983).

unions. The shift into mortgage lending after 1977 contributed to growth in loan size at credit unions, such that in March 2006 the average credit union loan was $10,903 (National Credit Union Administration 2006b).

During the 1940s, credit unions had little opportunity to compete for mortgage loans (as discussed later), but by 2006, such loans accounted for 32 percent of all assets. Nevertheless, credit unions continue to exhibit considerably greater emphasis on nonmortgage loans to individuals, which as of March 2006, accounted for 31 percent of all of their assets, compared to 8 percent for commercial banks, savings banks, and savings and loans (Federal Deposit Insurance Corporation 2006, 5).

Restricted Customer Base

An additional important factor differentiating credit unions from other depositories is that they have traditionally been required by state and federal law to choose their members from one group. Specifically, a credit union must focus on a group sharing one of three forms of affinity: (1) a common bond of association such as a fraternal or religious organization, (2) an occupation, or (3) residence in a single community, neighborhood, or rural district. In other words, unlike banks, savings and loans, and savings banks, credit unions face restrictions that limit with whom they may do business.

The apparent goal of this limitation is to protect the health of credit unions, as members are more likely to repay loans given that failure to repay harms fellow members of their own group (Robbins 2005, 3). While this repayment benefit may flow from the common bond requirement, there is also a serious disadvantage. Credit unions bound to accept members only from one employer or community are likely to suffer from a severe lack of diversification. Limited diversification may have contributed to high failure rates among small credit unions in the late 1970s and 1980s (Wilcox 2005, 2–3).

Tax Exemption

Tax exemption is the final, and most controversial, difference separating credit unions from other depositories. Credit unions are free from federal income taxes. All of a credit union's net earnings, meaning earnings after paying expenses, will be paid out to its owners (depositors) as interest and dividends or retained by the credit union as capital; none goes to federal taxes. In contrast, before a bank can make dividend payments to its owners (shareholders) or retain capital, it must pay a portion of its net earnings as federal income taxes.[7]

[7] Since credit unions do not raise capital by issuing equity to investors, they are at a disadvantage compared to banks, which can quickly raise additional capital in the equity markets.

While commercial banks have never been tax exempt and mutual savings and loans and mutual savings banks lost their tax-exempt status in the Revenue Act of 1951, credit unions remain tax exempt even though this status has been questioned by representatives of other portions of the financial services industry.

Section 501(c)(1) of the Internal Revenue Code grants tax-exempt status to federal credit unions. Section 501(c)(14) of the Code does the same for state-chartered credit unions as long as they do not have capital stock, are mutuals, and are nonprofit. In 1998, legislators explained and affirmed credit unions' tax-exempt status in the Credit Union Membership Access Act. The act stated that credit unions are tax exempt because they "are member-owned, democratically operated, not-for-profit organizations generally managed by volunteer boards of directors and because they have the specified mission of meeting the credit and savings needs of consumers, especially persons of modest means." (U.S. House of Representatives, Committee on Ways and Means 2005).

Credit unions are granted an advantage relative to their competitors because they are tax exempt. One measure of the size of this advantage is the total amount of taxes avoided. In a 2001 study, the Department of Treasury estimated that federal tax revenues of between $1.3 billion and $1.6 billion per year were lost due to credit unions' tax-exempt status (U.S. Department of the Treasury 2001a).

If credit unions paid this $1.6 billion to their depositors in the form of greater interest on deposits, the advantage would amount to about 30 basis points ($1.6 billion divided by $550 billion in total credit union deposits equals .29 percent). Alternatively, if they passed it along in lower loan interest rates, the basis point advantage would be about the same (as the dollar amount of loans and deposits are about the same).

Still, it seems likely that at least part of the tax advantage move in other directions. Some of the advantage is likely to be shifted toward capital holdings at credit unions and higher salaries and expenses.

While tax exemption may provide as much as a 30 basis point advantage that can be used to attract customers to a credit union, the advantage is undercut by two other considerations. The first consideration is that many small banks are organized as Subchapter S corporations. Within certain limits, S corporations pay no federal income taxes and can, like credit unions, devote all earnings either to dividends and interest, retained capital, or some of all three. As of March 2006, about 2,400 FDIC-insured institutions (commercial banks,

Instead, credit unions increase equity only by retaining earnings. The inability to quickly add to their equity can be a disadvantage if they wish to grow rapidly. This disadvantage, compared to banks, is offset somewhat because credit union earnings are not diminished by taxes, leaving more earnings that might be retained.

savings and loans, and savings banks), or 27 percent of all FDIC-insured institutions, were organized as S corporations (Federal Deposit Insurance Corporation 2006, 4–5).

The second consideration that limits, somewhat, the advantage of credit unions over non-tax-exempt banks is the current fairly low dividend tax rate of 15 percent. As an example, assume a taxed bank earns $100 million dollars, pays federal taxes (of $35 million at the top federal corporate tax rate of 35 percent), and pays out as dividends all of its remaining earnings. Its shareholders then pay taxes on the dividends at the current dividend tax rate of 15 percent. Ultimately, bank shareholders receive $55 million in earnings.

In contrast, shareholders (customers) in credit unions receive $65 million after taxes. While they do not face double taxation as corporate shareholders do, since the credit union pays no taxes before paying them interest or dividends, the tax rate on payments to credit union shareholders is the individual tax rate, which can be as high as 35 percent. The bottom line is that the tax advantage may be smaller than one might imagine.[8]

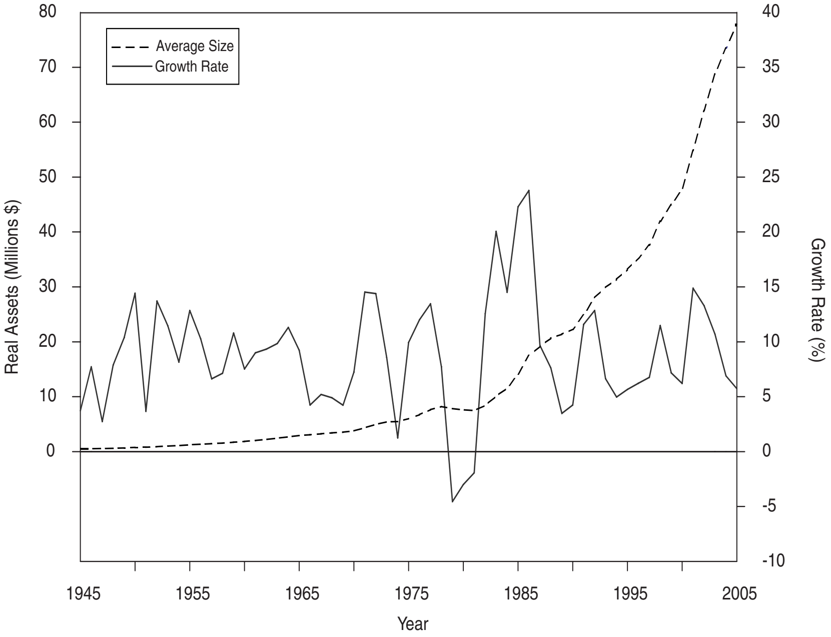
4. SIZE RELATIVE TO BANKS

As of December 2005, there were about 9,000 credit unions, 3,600 of which were state-chartered and 5,400 of which were federal credit unions (Credit Union National Association 2006). In sum, these 9,000 credit unions held $700 billion in assets. In comparison, there were 8,800 banks, savings and loans, and savings banks, together holding $10.9 trillion in assets. Several banks, individually, have assets that exceed the total assets of the sum of all credit unions. Nevertheless, with 87 million members as of 2005, credit unions are important competitors in the market for consumer loans and deposits (Credit Union National Association 2006).

Like banks, credit unions come in a wide range of sizes. Still, compared to banks, a much higher proportion of credit unions are very small. Approximately 46 percent of all credit unions are smaller than $10 million in assets (Credit Union National Association 2006). Only 79 banks, or 1 percent of all banks, have assets below this threshold. In part, this size difference reflects the long-standing focus of credit unions on retail deposits and loans compared to banks' focus on lending to firms and holding their deposits. Commercial loans account for only 2.5 percent of all credit union assets, though this amount has grown fairly rapidly in recent years. This contrasts with banks where

[8] Because these calculations of amounts received by bank and credit union shareholders are made assuming the highest tax rates (corporate and individual rates), they cannot be considered accurate for all shareholders. For example, credit union shareholders in lower tax brackets (presumably a majority of shareholders) will receive higher after-tax earnings.

Figure 1 Average Credit Union Size and Percentage Growth in Size



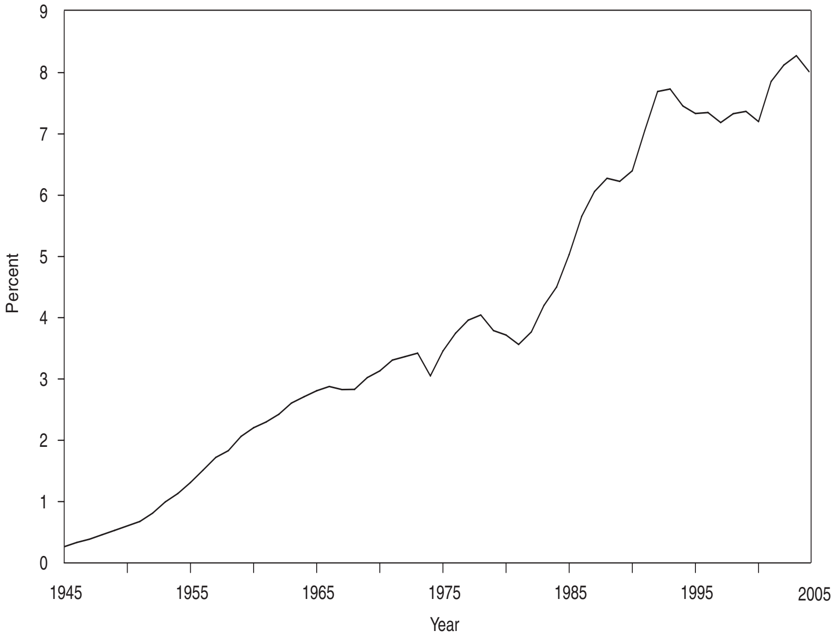
Notes: Commercial bank data are from FDIC. Credit union data are from Credit Union National Association.

commercial loans amount to about 22 percent of total assets (Credit Union National Association 2006).

Nevertheless, there are some large credit unions, even by bank standards. For example, 104 credit unions had assets greater than $1 billion as of June 2005 (Callahan and Associates 2006, 44–46).

While compared to banks, most credit unions are small, as Figure 1 shows. Yet, the size of the average credit union has grown significantly over the years. Figure 2 demonstrates that, relative to banks, credit union assets have grown fairly consistently. They increased from less than 1 percent of bank assets in 1944 to about 3.5 percent in 1981. Growth was especially rapid relative to banks during the 1980s and early 1990s, as membership restrictions were first being eased and banks were experiencing a widespread decrease in profits. More recently, growth has been less robust in relation to bank assets (see Figure 2).

Figure 2 Credit Union Assets as a Percent of Bank Assets



Notes: Commerical bank data are from FDIC. Credit union data are from Credit Union National Association.

5. REGULATORY FRAMEWORK

Historical Development

The first credit union in the United States operated under a special charter granted specifically to that credit union by the New Hampshire legislature in 1909.[9] As noted earlier, over a period of years following the 1909 chartering of the first credit union, other states passed laws giving credit-union-chartering authority to state banking agencies. Credit unions were required to seek state charters, otherwise they would have run afoul of state laws stipulating that deposits may be accepted only by entities holding government-issued charters.

[9] A 1922 study of cooperative lending associations notes that there were several credit union-like financial institutions operating even before 1909. For example, a cooperative association was created for employees of the Boston Globe in 1892. It gathered deposits from employees and made small-value loans to them. The Boston Globe Cooperative Association, and several other similar associations, operated without legal authority, leading the Massachusetts banking commissioner, Pierre Jay, to call for a law authorizing these activities (Whitney 1922, 17).

These early charter laws typically specified rules such as how credit unions might be founded, from whom they might gather deposits (credit union members), to whom they might make loans (typically only to members), who might become members (those sharing a community or occupational bond), and how an individual might become a member (by purchasing at least one share in the credit union) (Whitney 1922, 21–29). State banking agencies not only chartered credit unions, but also oversaw their operations, including sending examiners to review their activities (Whitney 1922, 48).

A federal charter for credit unions was created in 1934. At that time, 39 states had laws authorizing credit unions and there were 2,028 state credit unions (Giles 1951, 106; U.S. Bureau of the Census 1975, 1,049). The Federal Credit Union Act, enacted on June 26, 1934, created a federal agency to grant charters so that credit unions could be formed in any state regardless of whether that state offered charters to credit unions. The federal agency was placed in the Farm Credit Administration (Moody and Fite 1984, 120).

Over the years the agency responsible for federal credit unions was housed in government entities. It was moved to the FDIC in April 1942 and then in July 1948 to the Federal Security Agency—later renamed the Department of Health, Education, and Welfare (Croteau 1956, 121).

Current Regulatory Environment

In March 1970, legislation was enacted to create an independent agency, the National Credit Union Administration (NCUA) (Moody and Fite 1984, 304). The NCUA charters and supervises all federal credit unions. State credit unions continue to be chartered and examined by state agencies.

Also in 1970, credit unions gained federal deposit insurance when the Federal Credit Union Act was amended in October of that year to create the National Credit Union Share Insurance Fund (NCUSIF). The Fund is overseen by the NCUA. All federal credit unions must join the fund and state-chartered credit unions may choose to join.

The amendment that created the NCUSIF passed in spite of the objections of the credit union industry, which argued that insurance premiums would be high and losses from credit union failures had been quite low historically (Moody and Fite 1984, 304). While opposed in the beginning, insurance is now almost universal; only 191 state-chartered credit unions, out of the total of 3,600 such credit unions, operate without NCUSIF coverage (Callahan and Associates 2006, 19).

6. ADVANTAGES OF COMMON BOND LENDING DIMINISH

In the second half of the 20th century, a number of factors combined to slowly undercut the advantages of common bond lending relative to other forms of lending. First, nationwide credit-reporting agencies emerged and expanded in the 1970s, lowering the cost of lending to consumers with whom the lender had no direct contact. Credit-reporting agencies are companies that gather and sell information about individual consumers, information that can be used to predict the likelihood that a consumer will repay a debt. Second, in the 1980s and 1990s the home equity line of credit was developed and became widely available, allowing consumers to easily tap their home equity for small purchases. Third, credit card lending developed as a convenient, and reasonably low-cost means of making small, unsecured loans. And fourth, the advent of deposit insurance eliminated some of the social pressure that might be brought to bear on members who failed to repay loans.

Growing Availability of Creditworthiness Information

Until the 1970s, credit information was gathered largely only on a local basis, and on just a limited number of consumer variables. For example, Trans Union, one of several nationwide credit-reporting agencies, entered the credit-reporting business in 1969 by purchasing a local credit-reporting agency, and then expanding nationwide, and developing the computer technology to store and retrieve broad, detailed records on millions of consumers over the next several years (Trans Union 2006b). Before computers were employed, information was necessarily limited to a few items on each consumer, and was difficult to retrieve quickly.

Today, three credit-reporting agencies—Equifax Credit Information Services, Trans Union, and Experian Information Solutions—gather and maintain information that lenders consider predictive of debt repayment on the majority of United States consumers. Such information can include the consumer's debt and bill repayment history, bankruptcies, liens, number of loans and lines of credit, amount owed on each, home address, and employer name (Trans Union 2006a). For a fee, a reporting agency will supply this information about a consumer to a lender considering making a loan to that consumer, as well as a score that grades the consumer's creditworthiness based on the information. A lender might gather scores and the underlying information from two or three of the agencies before deciding to make a loan. Because many borrowers know that a default will mean a lowered credit score and higher interest rates on future loans, the presence of these credit agencies provides an incentive to repay. Because of the low cost and ease with which lenders can gather consumer creditworthiness information, and the repayment incentive their presence provides, the relative advantages of creditworthiness knowledge gained by maintaining direct contact with the borrower through a common bond and the motivation to repay produced by common bond relationships are reduced.

Home Equity Lines of Credit

The spread of home equity lines of credit (HELOCs) also tended to undercut the advantage of common bond lending. Prior to the mid-1980s spread of home equity loans, it was difficult for a consumer to make use of the equity built up in his or her house as collateral for a loan (Canner, Durkin, and Luckett 1998, 242). Second mortgages were often taken on homes to make a large, one-time purchase. But closing and other transaction costs made such loans expensive to use for making smaller and more frequent consumer purchases.

The development of HELOCs greatly expanded the ability of consumer borrowers to tap the equity in their homes and, thereby, collateralize borrowings for occasional consumer expenditures.[10] With a HELOC, the consumer bears the closing costs only once, but then can borrow as frequently as desired by simply writing a check against the HELOC. Further, such loans are far less risky than uncollateralized loans, and, therefore, allow lenders to offer consumers a lower rate of interest. The growing availability of HELOCs greatly diminished the need to employ the collateral substitution ability provided by common bond lending, at least for consumers with home equity.

Credit Card Lending Use Expands

Credit card lending makes use of a cost savings similar to that provided by HELOC lending. Once the credit card lending arrangement is established, costs of drawing down the credit card line are minimal for the lender and for the consumer. The credit card company makes an initial credit decision, based largely upon information from credit-reporting agencies, on whether to grant a loan to a consumer, at what rate, and how large a line. The consumer is then free to draw on the line in whatever increments he or she chooses, and to pay it down over a period largely determined by the consumer.

Credit card availability grew rapidly following a 1978 Supreme Court decision, which allowed credit card lenders to avoid state usury ceilings (Athreya 2001, 11–15). Because credit cards offer low-transaction-cost lending, based on detailed creditworthiness information from the files of credit reporting agencies, credit card lending offered an additional strong alternative to common bond lending.

Credit Unions Receive Federal Deposit Insurance

The 1970 application of federal deposit insurance to credit unions also tended to undercut an advantage of common bond lending. Before federal deposit

[10] Weinberg (2005) provides a thorough analysis of the growth in recent decades of consumer debt, the causes of this growth, and its consequences.

insurance was extended to credit unions, credit union members had reason to monitor one another's loan repayment. Any defaults increased the probability of the failure of the credit union and losses to its depositors. Sharing a common bond also meant that borrowers were likely to have frequent contact with one another and, therefore, the opportunity to bring pressure on delinquent borrowers. Additionally, frequent contact implies that a borrower might face significant embarrassment if he or she defaults.[11]

But once credit unions gained federal deposit insurance, the incentive to monitor repayment was reduced. With insurance in place, members are protected against loss of principal and interest if a default or several defaults cause the failure of the credit union. If the default is too small to cause the failure of the credit union but leads to a reduction in interest payments, a competitive market for deposits means that members can simply shift their deposits to another depository, paying an unreduced interest rate.

7. CREDIT UNIONS EVOLVE

As the benefits of common bond lending declined for reasons discussed previously, the credit union industry moved in directions that tended to deemphasize this form of lending. As long as common bonds were important to consumer lending, one would expect credit unions to operate with fairly small memberships in order to take advantage of the tighter bonds present in smaller groups. Once the advantage of the common bond is undercut, credit unions should tend to move toward the size of other depositories that are not focused on common bond lending. Further, one would expect credit unions to begin to place greater emphasis on other forms of lending. This movement is evident (1) in the growing average size of credit unions, encouraged to a degree by liberalizations that broadened the potential membership base of any given credit union, and (2) the move of credit unions into real estate, credit card, and business lending.

Common Bond Requirement Loosened

The strict common bond requirement for credit unions ended in 1982. In that year, the NCUA began to allow single credit unions to draw members from multiple groups. At the time, the failure rate of credit unions was at a historically high level. Failures were likely driven, in part, by the inability of credit unions to achieve diversification due to tight common bond membership limitations (Wilcox 2005, 1, 3). For example, a credit union with a membership

[11] Athreya (2004) finds evidence that fear of embarrassment or stigma remains today as an important incentive to repay. His finding contrasts with an often-cited view that the stigma has diminished.

comprised completely of employees of one firm is likely to face heavy losses if the employer goes out of business.

While in 1998 the Supreme Court determined that the NCUA had overstepped its statutory limits when it approved multiple groups for one credit union, this decision was nullified almost immediately by legislation. In 1998, shortly after the Supreme Court decision, Congress passed, and President Clinton signed, the Credit Union Membership Access Act, which authorized individual credit unions to serve multiple groups, within some restrictions. The credit union industry had lobbied aggressively for passage of this liberalization (Gill 1998).

The move to allow multiple groups along with mergers of credit unions, which were frequent in the 1980s, offered credit unions the opportunity to enlarge their membership base and size. As seen in the growth rates plotted in Figure 1, year-after-year growth in the average size of credit unions increased, in constant dollars, at historically high rates between 1982 and 1987, following the liberalization of membership requirements. Overall, average size increased from $7.5 million in 1981 to $77.7 million in 2005; though the annual growth rate after 1987 is about the same as it was before 1982.

In 1920, the average number of members per credit union was 200. By this measure, credit unions remained fairly small, even in 1960, with an average of 598 members per credit union. By 2005, the average credit union membership had risen to 10,000. Regarding lending decisions, there is probably little opportunity to use special knowledge gained from direct contact between members.[12] Further, it seems likely that embarrassment and social pressure will play a less important role in institutions with thousands of members than in small institutions in which members are known to one another.

Credit Unions Gain Mortgage Lending Powers

Federal credit unions gained the authority to make long-term (up to 30 years) mortgage real estate loans in 1977 with amendments made to the Federal Credit Union Act by Public Law 95-22. Before these amendments passed, federal credit unions were limited to providing short-term mortgage loans such as second mortgages and mobile home loans (Credit Union National Association 1978, 12).

State-chartered institutions in a number of states had this authority prior to 1977. Consequently, at the beginning of 1978, credit unions held $2.55 billion in first mortgages, which accounted for 6.2 percent of all credit union

[12] Karlan (2005) finds that for members of a group-lending organization in Peru, loans are more frequently repaid by members of groups who live close to one another and who are more culturally similar. This result occurs because such individuals are better able to monitor one another.

loans. Since 1977, first mortgage real estate lending has grown in importance as a percentage of credit union lending. As of 2005, mortgages amounted to 32 percent of all loans (Credit Union National Association 2006). As common bond lending became less necessary, credit unions moved more extensively into real estate lending.

Growth of Credit Card Lending

Similarly, credit card lending has grown in importance for credit unions, so that today such lending accounts for a large percentage of all unsecured lending by credit unions. Approximately 54 percent of credit unions offer credit cards, and the percentage rises from very low at the smallest credit unions to 98 percent at the largest (Credit Union National Association 2005). Since this type of lending is most important for the largest credit unions, (those with the largest membership) it seems unlikely that common bond lending can play much of a role in this new category of credit union assets.

Expanded Business Lending

During the last several years, credit unions expanded their business lending significantly, doubling the amount of such loans over the three years before 2005 (Credit Union National Association 2006). Specifically, business loans increased from $8.3 billion, in 2002, to $16.9 billion, or 2.5 percent of credit union assets, in 2005. In comparison, in 2005, business loans accounted for 22 percent of total assets at commercial banks (FDIC 2005). Since these loans are typically collateralized and made for business purposes, they are beyond the historical credit union emphasis on uncollateralized consumer lending.

Still, the extent of business lending by credit unions is limited. The 1998 Credit Union Membership Access Act set limits on credit union lending to businesses. Specifically, the sum of all business loans of a federal credit union may not exceed 1.75 percent of its net worth or 12.25 percent of its total assets, whichever is the lesser (National Credit Union Administration 2004, Section 723.16).

8. CONCLUSION

Credit unions once focused almost entirely on small-value, unsecured consumer lending. Other depositories, specifically commercial banks, savings banks, and savings and loans, emphasized lending to businesses or making large-dollar-value consumer loans based on real estate collateral. For unsecured consumer lending, the credit union model of taking deposits from and lending to a tightly knit group of members of a community or organization was an ideal business model in the early part of the 20th century, filling a niche in the marketplace at that time.

But the financial marketplace changed greatly in the second half of the 20th century. Computer databases and networks allowed the inexpensive collection and dissemination of consumer creditworthiness information. Home ownership spread and products developed that allowed consumers to borrow against their home equity for repeated small-value loans. Further, the availability and use of credit cards as a vehicle for unsecured consumer loans expanded immensely. The combination of these factors required credit unions to change as well.

Credit unions shifted away from their heavy emphasis on unsecured consumer loans into mortgage lending, other forms of collateralized consumer lending (such as auto loans), credit card lending, and business lending. Today, unsecured consumer loans account for only about 10 percent of all credit union loans. Auto loans are responsible for 37 percent, and mortgages comprise most of the remainder (Credit Union National Association 2005).

Additionally, while the very small credit union was well-suited to common bond lending, small size became less advantageous as common bond lending lost its advantage over the last several decades. Concurrently, the average number of members and size of credit unions increased significantly. Nevertheless, a large portion of credit unions remain quite small in comparison to commercial banks. By evolving with changes in the marketplace, the credit union industry remained healthy and grew, both in terms of dollar amounts and relative to other depositories.

REFERENCES

Athreya, Kartik. 2001. “The Growth of Unsecured Credit: Are We Better Off?” Federal Reserve Bank of Richmond *Economic Quarterly* 87 (3): 11–33.

Athreya, Kartik. 2004. “Shame as It Ever Was: Stigma and Personal Bankruptcy.” Federal Reserve Bank of Richmond *Economic Quarterly* 90 (2): 1–19.

Besley, Timothy, and Stephen Coate. 1995. “Group Lending, Repayment Incentives and Social Collateral.” *Journal of Development Economics* 46: 1–18.

Board of Governors of the Federal Reserve System. 1959. *All Bank Statistics: United States, 1896-1955.* Washington, D.C.: Federal Reserve System.

Calder, Lendol. 1999. *Financing the American Dream: A Cultural History of Consumer Credit*. Princeton, NJ: Princeton University Press.

Callahan and Associates. 2006. 2006 *Credit Union Directory*. Washington, D.C.: Callahan and Associates.

Canner, Glenn B., Thomas A. Durkin, and Charles A. Luckett. 1998. "Recent Developments in Home Equity Lending." *Federal Reserve Bulletin*. (April).Washington, D.C.: Federal Reserve Board.

Carruthers, Bruce G., Timothy W. Guinnane, and Yoonseok Lee. 2005. "The Passage of the Uniform Small Loan Laws." Mimeo. (March) Available at: http://www.lse.ac.uk/collections/economicHistory/seminars/Guinnane.pdf#search=%22uniform%20small%20value%20loan%20laws%22 (accessed on August 15, 2006).

Credit Union National Association. 1978. "Credit Unions: Progress for People." *1978 Yearbook*. Madison, WI: Credit Union National Association.

Credit Union National Association. 2005. *Credit Union Report: Year End 2005*. Madison, WI: Economics and Statistics Department Available at: http://advice.cuna.org/download/curepd05.pdf (accessed on August 15, 2006).

Credit Union National Association. 2006. "Long-Run Trends–1939-Present: Aggregates." Table. Available at: http://advice.cuna.org/download/us_totals.pdf (accessed on August 15, 2006).

Croteau, John T. 1956. *The Federal Credit Union: Policy and Practice*. New York, NY: Harper and Brothers.

Fama, E. F., and M. C. Jensen. 1983. "Agency Problems and Residual Claims." *Journal of Law and Economics* 26 (2): 327–49.

Federal Deposit Insurance Corporation. 2005. *Quarterly Banking Profile, Fourth Quarter 2005*. Available at: http://www2.fdic.gov/qbp/qbpSelect.asp?menuItem=QBP (accessed on August 15, 2006).

Federal Deposit Insurance Corporation. 2006. *Quarterly Banking Profile, First Quarter 2006*. Available at: http://www2.fdic.gov/qbp/qbpSelect.asp?menuItem=QBP (accessed on August 15, 2006).

Flannery, Mark J. 1974. *An Economic Evaluation of Credit Unions in the United States*. Research Report No.54. Boston, MA: Federal Reserve Bank of Boston.

Frederick, Donald A. 1997. “Co-ops 101: An Introduction to Cooperatives.” Cooperative Information Report 55, Rural Business-Cooperative Service, U.S. Department of Agriculture. Available at: http://www.rurdev.usda.gov/rbs/pub/cir55/cir55rpt.htm (accessed on August 15, 2006).

Giles, Richard Y. 1951. *Credit for the Millions: The Story of Credit Unions*. New York, NY: Harper and Brothers.

Gill, Buddy. 1998. “Credit Unions vs. Banks—Credit Unions Launch Grassroots Campaign to Lobby for Legislation in Congress.” *Campaigns & Elections* (July).

Karlan, Dean S. 2005. “Social Connections and Group Banking.” Center Discussion Paper No. 913. Economic Growth Center, Yale University. Available at: http://www.econ.yale.edu/growth_pdf/cdp913.pdf (accessed on August 15, 2006).

Mayers, D., and C. W. Smith. 1986. “Ownership Structure and Control: The Mutualization of Stock Life Insurance Companies.” *Journal of Financial Economics* 16 (1) (May): 73–98.

Mayers, D., and C. W. Smith. 2002. “Ownership Structure and Control: Property-Casualty Insurer Conversion to Stock Charter.” *Journal of Financial Services Research* 21 (1-2): 117–44.

Moody, J. Carroll, and Gilbert C. Fite. 1984. *The Credit Union Movement: Origins and Development, 1850-1980*. Dubuque, IA: Kendall/Hunt Publishing Company.

National Credit Union Administration. 2004. *Rules and Regulations* (April). Available at: http://www.ncua.gov/RegulationsOpinionsLaws/rules_and_regs/NCUA6.pdf (accessed on August 15, 2006).

National Credit Union Administration. 2006a. ”History of Credit Unions.” Available at: http://www.ncua.gov/AboutNCUA/historyCU.html (accessed on August 15, 2006).

National Credit Union Administration. 2006b. *Financial Performance Reports*. Available at: http://www.ncua.gov/data/FOIA/foia.html/2006MarchAggregate[1].zip/FedandStateAggregate.xls (accessed on August 15, 2006).

Prescott, Edward S. 1997. “Group Lending and Financial Intermediation: An Example.” Federal Reserve Bank of Richmond *Economic Quarterly* 83 (4): 23–48.

Robbins, Eric. 2005. “Credit Union Growth in the Tenth District: How Legal and Regulatory Changes Have Affected Credit Union Expansion.”

Federal Reserve Bank of Kansas City *Financial Industry Perspectives*. (July).

Trans Union, LLC. 2006a. “How to Read Your Credit Report.” Available at: http://www.truecredit.com/help/spotlight.jsp?cb=TransUnion&loc=1761 &bn=null.(accessed on August 15, 2006).

Trans Union, LLC. 2006b. “Company History.” Available at: http://www.transunion.com/corporate/aboutUs/whoWeAre/history. (accessed on August 15, 2006).

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

U.S. Bureau of Federal Credit Unions. 1948. *Federal Credit Unions, Report of Operations for the Year of 1948*. Federal Security Agency, Social Security Administration, Washington, D.C.: Bureau of Federal Credit Unions.

U.S. Department of Health, Education, and Welfare. 1965. *State-Chartered Credit Unions: 1965*. Washington, D.C.: Bureau of Federal Credit Unions.

U.S. Department of the Treasury. 1997. “Credit Unions.” A Study Prepared Pursuant to the Economic Growth and Regulatory Paperwork Reduction Act of 1996. Available at: http://www.treas.gov/press/releases/docs/cu_study.pdf. (accessed on August 15, 2006).

U.S. Department of the Treasury. 2001a. “Comparing Credit Unions With Other Depository Institutions.” A Study Prepared Pursuant to the Credit Union Membership Access Act of 1998. Available at: http://www.ustreas.gov/press/releases/reports/report30702.doc. (accessed on August 15, 2006).

U.S. Department of the Treasury. 2001b. “Credit Union Member Business Lending.” A study prepared pursuant to the Credit Union Membership Access Act of 1998. Available at: http://www.treasury.gov/press/releases/reports/mblstudy.doc. (accessed on August 15, 2006).

U.S. House of Representatives. 2005. *Advisory*. Committee on Ways and Means, No. FC-15. (November 3, 2005). Available at: http://waysandmeans.house.gov/hearings.asp?formmode=printfriendly& id=4130 (accessed on August 15, 2006).

Walter, John R. 2006. “Mixing Banking and Commerce.” Federal Reserve Bank of Richmond *Region Focus.* (Summer).

Weinberg, John A. 2005. “Borrowing by U.S. Households.” Federal Reserve Bank of Richmond *Annual Report*.

Whitney, Edson L. 1922. *Cooperative Credit Societies (Credit Unions) in America and in Foreign Countries*. U.S. Department of Labor, Bureau of Labor Statistics, No. 314 (November). Washington, D.C.: Government Printing Office.

Wilcox, James A. 2005. “Credit Union Failures and Insurance Fund Losses: 1971–2004.” Federal Reserve Bank of San Francisco *Economic Letter*, No. 2005-20 (August 19, 2005). Available at: http://www.frbsf.org/publications/economics/letter/2005/el2005-20.pdf (accessed on August 15, 2006).