Inflation and Changing Expenditure Shares

Alexander L. Wolman and Fan Ding

Inflation is an index of price changes for many goods. As such, the behavior of inflation is determined by the behavior of (1) price changes for individual goods, as well as (2) the weights that the index puts on the price changes of different goods. Most macroeconomic analyses of the time-series behavior of inflation—whether empirical or theoretical—implicitly emphasize the former determinant of inflation.[1] Theoretical analyses tend to focus on one-sector models in which there are no weights to shift, and empirical analyses tend to focus on the univariate properties of some broad inflation rate.

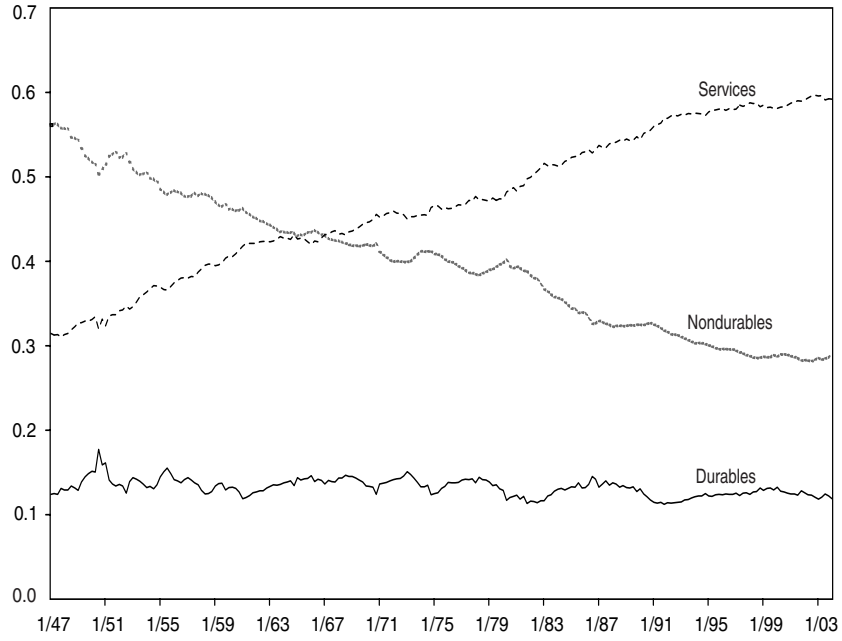
If rates of price change did not differ much across goods, then shifts in the weights would not matter much for inflation. In fact, there has been substantial variation in price change behavior across goods, and the weights on two of the three broad categories in consumption price indexes have shifted dramatically over the last 50 years (Figure 1). Those facts motivate us to investigate the importance of changing weights for three fundamental time-series properties of inflation: level, volatility, and persistence. The extent to which shifting weights are important for these properties may have implications for macroeconomic modeling. Suppose that inflation was highly persistent but that all of the persistence was accounted for by long-term shifts in the weights in the inflation measure. We might then conclude that in one-sector macroeconomic models, high inflation persistence is not a desirable feature.

We propose and implement two approaches to measuring the contribution of changing expenditure shares to inflation behavior. Both involve constructing an alternative inflation measure that holds fixed the weights on price

We wish to thank Bob Hetzel, Andreas Hornstein, Tom Humphrey, and Eric Nielsen for helpful comments and Sam Malek for his assistance. The views here are the author's and should not be attributed to the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Wolman (1999) is an example of an article that fits this description.

Figure 1 Sectoral Expenditure Shares



changes for different goods. We describe the behavior of the level, volatility, and persistence of the alternative inflation measures. The role of changing expenditure shares is then revealed by the divergence between the behavior of actual inflation and the fixed-weight measures. Neither approach leads to a dramatic revision in our understanding of post-war U.S. inflation; that is, the broad features of inflation over the past 50 years cannot be accounted for by changing expenditure shares. However, in more subtle ways, changing expenditure shares have been important for the behavior of inflation. For example, we attribute 15 basis points of quarterly inflation, on average, to changing expenditure shares over the period from 1947 to 2004. Expenditures have shifted to services, and the relative price of services has risen persistently over the last 50 years. This shift toward services has tended to make the overall inflation rate higher, other things equal. The caveat "other things equal" is important. Expenditure share shifts have been one factor influencing the behavior of inflation, but monetary policy has had the ability to counteract the effect of shifting expenditure shares on inflation. Thus, one could reinterpret the statement above as "in order to achieve the inflation behavior we have observed, monetary policy has had to counteract a 15-basis-point upward effect on inflation coming from the long-run shift in expenditures toward services."

It is important to make clear at the outset that we are not arguing that one should measure inflation by holding fixed the weights on different goods. It is well known that good price indexes from the standpoint of economic theory ought to have time-varying weights that reflect time-varying expenditure patterns. Our concern is instead one of fact-finding: Given the existence of changes in expenditure shares, to what extent can those changes account for the behavior of inflation? To answer this question, we construct alternative fixed-weight price indexes.

For the most part, recent literature on inflation in the United States has abstracted from the heterogeneity that underlies overall inflation. Notable exceptions are Clark (2003) and Bauer, Haltom, and Peterman (2004). Bauer, Haltom, and Peterman focus on the behavior of core inflation over the last 20 years. They decompose core inflation into contributions of different goods and services. These contributions are the product of expenditure shares and individual price changes. Bauer, Haltom, and Peterman find that just two components, rent and used vehicles, account for much of the decline in consumer price index (CPI) inflation over this period. Clark's emphasis is on inflation persistence, which we will discuss further. He contrasts the behavior of inflation persistence over time to the behavior of the persistence of disaggregated price changes. He finds that the persistence of disaggregated price changes tends to be lower than the persistence of inflation. Our article differs in its explicit emphasis on changing expenditure shares over time. Clark's findings, though, suggest that expenditure share behavior may be an important determinant of inflation persistence.

1. INFLATION IN THE UNITED STATES

The variables we are concerned with are all produced by the Bureau of Economic Analysis of the United States Department of Commerce. They are the price index for personal consumption expenditure; the subindexes for durable goods, nondurable goods, and services; and the expenditure shares for durable goods, nondurable goods, and services. Before turning to the behavior of these variables, it is useful to provide some background on price indexes, and, in particular, on the price index for personal consumption expenditure. (Henceforth, we will refer to this index as the PCE price index, and to its rate of change as PCE inflation.)

PCE inflation data are constructed from underlying price and quantity data for a large number of categories of goods and services. In turn, the price data for those underlying categories are constructed from more direct observation of prices on an even larger number of specific items (i.e., goods and services). The latter construction is performed mainly by the Department of Labor's Bureau of Labor Statistics. For the most part, the same item prices that form the basis for PCE inflation also form the basis for the more widely known CPI

inflation, which is produced by the Bureau of Labor Statistics. We focus here on PCE inflation for two reasons. First, the methodology used to produce the PCE inflation numbers corresponds more closely to notions of price indexes suggested by economic theory. Second, the PCE methodology makes it more straightforward to decompose inflation in a way that isolates the effect of changing expenditure shares.

The formula used to create the PCE inflation rate is known as a Fisher ideal index. We will first provide the formula and then interpret it.[2] We define π_t to be the PCE inflation rate in quarter t, $x_{i,t}$ to be the period t dollar expenditures on category i, and $\pi_{i,t}$ to be the rate of price change for category i from period $t-1$ to period t. The PCE inflation rate is

$$\pi_t = \sqrt{\left[\sum_{i=1}^{I} \omega_{i,t-1}\pi_{i,t}\right]\left[\sum_{i=1}^{I} \theta_{i,t}\pi_{i,t}\right]}, \tag{1}$$

where

$$\omega_{i,t} \equiv \frac{x_{i,t}}{\sum_{j=1}^{I} x_{j,t}}, \text{ and}$$

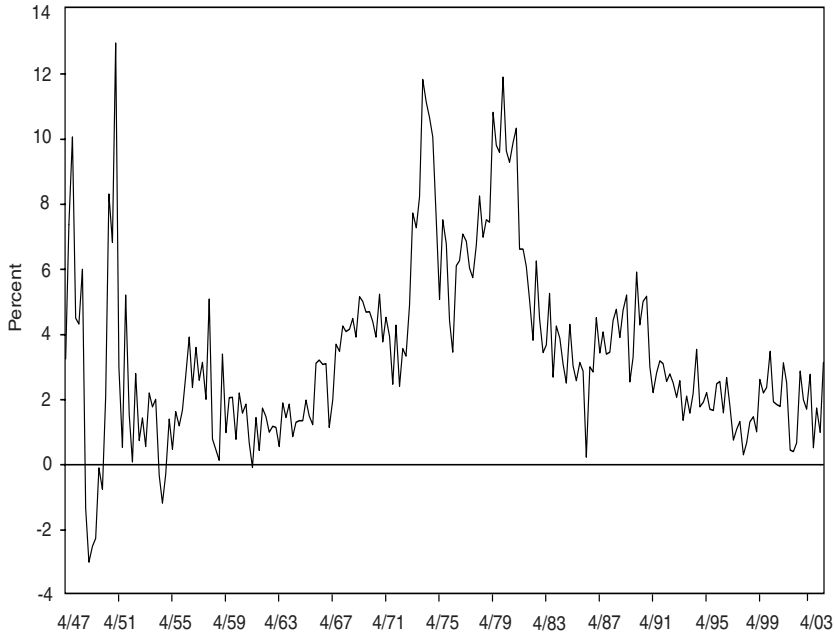
$$\theta_{i,t} \equiv \frac{x_{i,t}/\pi_{i,t}}{\sum_{j=1}^{I}\left(x_{j,t}/\pi_{j,t}\right)}, \text{ .for } i = 1, ..., I.$$

Both objects in square brackets in (1) are weighted averages of the rates of price change for each good and service. The weights, $\omega_{i,t-1}$, are simply the expenditure shares for category i in period $t-1$; thus, the first weighted average, $\sum_{i=1}^{I} \omega_{i,t-1}\pi_{i,t}$, measures the rate of price change for the basket of goods purchased in period $t-1$. The weights, $\theta_{i,t}$, are the hypothetical expenditure shares that are derived by combining period t real quantities with period $t-1$ prices. Thus, the second weighted average, $\sum_{i=1}^{I} \theta_{i,t}\pi_{i,t}$, measures the rate of price change in period t for the basket of goods purchased in period t. Finally, PCE inflation (π_t) is the geometric average of these two inflation rates.

It is clear from (1) that changes in expenditure shares on different goods and services are incorporated in the behavior of the PCE. In contrast, the CPI is a fixed-weight index; changes in expenditure shares are incorporated in the CPI only every two years. The precise way in which changing expenditure shares are incorporated in PCE inflation is somewhat complicated, as seen in (1). Fortunately, for our purposes, the true PCE inflation rate is well approximated by a simpler formula that aggregates prices for the three major spending categories using what is known as a Divisia index. The Divisia approximation

[2] See Webb (2004) and Clark (1999) for more detailed discussions of how the PCE price index is constructed.

Figure 2 PCE Inflation



to the PCE which we will use is

$$\pi_t^D = \sum_{i=N,D,S} \omega_{i,t-1}\pi_{i,t}, \tag{2}$$

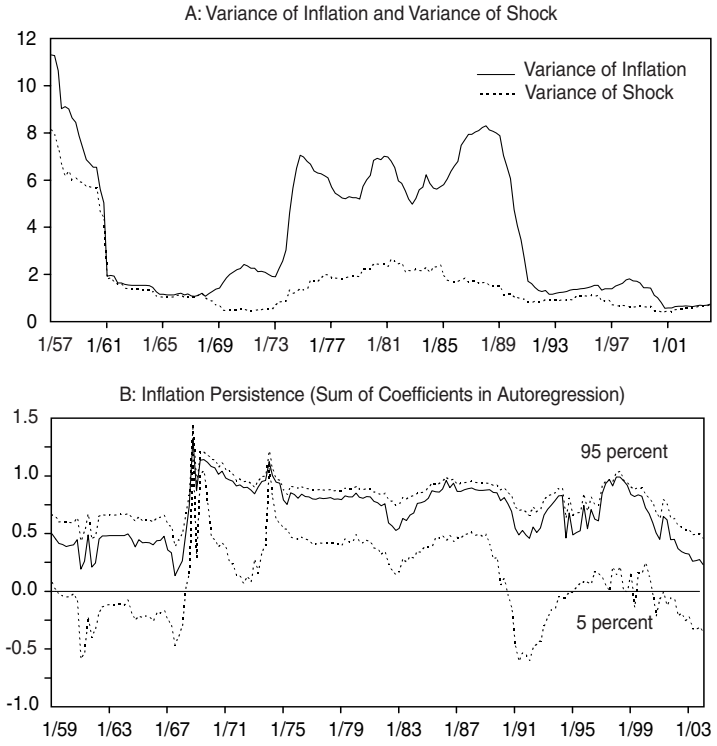
that is, the expenditure-share-weighted average of price changes for nondurable goods, durable goods, and services. This approximation is convenient because it allows us to easily decompose the behavior of inflation into the part accounted for by changing expenditure shares and the part accounted for by changing rates of price change for the main spending categories.

The Level of Inflation

Figure 2 displays the quarterly PCE inflation rate from 1947 to 2004, expressed in annualized percentage terms.[3] This figure displays the major facts about inflation in the United States. Inflation was highly volatile immediately after World War II, then declined and became more stable during the 1950s. In the

[3] In all figures, the month and year on the x-axis indicate the first month of the quarter represented by the tick mark.

Figure 3 Volatility and Persistence



mid-1960s, inflation began a steady rise that continued for the rest of the decade. The 1970s were characterized by high and volatile inflation, and then in the early 1980s inflation declined dramatically. Over the last 15 to 20 years, inflation has been low and stable, apart from a moderate increase in the late 1980s. The average PCE inflation rate from 1947 to the present has been 3.42 percent. Though these basic facts are clear, much about the behavior of the level of U.S. inflation remains in dispute. For example, economists agree that the Federal Reserve can determine the average level of inflation over periods of several years. Thus, there is consensus that the Federal Reserve could have brought about a much lower average inflation rate in the 1970s. However, there is no consensus about why the Fed behaved as it did. We direct interested readers to Hetzel (1998), Orphanides (2003), and Cogley and Sargent (2003) for an introduction to the vast literature analyzing that question.

Inflation Volatility

Panel A of Figure 3 displays two measures of inflation volatility. The first, the solid line, is the variance of inflation, measured over ten-year rolling windows ending at the date on the horizontal axis. For example, the entry labeled "4/79" is the sample variance of inflation from the third quarter of 1969 through the second quarter of 1979.

Variance is the most natural way to measure volatility. However, variance can be a misleading measure of volatility if a time series is serially correlated. For example, consider the first-order autoregressive process,

$$y_t = \rho y_{t-1} + \varepsilon_t, \tag{3}$$

where ε_t is an i.i.d. normal random variable with mean zero and variance v. The variance of y_t is $\text{var}(y) = \left(1 - \rho^2\right)^{-1} v$. Thus, even though v is the only source of random volatility in y, the autoregressive coefficient ρ contributes to the variance of y.

The effect of serial correlation (that is, persistence) on variance leads us to present a second measure of volatility along with variance. The dashed line is the variance of the residual in an autoregressive representation of inflation, where the autoregression is estimated by OLS, and the lag length is chosen by the Akaike information criterion (AIC). This residual variance can be thought of as a measure of the volatility that remains after taking out predictable variation in the series during the particular ten-year window. For both measures, volatility fell dramatically until 1961, then remained low until the early 1970s. It rose in the 1970s, fell in the 1980s, and has been historically low over the last five years. The fact that the variance of inflation rose much more than the shock variance from the late 1960s through the late 1980s suggests that there were changes in the serial correlation properties of inflation over this period. We consider these next.

Inflation Persistence

"Inflation persistence" refers to the degree to which a sudden change in the inflation rate tends to persist over time. As we just saw, persistence leads to higher variance, other things equal. In recent years much research has been devoted to estimating the persistence of inflation in the United States. This literature was spawned by Fuhrer and Moore (1995), who argued that inflation in the United States was characterized by high persistence and that models with forward-looking pricing behavior were unable to replicate the observed level of persistence. Fundamentally, however, interest in inflation persistence dates back to Lucas (1972) and Sargent (1971). These authors showed that the accuracy of econometric procedures for estimating Phillips curve slopes could be sensitive to the univariate persistence properties of inflation. Recent research on inflation persistence has, like Fuhrer and Moore, been concerned

with quantifying the degree of inflation persistence and then assessing whether and to what degree observed persistence is an inherent structural feature or an artifact of the particular monetary policy in place. The extent to which inflation persistence is structural has important implications for the consequences of alternative monetary policies.[4]

There are several ways to measure inflation persistence. Pivetta and Reis (2004) discuss the different measures in detail. In the case of the first-order autoregression discussed above, the different measures of persistence are all equivalent, and persistence is summarized by the parameter, ρ. For more complicated processes, the different measures can give different rankings of persistence. We will follow Levin and Piger (2003) and Clark (2003) in measuring inflation persistence by the sum of autoregressive coefficients in a univariate autoregressive representation of inflation.[5] If the sum of autoregressive coefficients is ρ, then $1/(1-\rho)$ represents the long-run effect of a permanent unit shock to the autoregression. That is, if in each period from $t = 0$ to ∞, the autoregression in (3) is hit by $\varepsilon_t = 1$, and $\varepsilon_t = 0$ for $t < 0$, then at $t = \infty$, we have $y_t = 1/(1-\rho)$.

Panel B of Figure 3 displays ten-year rolling-window estimates of PCE inflation persistence from the second quarter of 1959 to the first quarter of 2004. For each quarter, we take the ten years of prior data and estimate an autoregression for inflation, using the AIC to select lag length. The sum of autoregressive coefficients is then plotted in this panel, along with centered 90 percent confidence intervals constructed by semiparametric bootstrapping.[6]

Persistence fluctuates between 0.16 and 1.20 over the full sample. It was low until the late 1960s, then jumped up in late 1968 and early 1969, and remained high (roughly 0.8 or above) until 1999, apart from a brief period in 1983 and some rapid fluctuations between 1991 and 1995. In the last five years, our persistence measure has declined steadily, reaching 0.23 in the first quarter of 2004. The confidence intervals are quite wide. However, they encompass zero a much greater percentage of the time than they encompass unity, shedding some doubt on the conventional wisdom that inflation is inherently highly persistent.[7] The increase in inflation persistence in the late 1970s corresponds to the divergence (panel A of Figure 3) between the variance of inflation and the variance of the shock to the inflation autoregression. It is

[4] Different degrees of structural inflation persistence correspond to different degrees of price rigidity or other nominal frictions. Different specifications of nominal frictions, in turn, correspond to different real implications of changing policy rules.

[5] In the first-order example, the sum of coefficients is simply ρ.

[6] To generate the confidence intervals for a given quarter, we simulated 5000 samples by combining the estimated autoregressive coefficients with resampled residuals. These confidence intervals should be interpreted with caution; Hansen's (1999) grid bootstrap method deals more effectively with the bias associated with persistence being close to unity.

[7] This statement requires the caveat that the confidence intervals will be misleading when persistence is near unity.

not as easy to reconcile the joint behavior of these three objects later in the sample when the variance of inflation drops sharply. That is, the sharp drop in the variance of inflation without a sharp drop in the shock variance is not explained by a sharp drop in inflation persistence. Such a discrepancy can occur because, for autoregressions with more than one lag, the relationship between variance of the series and variance of the shock depends on the individual autoregressive coefficients, not just their sum.

2. SECTORAL INFLATION AND OVERALL INFLATION

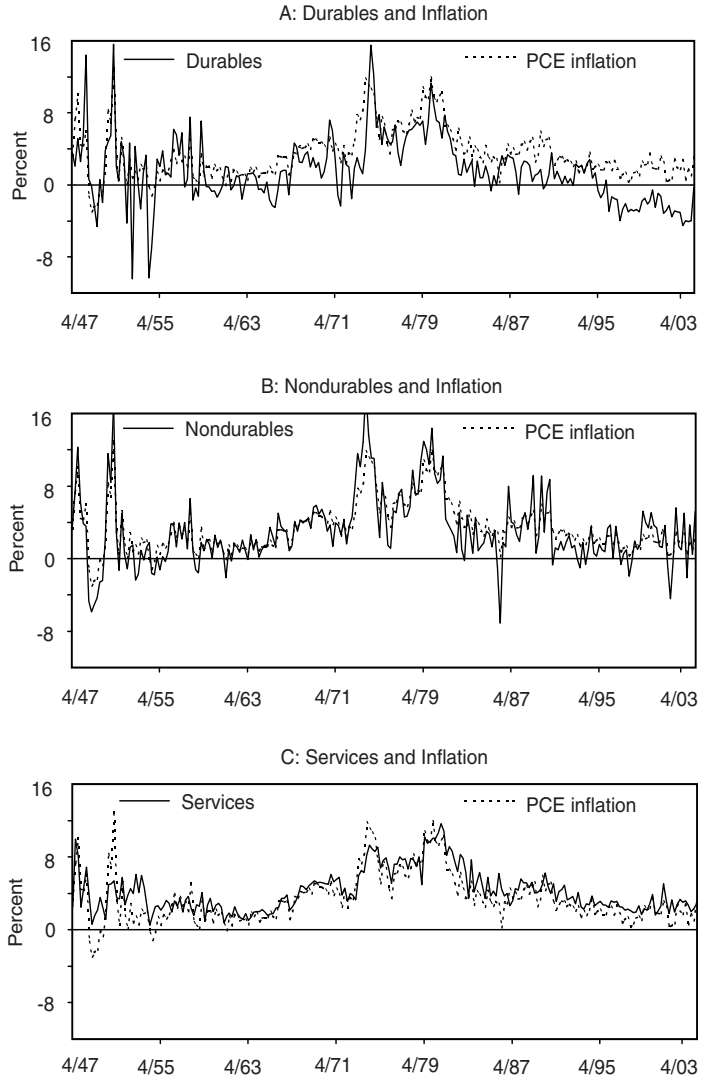
Having laid out the basic features of inflation behavior in the United States, we now turn to the components of inflation, expenditure shares, and price changes for the three consumer spending categories of durable goods, nondurable goods, and services. In this section we document the behavior of expenditure shares and price changes. The changes in expenditure shares over time and the variation in rates of price change across sectors then motivate our attempts in the next section to quantify the contribution of changing expenditure shares to the behavior of overall inflation.

Figure 1 plots expenditure shares for durable goods, nondurable goods, and services from 1947 to the present. Whereas the expenditure share for durable goods has fluctuated narrowly, between 12 and 18 percent, the shares of nondurables and services have respectively risen and fallen dramatically. In January 1947 services accounted for only 31 percent, and nondurable goods accounted for 56 percent of personal consumption expenditure. In January 2004, services accounted for 59 percent, and nondurable goods only 29 percent of personal consumption expenditure.

Figure 4 plots rates of price change for the three first-level components of personal consumption expenditure, together with the overall PCE inflation rate. Each series differs somewhat from overall inflation. Services price changes have generally been above PCE inflation, averaging 4.22 percent, compared to 3.42 percent for overall inflation. Durables price changes have generally been below PCE inflation, averaging 1.59 percent. The main distinguishing feature of nondurables price changes—which have averaged 3.09 percent—is that they have been more volatile than PCE inflation. This feature is reflected in Figure 5, which plots rolling-window variances of the sectoral rates of price change.[8] Figure 6 shows that the differences in rates of price

[8] Volatility of price changes of nondurables will not be a surprise to readers familiar with the concept of core PCE inflation. Core PCE inflation excludes food and energy prices, which are notoriously volatile and comprise a large share of nondurables expenditures. For short-run monetary policy purposes, core PCE inflation is generally preferred to overall PCE inflation.

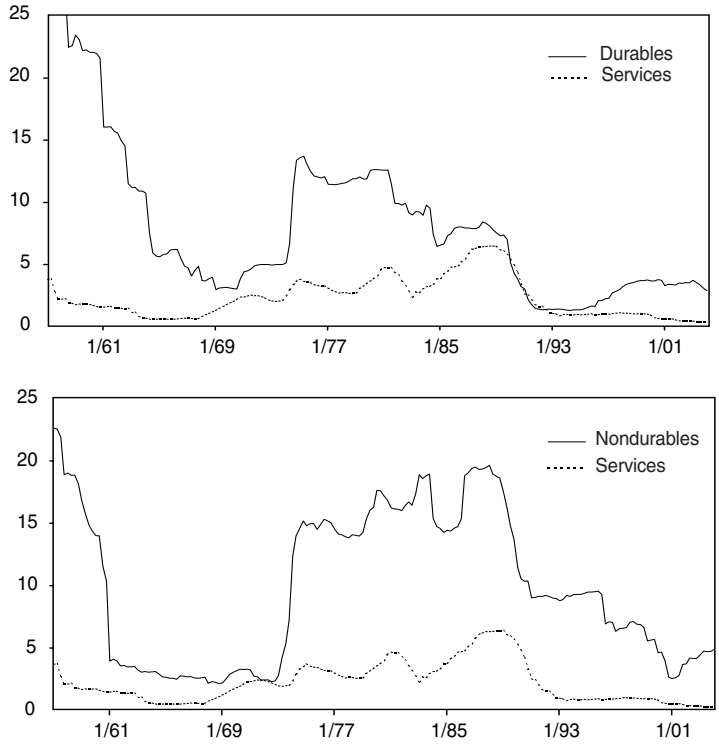
Figure 4 Sectoral Price Changes



change across sectors have cumulated significantly over time: the price index for services rose by a factor of eleven since 1947, whereas the price index for durables rose by less than a factor of three. In the last eight years, the price index for durable goods has actually been falling.

Figure 7 plots persistence for rates of price change of durables, nondurables, and services. The persistence measure is, again, the sum of auto-

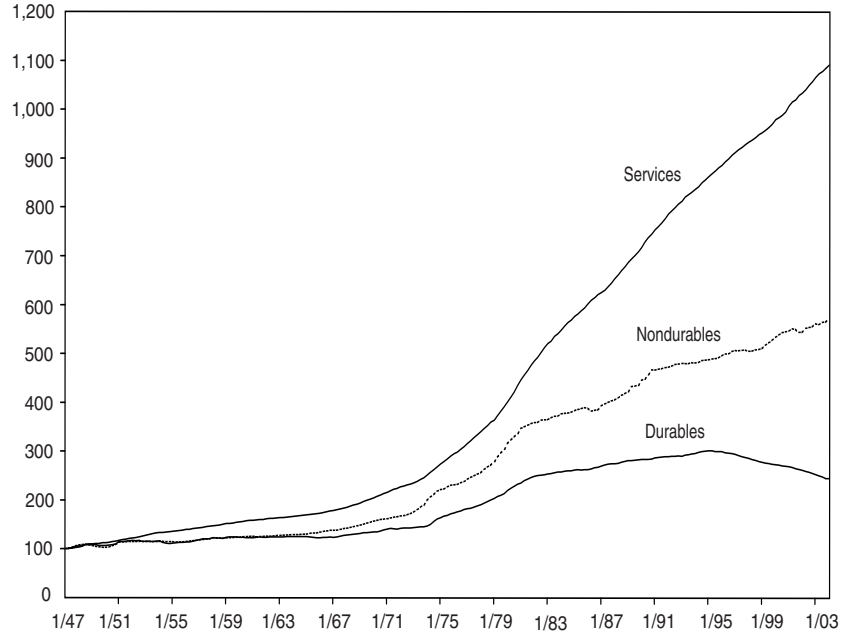
Figure 5 Variance of Sectoral Price Changes



regressive coefficients. The persistence measure moves broadly together across sectors, with services usually being the most persistent. Early in the sample, nondurables price changes are more persistent than durables price changes, but this ordering is reversed after about 1980. At the end of the sample, when persistence of PCE inflation is declining, the same is happening to rates of price change for services and nondurables, but persistence rises dramatically for durables price changes in 1998 and stays high until the present.

Together with the large swing in expenditure shares, differential behavior of price changes across sectors suggests that expenditure share changes may have been important contributors to the behavior of inflation. We will estimate this contribution in the next section. However, even if we find little contribution, the existence of expenditure shifts together with differing rates of price change across sectors is an important observation. Sectoral shifts and heterogeneous price behavior across sectors may have implications for

Figure 6 Sectoral Price Levels



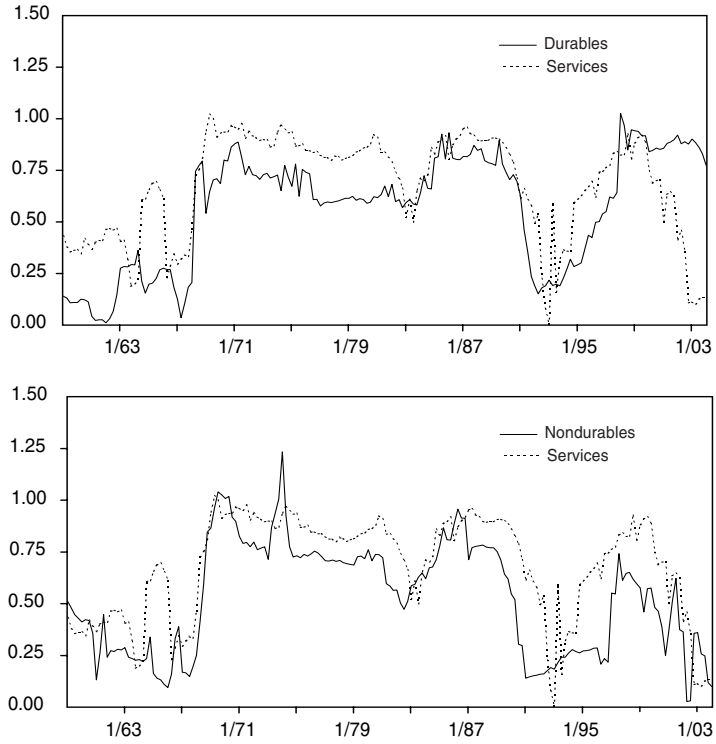
monetary policy. For example, the nature of optimal monetary policy may be sensitive to these factors.[9]

3. REINTERPRETING CHANGES IN THE BEHAVIOR OF INFLATION

To assess the importance of changing expenditure shares for the behavior of inflation, we construct two series that control for long-run shifts in expenditure shares. The first series we call "1947 inflation," and we create it by replacing the actual expenditure shares, $\omega_{i,t-1}$, in (2) with expenditure shares that fluctuate only transitorily around their 1947:4 levels. We generate 1947 inflation in two steps. First we estimate quadratic time trends for the expenditure shares under the restriction that the trends sum to one. Then we create a series of synthetic weights (expenditure shares) for each date in our sample by adding the 1947:4 value of the trend weight to the difference between the actual weight

[9] Aoki (2001), Erceg and Levin (2002), and Huang and Liu (2003) study cyclical fluctuations and monetary policy in multi-sector models. Wolman (2004) considers the optimal steady state inflation rate when there are relative price trends across sectors.

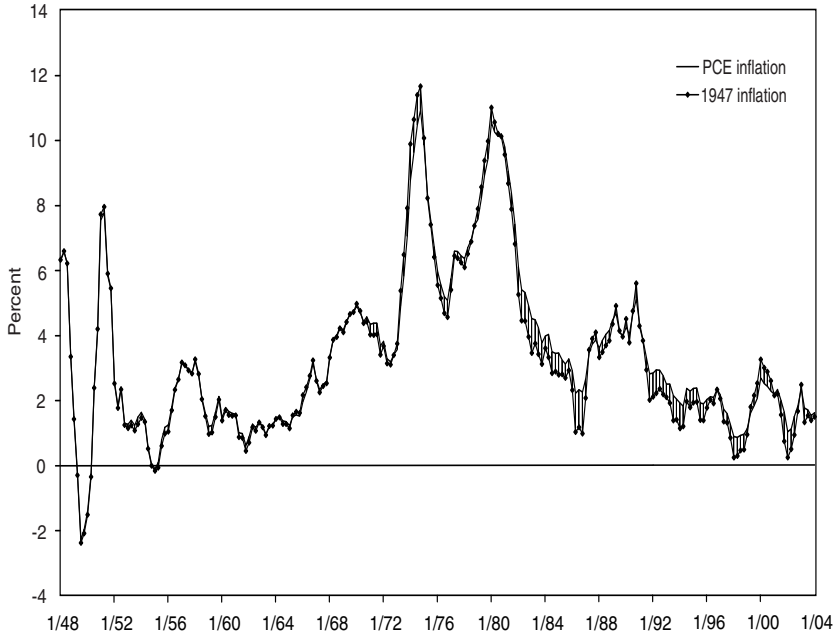
Figure 7 Persistence of Sectoral Price Changes



at each date and the trend weight estimated for that date. The initial values for the trend weights are 0.12 for durables, 0.31 for services, and 0.56 for nondurables. We allow for fluctuations around the trends because these may be independent of the long-run sectoral shifts we want to control for.

Our second approach to controlling for changing expenditure shares involves extracting the first principal component of the three sectoral rates of price change. The principal component is a weighted average of the three sectoral rates of price change, with the weights being chosen in order to maximize the variance of the weighted average. The weights are 0.76 for services, 0.21 for durables, and 0.03 for nondurables. The principal component can be viewed as the common component of the sectoral rates of price change. Because actual expenditure shares are not used to compute the principal component, they do not directly influence this series. Kapetanios (2002) suggests a similar measure as reflecting a notion of core inflation. The weighted median inflation measure emphasized by Bryan and Cecchetti (1994) is similar

Figure 8 1947 and PCE Inflation



in spirit to the first principal component in that it attempts to cut down the contribution of noisy components of inflation.[10]

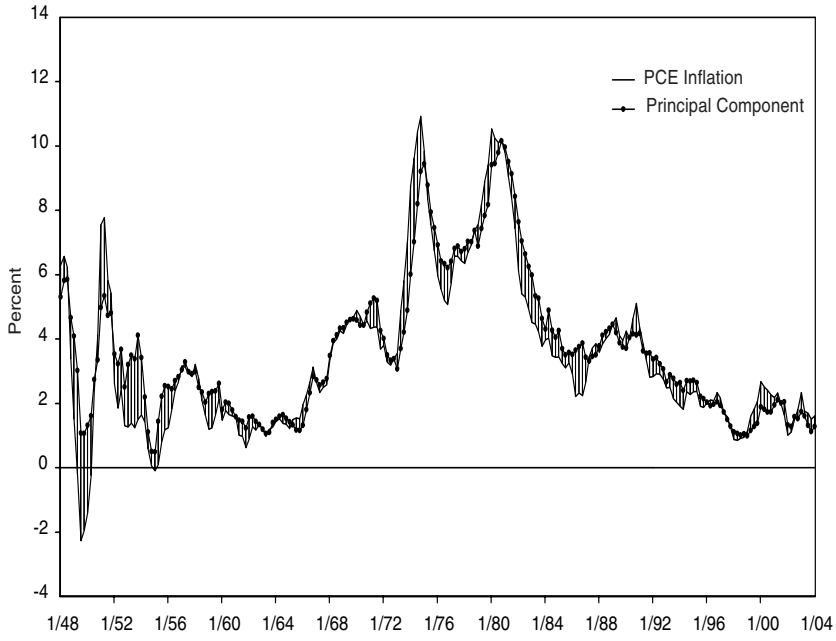
The Level of Inflation

Figure 8 displays the time series for 1947 inflation, and Figure 9 displays the first principal component of sectoral inflation. In each case we plot annual averages of the series and display them along with the corresponding series for actual PCE inflation. Both series share the broad patterns that characterize actual PCE inflation. If someone familiar with postwar U.S. inflation were shown either panel, it might not be difficult to convince them that it was a plot of actual inflation. However, there are some differences between both series and actual PCE inflation.

In the case of 1947 inflation, it is not surprising that these differences arise in the latter part of the sample, when the actual weights are quite different from the 1947 weights (services having risen and nondurables having fallen). Because nondurables inflation is more volatile than services inflation, the

[10] As a measure of core inflation, Bryan and Cecchetti (1994) use the weighted median of 36 components of the all-urban consumers CPI. This is the "central point, as implied by the CPI expenditure weights, in the cross-sectional histogram of inflation each month" (p. 203).

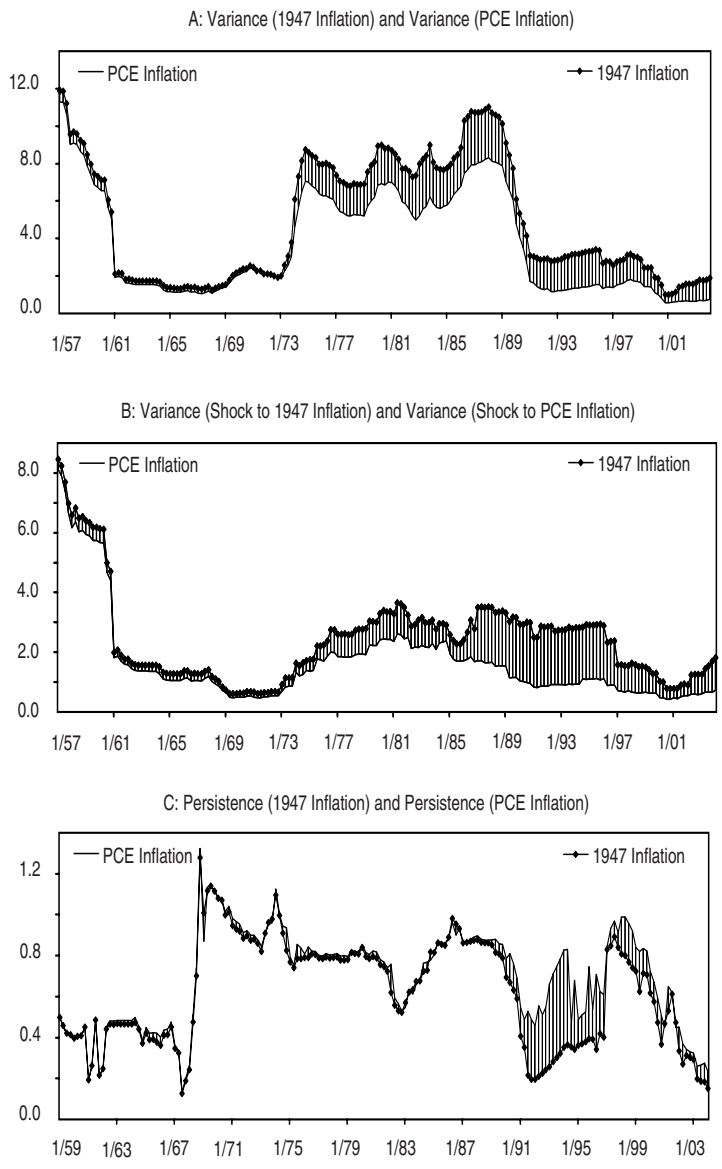
Figure 9 Principal Component and PCE Inflation



1947 inflation series with its higher weight on nondurables is noticeably more volatile than actual inflation in the last 20 years of the sample. In addition, because the average rate of price change for nondurables has been lower than that for services, 1947 inflation has a somewhat lower average level, 3.27 percent versus 3.42 percent. The lower level is obscured, however, by the higher volatility.

The principal component of sectoral price changes has a higher average than PCE inflation, at 3.64 percent. This is attributable to the high weight the principal component places on services. The high weight on services and low weight on volatile nondurables explains the fact that the principal component is less volatile than either PCE inflation or 1947 inflation. A notable feature of the principal component's behavior is that, unlike actual inflation and 1947 inflation, it is quite stable between 1983 and 1991. The other two series exhibit a sharp fall around 1986 and then a sharp rise followed by an additional steady increase. Referring to the sectoral price changes in Figure 4, we can understand this divergence as reflecting the fact that the volatility in the mid-to-late 1980s is largely accounted for by volatility in nondurables price changes.

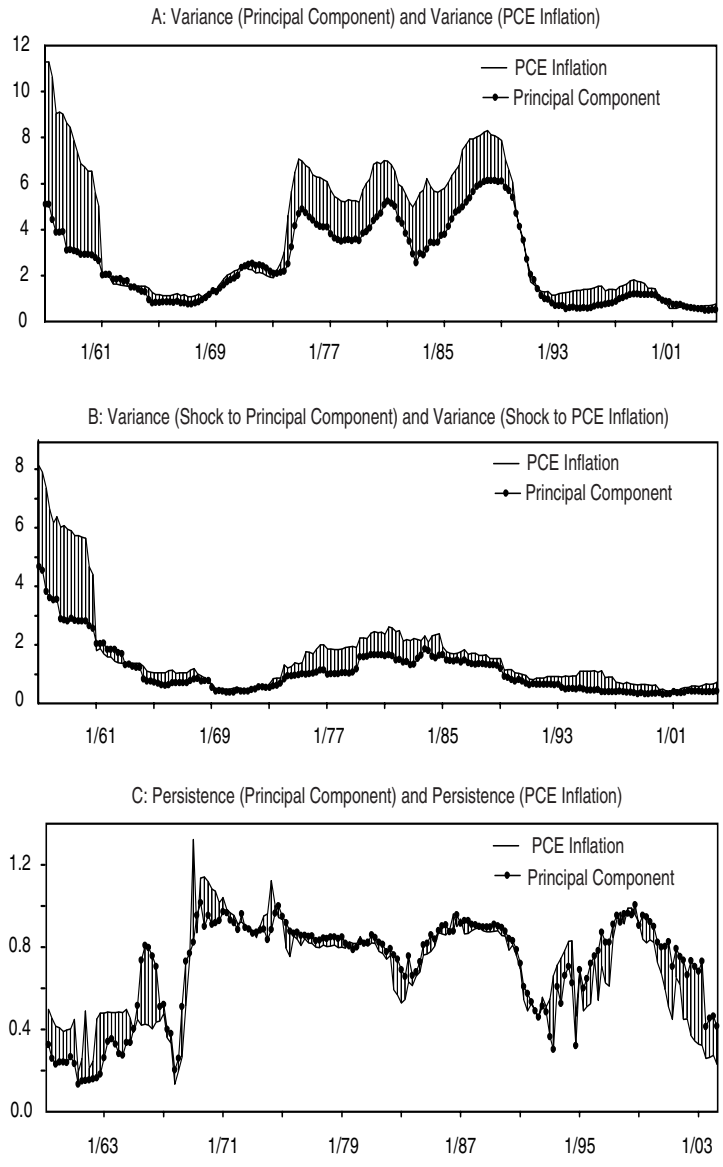
Figure 10 Volatility and Persistence of 1947 Inflation



Volatility and Persistence

Figures 10 and 11 display volatility and persistence of our alternative measures in the same way that Figure 3 displays volatility and persistence of actual inflation. Neither 1947 inflation nor the principal component of sectoral price

Figure 11 Volatility and Persistence of Principal Component



changes displays markedly different volatility patterns than does actual inflation. There are some minor differences across the series, however. Figures 10 and 11 confirm that the principal component has lower volatility than actual inflation or 1947 inflation. The inflation shock volatility displayed in the middle

panels behaves similarly for actual inflation and the principal component, declining smoothly from the mid-1970s until the early 1980s. In contrast, for 1947 inflation, there is a sharper decline in shock volatility, and it does not occur until the mid-1980s.

Rolling-window estimates of inflation persistence for the two new series are in the bottom panels of Figures 10 and 11. Over the first two-thirds of the sample, there is little difference between the persistence of 1947 inflation and the persistence of PCE inflation. This similarity is to be expected, because the underlying inflation series for the two figures do not differ much from each other. Since 1990, however, the two sets of estimates have diverged noticeably. For PCE inflation, persistence has been generally high over this period (with an average of 0.63), declining below 0.50 only in the last four years. In contrast, persistence of 1947 inflation has been generally *low* since 1990, averaging 0.45.

To some degree, the lower level of persistence in recent years for 1947 inflation is easy to explain. Nondurables has generally been the least persistent component of inflation (see Figure 7)—at least during the second half of the sample; therefore, because our 1947 inflation series places a relatively higher weight on nondurables later in the sample, this direct effect will make 1947 inflation more persistent than PCE inflation. However, this direct effect cannot explain all of the differences between the persistence of 1947 inflation and PCE inflation. The persistence of 1947 inflation is *not* simply the expenditure-share-weighted average of the persistence of the components. Our persistence measure has the flavor of a covariance, and, as such, it depends in a complicated manner on the covariance between sectoral rates of price change.

The bottom panel of Figure 11 plots the same rolling-window estimate of persistence for the principal component. Unlike 1947 inflation, the principal component places a very low weight on nondurables. Thus, it is not surprising that its persistence behaves quite differently than that of 1947 inflation. Although persistence of the principal component has declined in recent years, the decline has been smaller in magnitude than that of actual inflation; the relatively high weight on durables means that the increase in persistence of price changes of durables is reflected more in the principal component than in 1947 inflation. More generally, fluctuations in the persistence of the principal component have been smaller than fluctuations in the persistence of actual inflation or 1947 inflation.

4. CONCLUSION

We began by noting the dramatic changes in consumption expenditure shares that have occurred in the United States over the last 50 years. The fact that these shares serve as weights in consumption price inflation measures then led us to investigate the quantitative importance of shifts in expenditure shares for

the behavior of U.S. inflation. Using two different methods, we found that controlling for expenditure share changes led to a picture of U.S. inflation over the last 50 years that was somewhat—but not dramatically—different from the picture provided by actual PCE inflation. This analysis is exploratory only. That changing expenditure shares do not account for much of the behavior of inflation does not mean that those changes are inconsequential for monetary policy. Large changes in expenditure shares, together with trend changes in relative prices across sectors (as displayed in Figure 6) may interact with other differences across sectors in a way that has important implications for monetary policy. For example, if the nature of price stickiness differs systematically across sectors (as tentatively suggested by the work of Bils and Klenow [2004]) or if money demand varies systematically across expenditure types, then the monetary policy prescriptions from one-sector models may differ markedly from those in models with multiple categories of consumption.

REFERENCES

Aoki, Kosuke. 2001. "Optimal Monetary Policy Response to Relative Price Changes." Journal of Monetary Economics 48 (December): 55–80.

Bils, Mark, and Pete Klenow. 2004. "Some Evidence on the Importance of Sticky Prices." *Journal of Political Economy* 112 (October): 947–85.

Bauer, Andrew, Nicholas Haltom, and William Peterman. 2004. "Examining Contributions to Core Consumer Inflation Measures." Federal Reserve Bank of Atlanta Working Paper 04-07.

Bryan, Michael, and Stephen Cecchetti. 1994. "Measuring Core Inflation." *Monetary Policy*, ed. N. Gregory Mankiw. Chicago: University of Chicago Press: 195–215.

Clark, Todd E. 1999. "A Comparison of the CPI and the PCE Price Index." Federal Reserve Bank of Kansas City *Economic Review* 84 (Third Quarter): 15–29.

_____. 2003. "Disaggregate Evidence on the Persistence of Consumer Price Inflation." Federal Reserve Bank of Kansas City Research Working Paper 03-11.

Cogley, Timothy, and Thomas J. Sargent. 2003. "The Conquest of U.S. Inflation: Learning, Model Uncertainty, and Robustness." Manuscript.

Erceg, Christopher, and Andrew Levin. 2002. "Optimal Monetary Policy with Durable and Nondurable Goods." FRB International Finance

Discussion Paper 748 and ECB Working Paper 179.

Fuhrer, Jeffrey, and George Moore. 1995. "Inflation Persistence." *Quarterly Journal of Economics* 110 (February):127–59.

Hansen, Bruce E. 1999. "The Grid Bootstrap and the Autoregressive Model." *The Review of Economics and Statistics* 81 (November): 594–607.

Hetzel, Robert L. 1998. "Arthur Burns and Inflation." Federal Reserve Bank of Richmond *Economic Quarterly* 84 (Winter): 21–44.

Huang, Kevin, and Zheng Liu. 2003. "Inflation Targeting: What Inflation Rate to Target?" *Journal of Monetary Economics*. Forthcoming.

Kapetanios, George. 2002. "Modelling Core Inflation for the UK Using a New Dynamic Factor Estimation Method and a Large Disaggregated Price Index Dataset." Queen Mary, University of London, Department of Economics Working Paper 471.

Levin, Andrew, and Jeremy Piger. 2002. "Is Inflation Persistence Intrinsic in Industrial Economies?" Federal Reserve Bank of St. Louis Working Paper 23E.

Lucas, Robert E. 1972. "Econometric Testing of the Natural Rate Hypothesis." *The Econometrics of Price Determination*, ed. O. Eckstein. Washington, D.C.: Board of Governors of the Federal Reserve System: 50–9.

Orphanides, Athanasios. 2003. "The Quest for Prosperity Without Inflation." *Journal of Monetary Economics* 50 (April): 633–63.

Pivetta, Frederic, and Ricardo Reis. 2004. "The Persistence of Inflation in the United States." Manuscript.

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

Webb, Roy. 2004. "Which Price Index Should a Central Bank Employ?" Federal Reserve Bank of Richmond *Economic Quarterly* 90 (Spring): 63–76.

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

_____. 2004. "The Optimal Rate of Inflation with Trending Relative Prices." Manuscript.

On the Aggregate Labor Supply

Yongsung Chang and Sun-Bin Kim

Issues of labor supply are at the heart of macroeconomic studies of large cyclical fluctuations. The population puts forth more work effort in booms than in slumps. Economists' explanations of this phenomenon range from a pure market-clearing supply-and-demand view at one extreme to a dismissal of almost any role of supply and of market clearing at the other extreme. Disagreement is intense because labor markets' failure to clear may create a strong case favoring activist macroeconomic policy. According to the equilibrium business cycle models led by Lucas and Rapping (1969), people work more hours in some years than in others because the market rewards them for this pattern. Even in a non-equilibrium model in which the role of labor supply is dismissed in the short run, its slope is still important for the welfare cost of departing from the supply schedule. Labor supply elasticity is also crucial in evaluating the effect of taxes and government spending (e.g., Auerbach and Kotlikoff 1987; Judd 1987).

Figure 1 shows the cyclical components of total hours worked and wages for the U.S. economy for 1964:I–2003:II (detrended using the Hodrick-Prescott filter). Hours worked represent the total hours employed in the nonagricultural business sector. The wages are real hourly earnings of the production and nonsupervisory workers. Fluctuations of hours of work are much greater than those of wages.[1] If the intertemporal substitution hypothesis were to explain fluctuations in hours, it would require a labor supply elasticity beyond

We would like to thank Andreas Hornstein, Thomas Humphrey, Yash Mehra, and Pierre Sarte for their helpful comments. The views expressed herein are not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Moreover, wages are not strongly correlated with hours, casting further doubt on the intertemporal substitution mechanism. While the contemporaneous and dynamic correlations between hours and wages are important for business cycle analysis, we focus on the slope of the labor supply schedule only in this article. See Chang and Kim (2004b) on this issue.

the admissible estimates from the empirical micro studies, which are typically less than 0.5.[2]

In this article, we demonstrate both qualitatively and quantitatively how the slope of the aggregate labor supply schedule is determined by the reservation wage distribution, rather than by the willingness to substitute leisure intertemporally.[3] Based on our recent studies (Chang and Kim 2004a, 2004b), we present a fully specified general equilibrium model economy where the reservation wage distribution is nondegenerate. While the model is parsimonious, it provides a laboratory in which we can investigate the mapping from individual to aggregate labor supply functions. The model economy is populated by many workers who face uninsurable idiosyncratic productivity shocks—as demonstrated in Aigagari's (1994) incomplete capital market—and make decisions on the labor market participation—as demonstrated by Rogerson's (1985) study of indivisible labor. The cross-sectional distributions of earnings and wealth are comparable to those in the U.S. data. We find that the aggregate labor supply elasticity of such an economy is around one, even though the intertemporal substitution elasticity of leisure at the individual level is assumed to be 0.4. This aggregate elasticity is greater than the typical micro estimates but smaller than those often assumed in the aggregate models.

The article is organized as follows: Section 1 provides various models of aggregate labor supply based on individuals' work decisions. Section 2 presents illustrative examples that demonstrate how the aggregate labor supply depends on the reservation wage distribution. Section 3 lays out the model economy where the reservation wage distribution is dispersed. In Section 4, we calibrate the model parameters using various microdata and investigate the properties of aggregate labor supply of the model. Section 5 summarizes our findings.

1. LABOR SUPPLY: INDIVIDUAL VERSUS AGGREGATE

In this section, we consider various models on individuals' labor supply decisions and derive the corresponding aggregate labor supply schedules. For

[2] In his survey paper, Pencavel (1986) reports that most estimates are between 0.00 and 0.45 for men. In their parallel survey of research on the labor supply of women, Killingsworth and Heckman (1986) present a wide range of estimates, from -0.3 to 14.0; they do not venture a guess as to which is correct but conclude that the elasticity is probably somewhat higher for women than men. See Blundell and MaCurdy (1999) for a more recent review of the literature. An alternative (equilibrium) approach is to introduce shifts in labor supply through shifts in preference (Bencivenga 1992), home technology (Benhabib, Rogerson, and Wright 1991; Greenwood and Hercowitz 1991), or government spending (Christiano and Eichenbaum 1992).

[3] Hansen's (1985) indivisible labor economy, based on the theory of employment lotteries by Rogerson (1988), generates a very high aggregate labor supply elasticity—in fact, infinity—regardless of individual labor supply elasticity. However, the existence of employment lotteries is not strongly supported by the data, as the persons with greater hours or greater earnings per hour consume more. Our analysis illustrates that such an economy is a special case where the reservation wage distribution is degenerate.

the moment, we abstract from the intertemporal decisions. Hence, models in this section are static and of partial equilibrium. We will study a fully specified dynamic general equilibrium model in Section 3.

Homogeneous Agents with Divisible Labor

Suppose there is measure one of identical agents with the following preferences over consumption, c, and hours worked, h:

$$U = \max_{c,h} \; lnc - B\frac{h^{1+1/\gamma}}{1+1/\gamma} \tag{1}$$

subject to

$$c = wh + ra, \tag{2}$$

where w is the hourly wage; r, the interest rate; and a, asset holdings. The first order condition for hours of work is

$$Bh^{1/\gamma} = \frac{w}{c}. \tag{3}$$

The marginal disutility from additional hours of work equals the marginal utility of consumption from income earned. The labor supply function can be written as

$$h = \left(\frac{w}{Bc}\right)^{\gamma}. \tag{4}$$

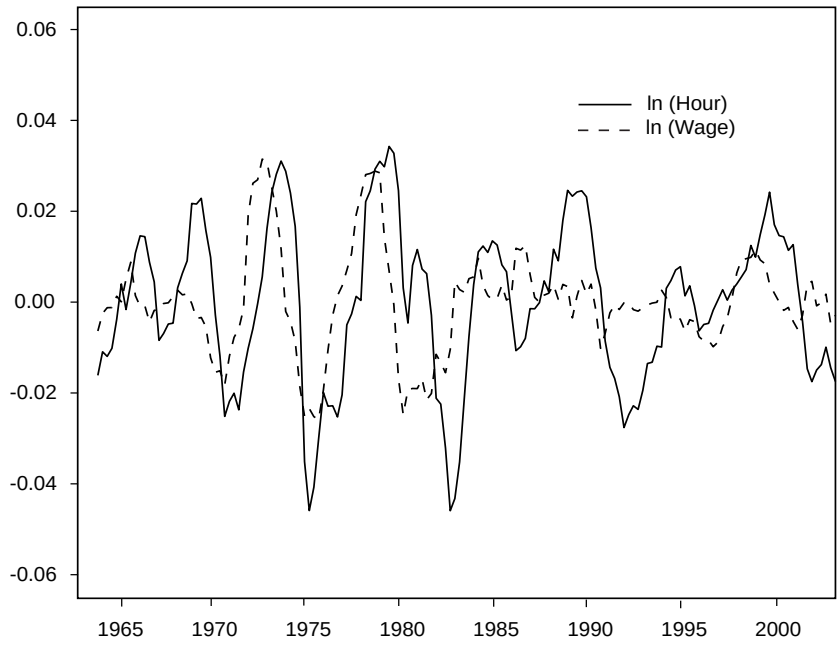
The Frisch elasticity—elasticity of hours with respect to wage holding wealth (consumption) constant—is γ. With homogeneous agents, the aggregate labor supply elasticity is also γ. According to the empirical micro studies, the labor supply is inelastic since a typical value of γ is less than 0.5. As Figure 1 illustrates, inelastic labor supply is hard to reconcile with the fact that hours fluctuate greatly without much variation in wages.

Homogeneous Agents with Indivisible Labor

A large fraction of cyclical fluctuations of total hours worked reflects the decisions to work or not (the so-called extensive margin), whereas the micro elasticities reflect the variation of hours for employed workers (intensive margin). The indivisible labor model has been developed to highlight the extensive margin of labor supply.[4] Suppose an agent supplies $\bar{h}$ hours if he works and zero hours otherwise. With homogeneous agents, the labor supply decision is

[4] In general, the labor supply decision operates on both the extensive and intensive margins. However, workers are rarely allowed to choose completely flexible work schedules or to supply a small number of hours.

Figure 1 Cyclical Components of Total Hours and Wages



Notes: Hours worked represent the nonagricultural private sector. Wage is real hourly earnings for nonsupervisory and production workers.

randomized (see Hansen 1985 and Rogerson 1988). An agent chooses probability of working, p, and the expected utility is $p(lnc - B\frac{\bar{h}^{1+1/\gamma}}{1+1/\gamma}) + (1 - p)(lnc - 0)$. Then the agent's maximization problem with the existence of a complete insurance market is

$$U = \max_{c,p} \; lnc - pB\frac{\bar{h}^{1+1/\gamma}}{1+1/\gamma},$$

subject to

$$c = wp\bar{h} + ra.$$

The equilibrium value of p is equal to the fraction of agents that work, and the aggregate labor supply is given by $H = p\bar{h}$. The aggregate labor supply elasticity is infinite, as the stand-in agent's utility is linear in p. While the aggregate labor supply is infinitely elastic in this environment, the underlying assumptions—homogeneity and complete market—are vulnerable even to casual empiricism.

Heterogeneous Agents with Indivisible Labor

Suppose workers differ in both preference (B) and asset holdings (a) and that the complete insurance of idiosyncratic risks is not available. With indivisible labor, the agent, i, works if

$$\log(w\bar{h} + ra_i) - B_i \frac{\bar{h}^{1+1/\gamma}}{1+1/\gamma} \geq \log(ra_i). \tag{5}$$

The reservation wage, $\widetilde{w}$, is

$$\widetilde{w} = \frac{ra_i}{\bar{h}}\Big(\exp(B_i\Delta) - 1\Big), \tag{6}$$

where $\Delta = \frac{\bar{h}^{1+1/\gamma}}{1+1/\gamma}$ is a constant, independent of individual characteristics. Workers with high B_i (those who value leisure more relative to commodity consumption) exhibit a higher reservation wage. The richer (a_i) a worker is, the higher his reservation wage. In general the reservation wage depends on various dimensions of cross-sectional heterogeneity. In Section 3, we investigate the fully specified dynamic general equilibrium model where the shape of $\Phi(\widetilde{w})$ is parsimoniously characterized by the microdata and depends on the agent's earnings ability as well as wealth.

2. ILLUSTRATIVE EXAMPLES

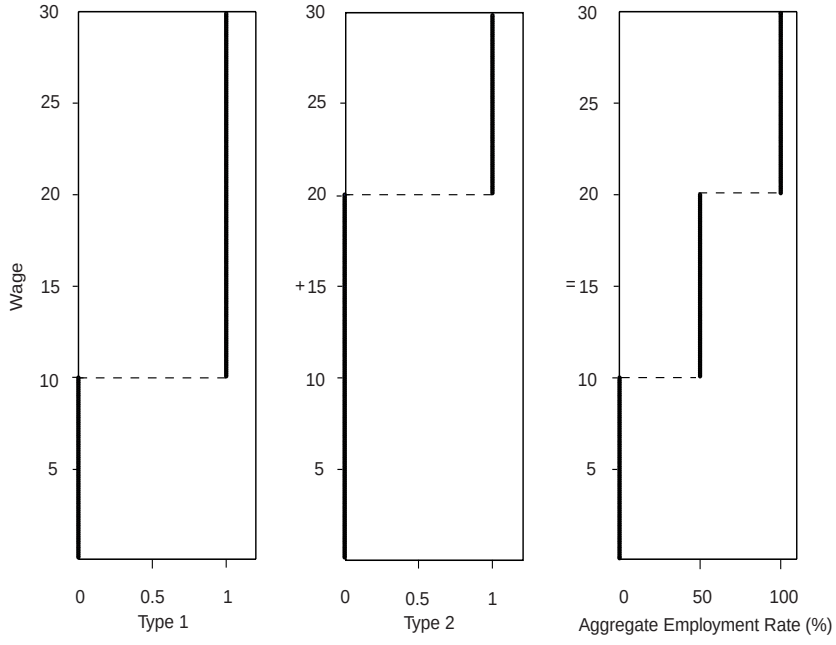
Before we present our formal analysis, we provide the examples that illustrate the dependence of aggregate labor supply on the shape of reservation wage distribution. Suppose that equal numbers of two types of workers exist in the economy, with reservation wages of \$10 and \$20, respectively. Suppose also that labor supply is indivisible in the sense that a worker supplies one unit of labor if he works. Figure 2 shows that the aggregate labor supply—the horizontal sum of individual labor supply—can have two elasticities. At a wage rate of \$10 and \$20, the elasticity is infinity. Otherwise, it is zero. Whenever a mass in the reservation wage distribution exists, the aggregate labor supply elasticity can take a large value. Suppose that many types of workers exist and that a worker works $\bar{h}$ hours if the market wage, w, exceeds the reservation wage, $\widetilde{w}$:

$$h(w) = \begin{cases} \bar{h} & \text{if } w \geq \widetilde{w}, \\ 0 & \text{otherwise.} \end{cases}$$

The aggregate labor supply function, $H(w)$, is

$$H(w) = \int_0^w \bar{h}\phi(\widetilde{w})d\widetilde{w} = \Phi(w)\bar{h}.$$

Figure 2 Individual and Aggregate Labor Supply



The aggregate labor supply elasticity, $\Gamma(w) = \frac{H'(w)w}{H(w)}$, is

$$\Gamma(w) = \frac{\Phi'(w)w}{\Phi(w)}.$$

The aggregate elasticity depends on the concentration of workers—the marginal density, $\Phi'(w)$, relative to the cumulative density, $\Phi(w)$. In the two-type workforce example, the aggregate elasticity is infinity where there is a mass of workers, ($\Phi'(10) = \Phi'(20) = \infty$), and zero elsewhere. In the lottery economy of Hansen (1985) and Rogerson (1988), the reservation wage distribution is degenerate (as the agents are identical) at the equilibrium wage rate ($\Phi'(w) = \infty$), and the aggregate elasticity becomes infinity.

The aggregate elasticity depends on the relative concentration of workers even when workers are allowed to work longer hours at higher wages. Suppose the labor supply of a worker is

$$h(w; \widetilde{w}) = \begin{cases} \bar{h} + \hat{h}(w; \widetilde{w}) & \text{if } w \geq \widetilde{w}, \\ 0 & \text{otherwise.} \end{cases}$$

Here $\hat{h}(w; \widetilde{w})$ is hours worked beyond the minimum hours, $\bar{h}$, and satisfies $\hat{h}(w; \widetilde{w}) \geq 0$ with equality when $w = \widetilde{w}$ and $\hat{h}'(w; \widetilde{w}) > 0$.[5] The aggregate labor supply function is

$$H(w) = \int_0^w h(w; \widetilde{w})\phi(\widetilde{w})d\widetilde{w} = \bar{h}\Phi(w) + \int_0^w \hat{h}(w; \widetilde{w})\phi(\widetilde{w})d\widetilde{w},$$

where the total hours worked consists of the sum of the extensive margins, $\bar{h}\Phi(w)$, and that of intensive margins, $\int_0^w \hat{h}(w; \widetilde{w})\phi(\widetilde{w})d\widetilde{w}$. Given that

$$H'(w) = h(w; w)\phi(w) + \int_0^w h'(w; \widetilde{w})\phi(\widetilde{w})d\widetilde{w},$$

the aggregate elasticity is

$$\Gamma(w) = \frac{[\bar{h}\phi(w) + \int_0^w h'(w; \widetilde{w})\phi(\widetilde{w})d\widetilde{w}]w}{\int_0^w h(w; \widetilde{w})\phi(\widetilde{w})d\widetilde{w}}.$$

For illustrative purposes, suppose the individual labor supply elasticity, γ, is constant across workers and wages: $\gamma = \dfrac{h'(w; \widetilde{w})w}{h(w; \widetilde{w})}$. Substituting $h'(w; \widetilde{w})$ with γ, the aggregate elasticity can be again expressed as the sum of the relative concentration of reservation wages and the individual elasticity:

$$\Gamma(w) = \frac{\bar{h}\Phi'(w)w}{\bar{h}\Phi(w) + \int_0^w \hat{h}(w; \widetilde{w})\phi(\widetilde{w})d\widetilde{w}} + \gamma.$$

These examples illustrate two important aspects of aggregate labor supply: the aggregate elasticity can be different from that of microelasticity and the aggregate labor supply elasticity is not time-invariant because the reservation wage distribution evolves over time as the wealth distribution and the level of employment change over time. However, these examples are silent about the magnitude of the aggregate labor supply elasticity for which the exact shape of the empirical reservation wage distribution must be uncovered. In the next section, we present a model economy—a simplified version of Chang and Kim (2004a)—where the reservation wage distribution, $\Phi(\widetilde{w})$, is determined by the asset accumulation of households that face different types of uninsurable income risks.[6] While the model is parsimonious, it allows for a complete characterization of the reservation wage distribution.

3. A FULLY SPECIFIED GENERAL EQUILIBRIUM MODEL

There is a continuum (measure one) of workers who have identical preferences but different productivity. Individual productivity varies exogenously

[5] The minimum-hours restriction can be easily justified, for example, by fixed costs, such as commuting time.

[6] In Chang and Kim (2004a), the economy consists of many households made up of a husband and wife. Here we present a model that is populated by many single-agent households.

according to a stochastic process with a transition probability distribution function, $\pi_x(x'|x) = Pr(x_{t+1} \leq x'|x_t = x)$. A worker maximizes his utility over consumption, c_t, and hours worked, h_t:

$$U = \max_{\{c_t, h_t\}_{t=0}^{\infty}} E_0 \left\{ \sum_{t=0}^{\infty} \beta^t u(c_t, h_t) \right\},$$

with

$$u(c_t, h_t) = lnc_t - B\frac{h_t^{1+1/\gamma}}{1+1/\gamma},$$

subject to

$$a_{t+1} = w_t x_t h_t + (1 + r_t)a_t - c_t. \tag{7}$$

Workers trade claims for physical capital, a_t, which yields the rate of return, r_t, and depreciates at the rate, δ. The capital market is incomplete. Physical capital is the only asset available to workers who face a borrowing constraint, $a_t \geq \bar{a}$ for all t. We abstract from the intensive margin and assume that the labor supply is indivisible. If employed, a worker supplies $\bar{h}$ units of labor and earns $w_t x_t \bar{h}$, where w_t is wage rate per effective unit of labor.

The representative firm produces output according to a Cobb-Douglas technology in capital, K_t, and efficiency units of labor, L_t:[7]

$$Y_t = F(L_t, K_t, \lambda_t) = \lambda_t L_t^{\alpha} K_t^{1-\alpha},$$

where λ_t is the aggregate productivity shock with a transition probability distribution function, $\pi_\lambda(\lambda'|\lambda) = Pr(\lambda_{t+1} \leq \lambda'|\lambda_t = \lambda)$.[8]

The value function for an employed worker, denoted by V^E, is

$$V^E(a, x; \lambda, \mu) = \max_{a' \in \mathcal{A}} \left\{ lnc - B\frac{\bar{h}^{1+1/\gamma}}{1+1/\gamma} + \beta E\Big[\max \left\{ V^E(a', x'; \lambda', \mu'), V^N(a', x'; \lambda', \mu') \right\} \Big| x, \lambda \Big] \right\},$$

subject to

$$c = wx\bar{h} + (1 + r)a - a',$$

[7] This production function implicitly assumes that workers are perfect substitutes for each other. While this assumption abstracts from reality, it greatly simplifies the labor market equilibrium.

[8] In this model economy, the technology shock is the only aggregate shock. This restriction does not necessarily reflect our view on the source of the business cycles. As we would like to show that the preference residual contains a significant specification error rather than true shifts in preferences, we intentionally exclude shocks that may shift the labor supply schedule itself (e.g., shifts in government spending or changes in the income tax rate) from the present article.

$$a' \geq \bar{a}, \text{ and}$$

$$\mu' = \mathbf{T}(\lambda, \mu),$$

where $\mathbf{T}$ denotes a transition operator that defines the law of motion for the distribution of workers, $\mu(a, x)$.[9] The value function for a nonemployed worker, denoted by $V^N(a, x; \lambda, \mu)$, is defined similarly with $h = 0$. Then, the labor supply decision is characterized by

$$V(a, x; \lambda, \mu) = \max_{h \in \{0, \bar{h}\}} \left\{ V^E(a, x; \lambda, \mu), V^N(a, x; \lambda, \mu) \right\}.$$

Equilibrium consists of a set of value functions, $\{V^E(a, x; \lambda, \mu), V^N(a, x; \lambda, \mu), V(a, x; \lambda, \mu)\}$; a set of decision rules for consumption, asset holdings, and labor supply, $\{c(a, x; \lambda, \mu), a'(a, x; \lambda, \mu), h(a, x; \lambda, \mu)\}$; aggregate inputs, $\{K(\lambda, \mu), L(\lambda, \mu)\}$; factor prices, $\{w(\lambda, \mu), r(\lambda, \mu)\}$; and a law of motion for the distribution $\mu' = \mathbf{T}(\lambda, \mu)$ such that:

1. Individuals optimize:

   Given $w(\lambda, \mu)$ and $r(\lambda, \mu)$, the individual decision rules—$c(a, x; \lambda, \mu)$, $a'(a, x; \lambda, \mu)$, and $h(a, x; \lambda, \mu)$—solve $V^E(a, x; \lambda, \mu)$, $V^N(a, x; \lambda, \mu)$, and $V(a, x; \lambda, \mu)$.

2. The representative firm maximizes profits:

   $$w(\lambda, \mu) = F_1\big(L(\lambda, \mu), K(\lambda, \mu), \lambda\big), \text{ and}$$

   $$r(\lambda, \mu) = F_2\big(L(\lambda, \mu), K(\lambda, \mu), \lambda\big) - \delta$$

   for all (λ, μ).

3. The goods market clears:

   $$\int \left\{a'(a, x; \lambda, \mu) + c(a, x; \lambda, \mu)\right\} d\mu = F\big(L(\lambda, \mu), K(\lambda, \mu), \lambda\big) + (1 - \delta)K$$

   for all (λ, μ).

4. Factor markets clear:

   $$L(\lambda, \mu) = \int x h(a, x; \lambda, \mu) d\mu, \text{ and}$$

   $$K(\lambda, \mu) = \int a d\mu$$

[9] Let $\mathcal{A}$ and $\mathcal{X}$ denote sets of all possible realizations of a and x, respectively. The measure $\mu(a, x)$ is defined over a σ-algebra of $\mathcal{A} \times \mathcal{X}$.

Table 1 Parameters of the Benchmark Model Economy

Parameter	Description
$\alpha = 0.64$	Labor share in production function
$\beta = 0.9785504$	Discount factor
$\gamma = 0.4$	Individual labor supply elasticity with divisible labor
$B = 151.28$	Utility parameter
$\bar{h} = 1/3$	Labor supply if working
$\bar{a} = -2.0$	Borrowing constraint
$\rho_x = 0.939$	Persistence of idiosyncratic productivity shock
$\sigma_x = 0.287$	Standard deviation of innovation to idiosyncratic productivity
$\rho_\lambda = 0.95$	Persistence of aggregate productivity shock
$\sigma_\lambda = 0.007$	Standard deviation of innovation to aggregate productivity

for all (λ, μ).

5. Individual and aggregate behaviors are consistent:

$$\mu'(A^0, X^0) = \int_{A^0, X^0} \left\{ \int_{\mathcal{A},\mathcal{X}} 1_{a'=a'(a,x;\lambda,\mu)} \, d\pi_x(x'|x) d\mu \right\} da' dx'$$

for all $A^0 \subset \mathcal{A}$ and $X^0 \subset \mathcal{X}$.

4. QUANTITATIVE ANALYSIS

Calibration

We briefly explain the choice of the model parameters. The unit of time is a business quarter. We assume that x follows an AR(1) process: $lnx' = \rho_x lnx + \varepsilon_x$, where $\varepsilon_x \sim N(0, \sigma_x^2)$. As we view x as reflecting a broad measure of earnings ability in the market, we estimate the stochastic process of x based on the wages from the Panel Study of Income Dynamics (PSID) for 1979–1992. The values of $\rho_x = 0.939$ and $\sigma_x = 0.287$ reflect the persistence and standard deviation of innovations to individual wages.[10] The other parameters of the article are in accordance with the business cycle analysis and empirical labor supply literature. A working individual spends one-third of her discretionary time: $\bar{h} = 1/3$. The individual compensated labor supply elasticity of hours, γ, is 0.4. The labor share of output, α, is 0.64, and the depreciation rate, δ, is 2.5 percent. We search for the weight parameter on leisure, B, such

[10] These are maximum-likelihood estimates of Heckman (1979), correcting for a sample selection bias. Our estimate for income shocks does not purge the life-cycle effect. In our companion paper, Chang and Kim (2004a), we use both cases. When the life-cycle effect (accounted for by observed characteristics such as age, education, and sex) is purged, the aggregate labor supply elasticity becomes slightly bigger because the reservation wage distribution becomes less dispersed.

Table 2 Characteristics of Wealth Distribution

	Quintile					
	1st	2nd	3rd	4th	5th	Total
PSID						
Share of wealth	-0.52	0.50	5.06	18.74	76.22	100
Group average/population average	-0.02	0.03	0.25	0.93	3.81	1
Share of earnings	-7.51	11.31	18.72	24.21	38.23	100
Model						
Share of wealth	-2.05	2.46	10.22	23.88	65.49	100
Group average/population average	-0.10	0.12	0.51	1.19	3.27	1
Share of earnings	9.70	15.06	19.01	23.59	32.63	100

Notes: The PSID statistics reflect the family wealth and earnings levels published in their 1984 survey.

that the steady state employment rate is 60 percent, the current population survey average for 1967:II–2000:IV. The discount factor, β, is chosen so that the quarterly rate of return to capital is 1 percent. The aggregate productivity shock, λ_t, follows an AR(1) process: $ln\lambda' = \rho_\lambda ln\lambda + \varepsilon_\lambda$, where $\varepsilon_\lambda \sim N(0, \sigma_\lambda^2)$. We set ρ_λ equal to 0.95 and σ_λ equal to 0.007, following Kydland and Prescott (1982). Table 1 summarizes the parameter values of the benchmark economy.

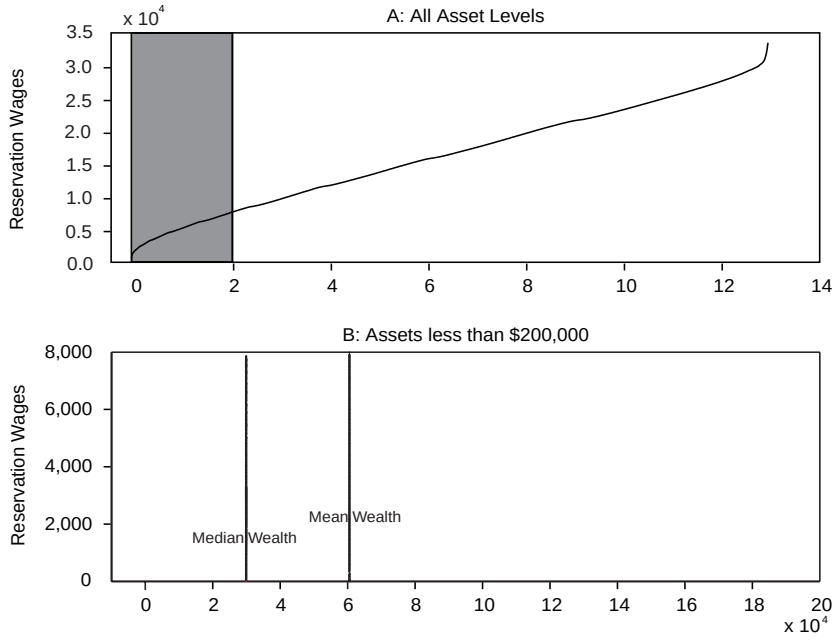
Cross-Sectional Earnings and Wealth Distribution

As we investigate the aggregation issue, it is desirable for the model economy to possess a reasonable amount of heterogeneity. We compare cross-sectional earnings and wealth—two important observable dimensions of heterogeneity in the labor market—found in the model and in the data.

Table 2 summarizes both the PSID and the model's detailed information on wealth and earnings. Family wealth in the PSID (1984 survey) reflects the net worth of houses, other real estate, vehicles, farms and businesses owned, stocks, bonds, cash accounts, and other assets. For each quintile group of wealth distribution, we calculate the wealth share, ratio of group average to economy-wide average, and the earnings share.

In both the data and the model, the poorest 20 percent of families in terms of wealth distribution were found to own virtually nothing. In fact, households in the first quintile of wealth distribution were found to be in debt in both the model and the data. The PSID found that households in the fourth and fifth quintile own 18.74 and 76.22 percent of total wealth, respectively, while, according to the model, they own 23.88 and 65.49 percent, respectively. The average wealth of those in the fourth and fifth quintile is, respectively, 0.93 and 3.81 times larger than that of a typical household, according to the

Figure 3 Reservation Wage Schedule



Notes: The graphs denote the reservation wage schedule of the benchmark model. Wages (quarterly earnings) and assets are in 1983 dollars.

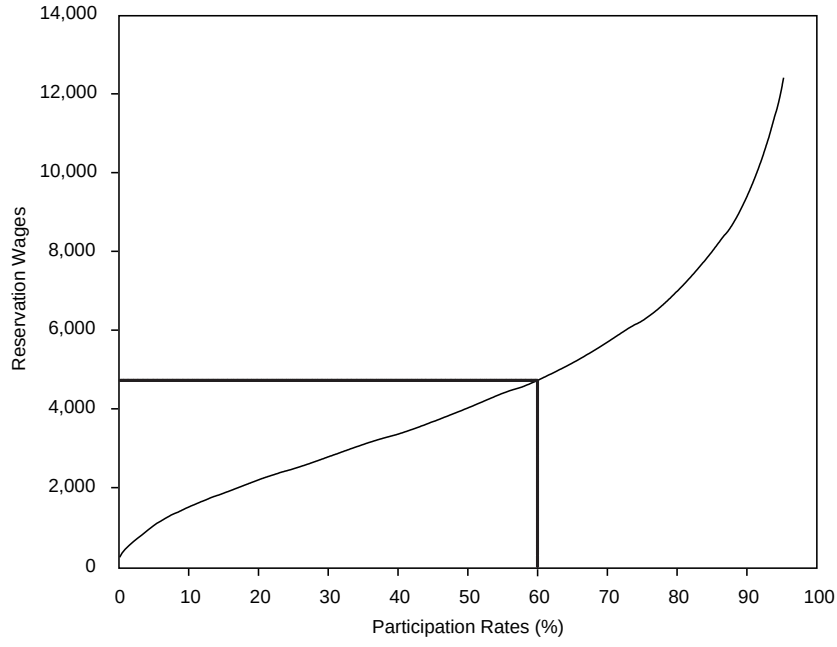
PSID. These ratios are 1.19 and 3.27 according to our model. The fourth and fifth quintile groups of the wealth distribution earn, respectively, 24.21 and 38.23 percent of total earnings, according to the PSID. The corresponding groups earn 23.59 and 32.63 percent, respectively, in the model.

Overall, the wealth distribution is found to be more skewed in the data. In particular, our model fails to match the highly concentrated wealth found in the right tail of the distribution. In the PSID, the top 5 percent of the population controls about half of total wealth (not shown in Table 2), whereas, in our model, they possess only 20 percent of total wealth. Since our primary objective is not to explain the top 1 to 5 percent of the population, we argue that the model economy presented in this article possesses a reasonable degree of heterogeneity, thus making it possible to study the effects of aggregation in the labor market.

Reservation Wage Distribution

The reservation wage distribution is crucial for the mapping from individual to aggregate labor supply. In Figure 3, we plot the reservation wage schedule

Figure 4 Reservation Wages and Participation Rates



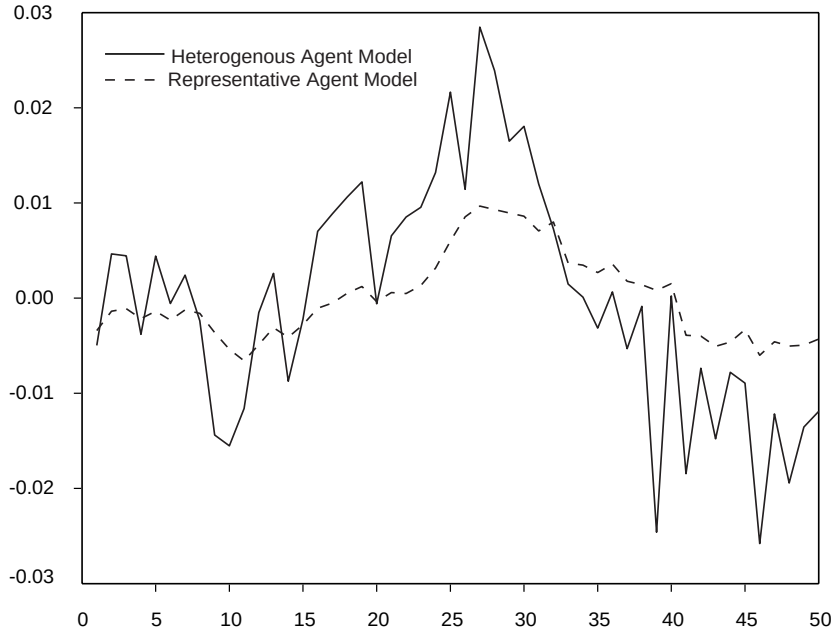
Notes: The graph denotes the inverse cumulative distribution functions of reservation wages. Wages are quarterly earnings in 1983 dollars.

of the benchmark model for all asset levels (panel A) and for assets less than $200,000 (panel B). At a given asset level, workers with wages (productivity) above the line choose to work. The reservation wage increases as the asset level increases. To illustrate, we adjust the units so that the mean asset of the model matches the average asset in the 1984 PSID survey, $60,524; thus, the values are in 1983 dollars.[11] Consider a worker whose assets are $29,880, the median of the wealth distribution from the model. According to the model, he is indifferent between working and not working at quarterly earnings of $3,287. Another worker whose assets are equivalent to the average asset holding of the economy, $60,524 (which belongs to the 66th percentile of the wealth distribution in our model and to the 72nd percentile in the PSID), is indifferent about working at $4,273 per quarter.

In Figure 4 we plot the inverse cumulative distribution of reservation wages of the model. In practice, the reservation wage distribution is neither observed nor constant over time. Based on the reservation wage schedule and

[11] The mean asset in our model is 14.48 units. The reservation wages in the vertical axis reflect quarterly earnings (the reservation wage rate multiplied by $\bar{h}$).

Figure 5 Total Hours Worked from the Models



invariant distribution, $\mu(x, a)$, we can infer the responsiveness of labor market participation. In Table 3 we compute the elasticities of participation with respect to the reservation wage around the steady state. These values may be viewed as the aggregate labor supply elasticity with zero wealth effect as they assume the *entire* wealth distribution held is constant. For the model economy, the elasticities are 1.12, 1.05, and 0.97, respectively, at the employment rates of 58, 60, and 62 percent. Overall, these values are bigger than typical micro estimates, but they remain in a moderate range. In particular, a very high elasticity—in fact, infinity—generated by a lottery economy with a homogeneous workforce (in which the reservation wage distribution is degenerate) does not survive serious heterogeneity.

Finally, we would like to emphasize that, when labor supply is indivisible, the slope of the aggregate labor supply schedule is mostly determined by the distribution of reservation wages rather than by the willingness to substitute leisure intertemporally. In fact, the aggregate labor supply is independent of γ in our economy. With a binary choice of hours, utility of market participants and non-participants differs by a constant term, $B\frac{\bar{h}^{1+1/\gamma}}{1+1/\gamma}$ [Recall (6)]. Given γ, we adjust B (the weight parameter on disutility from working) to match the 60 percent employment rate in the steady state, leaving the above constant

Table 3 Labor Supply Elasticity Implied by the Reservation Wage Distribution

Employment Rate		
$E = 58\%$	$E = 60\%$	$E = 62\%$
1.12	1.05	0.97

Notes: The numbers reflect the elasticity of the labor market participation rate with respect to reservation wage (evaluated at employment rates of 58, 60, and 62 percent) based on the reservation wage distribution in the steady state.

term unchanged. As a result, the steady state reservation-wage distribution remains the same regardless of γ.

Comparison with the Representative Agent Model

We compare the volatility of hours from our model economy to that of the representative agent economy. Both model economies will be subject to identical stochastic aggregate productivity shocks that resemble that of the post-war total factor productivity (Solow residual).

The value function of the representative agent, $V^R(K, \lambda)$, is

$$V^R(K, \lambda) = \max_{C,H} \left\{ lnC - B\frac{H^{1+1/\gamma}}{1+1/\gamma} + \beta E\left[V^R(K', \lambda')\middle|\lambda\right] \right\},$$

subject to

$$K' = F(K, H, z) + (1-\delta)K - C.$$

Except for β, the same parameter values are used, $\beta = 0.99$.[12] Fluctuation of the heterogeneous agent model is solved by the method developed by Krusell and Smith (1998). Figure 5 shows the sample paths of total hours worked (percentage deviations from the steady states), respectively, from the heterogeneous agent economy and the representative agent economy, $\gamma = 0.4$. In the face of aggregate productivity shocks whose stochastic process resembles that of the post-war total factor productivity, hours of work from the heterogeneous agent economy exhibit a much greater volatility than those of the representative agent model.

[12] B is a free parameter in a sense that it does not affect the dynamics around the steady state.

5. SUMMARY

We demonstrate that, at the aggregate level, the labor supply elasticity can significantly depart from the microelasticity. In an economy where households make decisions on labor market participation, the slope of the aggregate labor supply curve is determined by the distribution of reservation wages rather than by the willingness to substitute leisure intertemporally. We present a model economy where households face uninsurable idiosyncratic income shocks. While the model is parsimonious, the cross-sectional distributions of earnings and wealth are comparable to those in the U.S. data. We find that the aggregate labor supply elasticity of such an economy is around 1.0—despite the low intertemporal substitution elasticity of leisure, assumed to be 0.4. The equilibrium approach of business cycle analysis has been criticized on the grounds that it requires an elasticity higher than the intertemporal substitution elasticity estimated from the microdata. Our analysis shows that, while the aggregate labor elasticity can depart from a microelasticity, it remains in a moderate range as the reservation wage distribution is dispersed.

REFERENCES

Aiyagari, Rao S. 1994. "Uninsured Idiosyncratic Risk and Aggregate Savings." *Quarterly Journal of Economics* CIX: 659–83.

Auerbach, A. J., and L. J. Kotlikoff. 1987. *Dynamic Fiscal Policy.* Cambridge, U.K.: Cambridge University Press.

Bencivenga, Valerie. 1992. "An Econometric Study of Hours and Output Variation with Preference Shocks." *International Economic Review* (33): 448–71.

Benhabib, Jess, Richard Rogerson, and Randall Wright. 1991. "Homework in Macroeconomics: Household Production and Aggregate Fluctuations." *Journal of Political Economy* 99: 1166–87.

Blundell, Richard, and Thomas MaCurdy. 1999. "Labor Supply: A Review of Alternative Approaches." In *Handbook of Labor Economics*, Vol. 3A. Ed. O. Ashenfelter and D. Card. Amsterdam: North Holland: 1599–695.

Chang, Yongsung, and Sun-Bin Kim. 2004a. "From Individual to Aggregate Labor Supply: A Quantitative Analysis Based on a Heterogeneous Agent Macroeconomy." Manuscript, Federal Reserve Bank of Richmond.

_____. 2004b. "Heterogeneity and Aggregation in the Labor Market: Implications for Aggregate Preference Shocks." Manuscript,

Federal Reserve Bank of Richmond.

Christiano, Lawrence J., and Martin Eichenbaum. 1992. "Current Real-Business Cycle Theories and Aggregate Labor-Market Fluctuations." *American Economic Review* 82: 430–50.

Greenwood, Jeremy, and Zvi Hercowitz. 1991. "The Allocation of Capital and Time over the Business Cycle." *Journal of Political Economy* 99: 1188–215.

Hansen, Gary D. 1985. "Indivisible Labor and the Business Cycle." *Journal of Monetary Economics* 16: 309–27.

Heckman, James. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–62.

Judd, Kenneth L. 1987. "The Welfare Cost of Factor Taxation in a Perfect Foresight Model." *Journal of Political Economy* 95(4): 675–709.

Killingsworth, Mark R., and James Heckman. 1986. "Female Labor Supply." *Handbook of Labor Economics,* Vol. 1. Ed. O. Ashenfelter and R. Layards. Amsterdam: North Holland: 103–204.

Krusell, Per, and Anthony Smith. 1998. "Income and Wealth Heterogeneity in the Macroeconomy." *Journal of Political Economy* 106: 867–96.

Kydland, Finn E., and Edward Prescott. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica* 50: 1345–70.

Lucas, Robert E. Jr., and Leonard Rapping. 1969. "Real Wages, Employment, and Inflation." *Journal of Political Economy* 77: 721–54.

Pencavel, John. 1986. "Labor Supply of Men." *Handbook of Labor Economics,* Vol. 1. Ed. O. Ashenfelter and R. Layards. Amsterdam: North Holland: 3–102.

Rogerson, Richard. 1988. "Indivisible Labor, Lotteries and Equilibrium." *Journal of Monetary Economics* 21: 3–16.

Depression-Era Bank Failures: The Great Contagion or the Great Shakeout?

John R. Walter

Deposit insurance was created, at least in part, to prevent unfounded bank failures caused by contagion. The legislation that created the Federal Deposit Insurance Corporation (FDIC) was driven by the widespread bank failures of the Great Depression. In the years immediately before the 1934, when the FDIC began insuring bank deposits, over one-third of all extant banks failed. Many observers argue that these failures occurred because the banking industry is inherently fragile since it is subject to contagion-induced runs. Fragility arises because banks gather a large portion of their funding through the issuance of liabilities that are redeemable on demand at par, while investing in illiquid assets. Specifically, loans, which on average account for 56 percent of bank assets, tend to be made based on information that is costly to convey to outsiders. As a result, if a significant segment of bank customers run, that is, quickly require the repayment of their deposits, the bank is unlikely to be able to sell its assets except at a steep discount. Bank failure can result.

But do Depression-era bank failures imply the need for government-provided deposit insurance, or is there another explanation of the failures other than contagion and inherent fragility? Some observers question the view that banks are inherently fragile. They argue instead that the banking industry developed various market-based means of addressing runs such that the danger of failure was reduced. They also argue that the banks that failed

The author benefited greatly from comments from Tom Humphrey, Ned Prescott, John Weinberg, and Alex Wolman. Able research assistance was provided by Fan Ding. The views expressed herein are not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

in response to runs were weak and likely to fail regardless of runs (Calomiris and Mason 1997; Benston and Kaufman 1995).

If not fragility, what might explain the widespread failures before 1934? One possible explanation is that the banking industry was experiencing a shakeout, not unusual in industries that have previously enjoyed significant growth. The number of banks had grown briskly from the mid-1880s until 1921. Beginning in 1921, bank failures increased significantly, such that the number of banks began a precipitous decline that continued until 1934. There are reasons to think that the industry had become overbuilt and that macroeconomic shocks, in conjunction with overbuilding, produced a retrenchment in the industry that lasted for the next 12 years. Indeed, many authors point to the relationship between bank failures and weakening economic conditions[1]. This article suggests that overbuilding could have made the banking industry all the more sensitive to macroeconomic shocks.

A number of other industries provide examples of growth followed by shakeouts, the most recent of which is the telecom industry. If a large portion of Depression-era banking failures were the result of a shakeout rather than contagion, an important argument for deposit insurance is undercut.

Though the termination of bank failures and the creation of the FDIC in 1934 occurred simultaneously, implying that contagion must have been at work, other explanations are just as credible. First, deposit insurance augmented the profits of risky banks, protecting them from failure. Second, the creation of deposit insurance undercut a market process that caused supervisors to close troubled banks quickly.

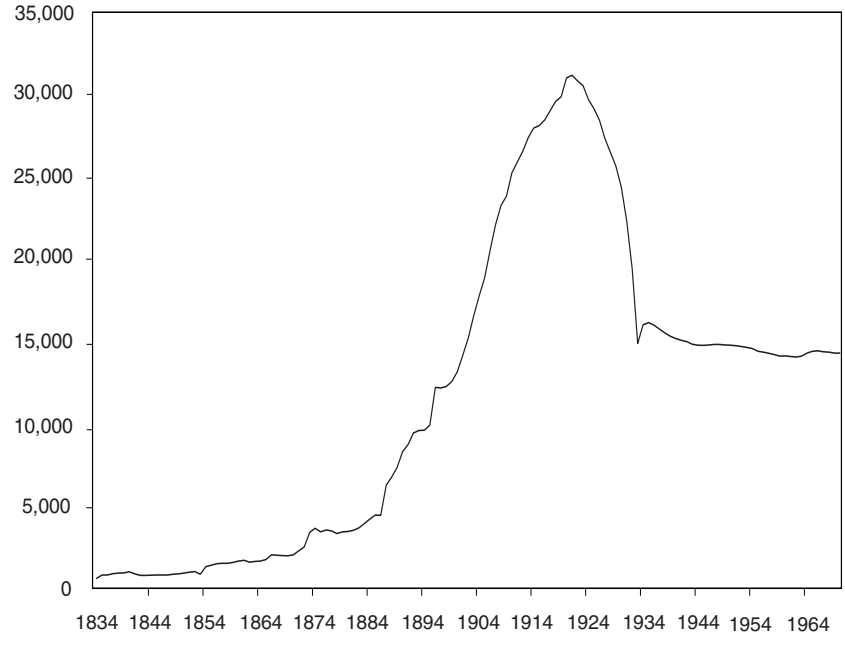
1. GROWTH IN THE NUMBER OF BANKS

The number of U.S. banks grew rapidly from 1887 until 1921 (Figure 1). Much of the increase coincided with improving economic conditions. Yet, commentators also claim that a good portion of the increase resulted from a statutory change that lowered the minimum capital required to form a new bank as well as careless application of entry standards by regulators. Many of the new banks were viewed by commentators as being ill-prepared for the business of banking. In other words, too many banks were formed without adequate financial or managerial resources. The banking market was overbanked.

As shown in Figure 1, the number of banks began growing rapidly in the late 1880s. The initial run-up in the number of banks followed an economic recovery occurring in 1885 and 1886. The increase in the number of banks was rapid enough, and the size of new banks small enough, to drive down the U.S. average bank size fairly significantly. The average size bank shrank from

[1] Temin (1976, 83–95) discusses banking failures that resulted from macroeconomic weakness during the Depression.

Figure 1 Number of Banks

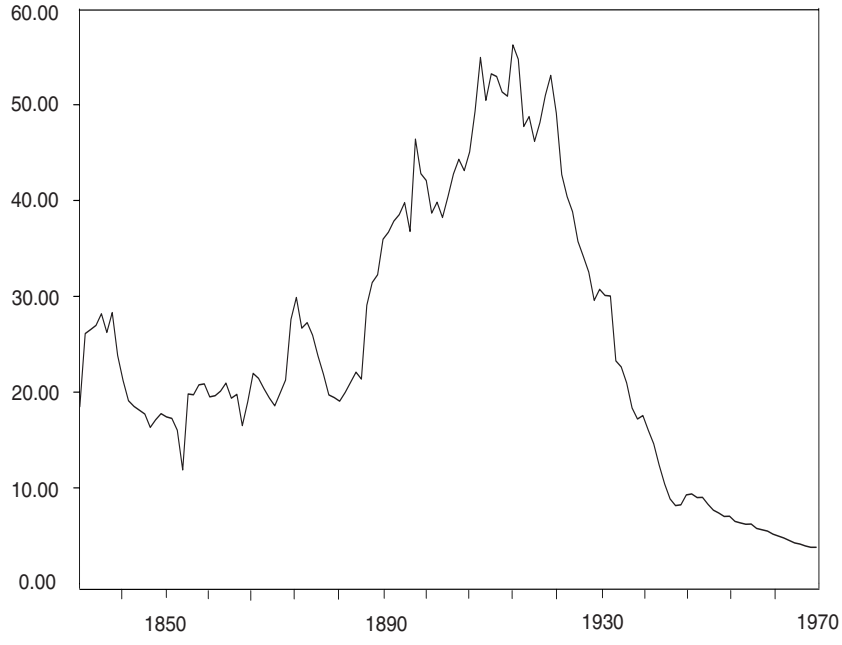


$1.04 million in 1886 to a low of $660,000 in 1896 and did not return to its 1886 level until 1916. The growth in number of banks was much faster than the pace of economic growth, so that the increase in the number of banks is still quite apparent even when the number of banks is deflated by the level of real GDP (Figure 2).

Most commentators focus on the increase in the number of banks, especially of very small banks, after the beginning of the 20th century. Figure 1 shows that the growth in the number of banks was indeed rapid from 1900 until 1921. An important explanation for the growth in the number of banks during these two decades was the reduction in the minimum capital required to form a bank (Mengle 1990; Wheelock 1993). Specifically, the Currency Act of 1900 lowered from $50,000 to $25,000 the minimum capital needed from investors to start a national bank. In turn, over the next ten years, two-thirds of newly formed banks were quite small, averaging capital of only slightly more than the minimum $25,000 (Mengle 1990, 6).

Beyond this reduction in minimum capital, regulatory laxity was also thought to have contributed to the rapid increase in the number of banks. For example, Federal Reserve analysts concluded that during the first two decades of the 20th century "insufficient attention was paid to the qualifications of those

Figure 2 Number of Banks Divided by Real GDP (2000 Dollars)

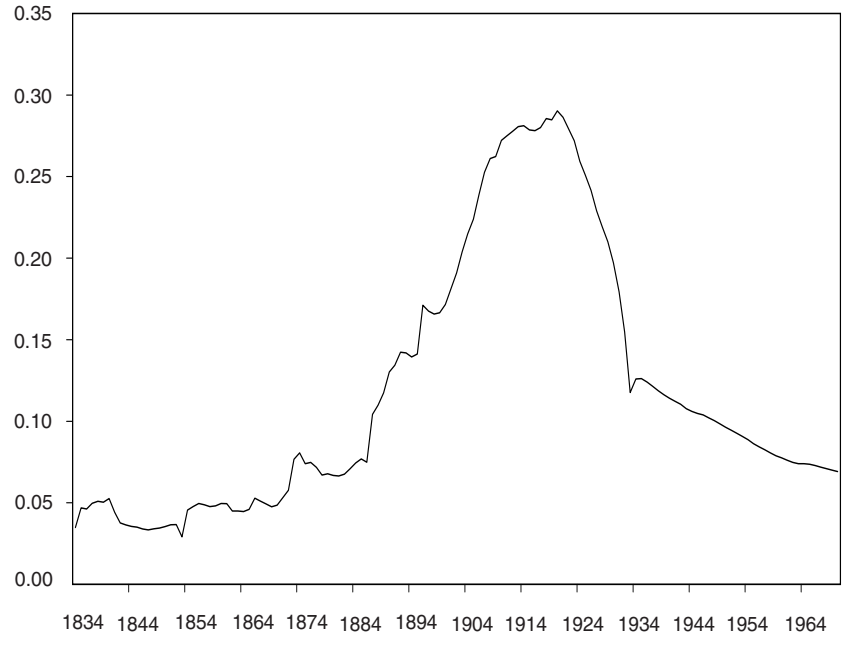


to whom charters were granted" (Federal Reserve Board 1933, 63–65). These observers saw the banking industry as overbuilt by 1920 (Federal Reserve Board 1933, 67).

Most of these new small banks were formed in small towns and rural communities—especially in the corn and cotton belts of the country. Rising prices of farm commodities along with rising farm real estate values may have played a significant role in the attractiveness of rural banking to new investors (Federal Reserve Board 1933, 65). More generally, economic growth was strong between 1887 and 1920, with the annual rate of growth of GDP averaging over 6 percent for the period. GDP growth was especially strong during 1916, and in the two years of the U.S. participation in World War I, 1917 and 1918.

The formation of state deposit insurance systems in a number of states may also have contributed to a perception of safety and allowed the rapid growth of new small banks. Following the banking panic of 1907, eight states adopted such programs for state-chartered banks (Wheelock 1993, 860). Using regression analysis, Wheelock (1993, 865) finds that the presence of deposit insurance systems encouraged bank formation.

Figure 3 Number of Banks per Thousand Persons

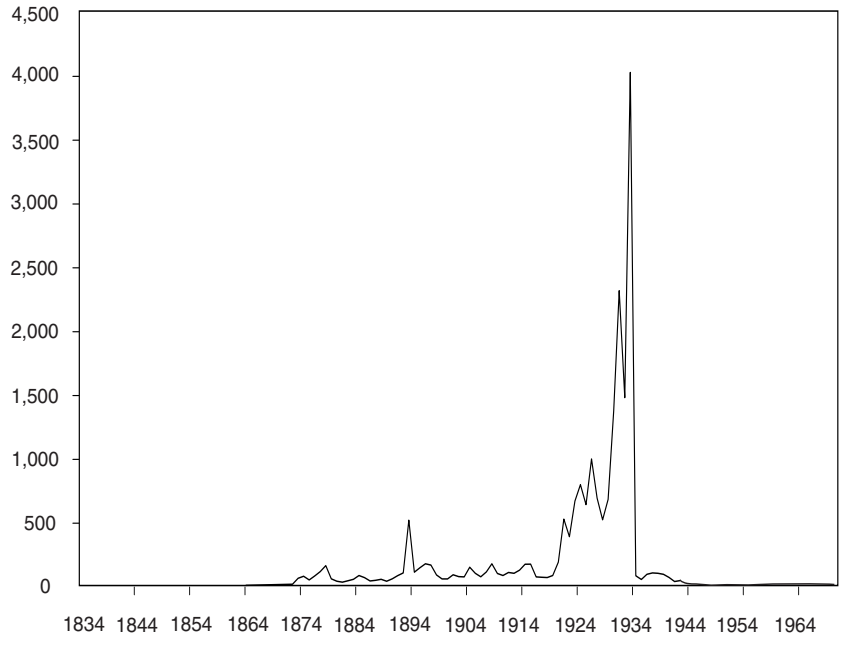


An important trend associated with rapid entry was gradually declining banking profitability. Contemporary writers blamed declining returns on increased competition from new banks. For example, a 1933 Federal Reserve study claimed that "the overbanked condition, which reached its peak shortly after 1920, caused units struggling for existence to increase services to their clients, thus adding to expenses. It served to introduce into many banks high risk and marginal business" (Federal Reserve Board 1933, 67). Net profits relative to assets for the industry fell fairly consistently from 1900 through 1920, from about 2.55 percent to about 1.70 percent (Federal Reserve Board 1933, 67).[2] Bank regulators also noted declining loan standards as rapid entry occurred (Federal Reserve Board 1933, 4).

By several measures the banking industry appears to have become overbuilt. As noted earlier, the numbers grew faster than did overall economic growth. Additionally, as shown in Figure 3, the number of banks increased much more rapidly than did U.S. population. The expansion in the number of competitors may also have been driving down bank profitability and loan

[2] Earnings figures are for national banks only. Until the formation of the Federal Deposit Insurance Corporation, figures on state bank earnings were not collected in any uniform manner.

Figure 4 Number of Bank Failures



standards beginning as early as 1900. This declining profitability implies that even without the agricultural shocks of the 1920s, and the macroeconomic shocks of the Depression, the banking industry would have experienced some retrenchment.

2. 1921–1933 BANKING FAILURES

Banking failures, which began growing in number in the early 1920s, coincide with regional agricultural problems, and later with broader economic problems of the Great Depression. Analysts argue that contagion and branching restrictions account for a significant portion of the failures.

The number of U.S. banks peaked in 1921 at about 31,000 banks. The year also produced the beginning of a period of rapid annual rates of bank failure. In 1921 there were 505 failures. As shown in Figure 4, during most of the next eight years failures remained between 500 and 1,000 per year. From 1930 through 1933 failure rates were in the thousands.

The shift in banking failures in 1921 was precipitated by widespread crop failures of that year. Farm problems were evident in falling farm real estate values in corn- and cotton-producing regions. Real estate values fell each year

from 1921 through 1930 (White 1984, 126). The bank failures of the 1920s were heaviest in states with the most rapid growth prior to the 1920s (Wicker 1996, 7). Difficulties suffered by farmers in the Midwest seem to have driven much of the failure.

Between 1921 and 1930, half of all small banks in agricultural regions failed. Larger banks, however, suffered much less. For example, on average between 1926 and 1930, 74 percent of the smallest banks, those with assets less than $150,000, had weak profits. Here, weak profits are defined as return on equity of less than 6 percent. In contrast, only 21 percent of those banks in the largest size category, with assets greater than $50 million, produced profit rates averaging below 6 percent between 1926 and 1930.

A regulatory shift may also account for the disappearance of some small banks. In the early 1920s the Comptroller of the Currency, the agency that regulates national banks, dropped its branching prohibitions (Mengle 1990, 6). In turn, the number of bank branches grew from 1,400 to 3,500 between 1921 and 1930 (Calomiris and White 2000, 170). These new branches would have brought fresh competition to banking markets, and since branch banks probably had advantages in diversity and scale over small unit banks, unit banks would have been imperiled. Therefore, this liberalization of branching restrictions acted as a shock to small bank profitability, occurring in the 1920s.

Some contemporary commentators claim that improving transportation technology accounts for the decline of small banks, which were once protected from competition by the costs their customers faced to travel to other towns and cities to conduct banking business. These small banks were suddenly faced with new competition once customers' travel costs fell. The growing availability of the car opened the opportunity to purchase services, including banking services, in central cities (Wheelock 1993).

While the number of banking failures (and therefore the decline in the number of banks) grew rapidly in the 1920s, it grew even more rapidly after the onset of the Depression. Between 1930 and 1932, the number of failures per year averaged 1,700. In 1933, slightly more than 4,000 banks failed.

Contagion is often cited to explain the rapid pace of failures between 1930 and 1933. Contagion could work as follows. A prominent bank fails, and because there is no federal deposit insurance protection, depositors of the failed bank suffer losses.[3] Customers of other banks learn of the failure, believe that their bank might suffer the same fate, and *run* their banks—i.e., demand cash repayment of their deposits. Since bank assets are typically tied up in loans and securities, to meet these demands for cash, banks must liquidate these assets. If many banks attempt to sell their securities, prices will fall, and banks will suffer losses on the sales. Further, because outsiders have

[3] Between 1907 and 1917, eight states created deposit insurance systems (White 1983, 207–18). By 1928, all of these state systems had failed.

difficulty determining the worth of bank loans, they will sell at a steep discount, i.e., at firesale prices. Therefore runs of otherwise healthy banks could cause such banks to suffer losses large enough that they would be unable to meet all depositor demands, creating failures of the banks experiencing the runs. The process would become a cycle, spreading widely.

Friedman and Schwartz (1963) identify three banking crises during the Depression involving widespread runs. During these crises, they, along with others, argue that waves of widespread runs created by a "contagion of fear" produced bank illiquidity and the failure of otherwise healthy banks. They hold that much of the bank failure was the result of such contagion and that healthy banks failed as a result.

Branching restrictions are also viewed as an important explanation of bank failures (Mengle 1990, 7–8). While in the 1920s branching restrictions were liberalized for national banks, most states placed severe restrictions on branching, or banned it altogether. As of 1929, 22 states prohibited branching and another ten states restricted it somewhat (Mengle 1990, 6). Branching restrictions prevented banks from diversifying their lending, forcing them to concentrate their lending in one geographic area. The lack of diversity made banks more susceptible to failure caused by localized economic weaknesses. Therefore, oddly enough, both branching restrictions and the branching liberalization, discussed earlier, could have contributed to bank failures.

3. CONTAGION IS AN INCOMPLETE EXPLANATION OF FAILURES

Contagion is an incomplete explanation of banking failures for two reasons. First, contagion was not a factor in the failures that occurred in the 1920s. The bank failures of the 1920s were not caused by and did not create banking panics, that is, banking runs or heavy withdrawals of currency. There was no increase in currency in circulation during the period (Wicker 1996, 1). If factors other than contagion were important in the 1920s, it seems likely that these same factors would also be at work in the 1930s as well. Therefore, contagion is unlikely to account for at least some of the failures in the Depression years of the 1930s.

Second, in their influential work on U.S. banking and monetary history, Friedman and Schwartz (1963) maintain that the bank failures of the 1930s were contagion-induced. Later reviews, however, cast doubt on the contagion explanation. Using regression analysis to identify factors that accounted for individual bank failures during the 1930 crisis, which occurred in November and December of that year, White (1984) finds that these failures were no different from those occurring in the 1920s. In other words, he finds no evidence that these failures were driven by contagion-sparked illiquidity. Instead, these

failures were the result of bad loans. Friedman and Schwartz (1963) give special emphasis to the failure of the Bank of the United States in December 1930 as a cause of contagion-induced bank failures. Wicker (1996, 36–38) questions this view, concluding that the bank's failure did not lead to major deposit withdrawals and could not have accounted for the failures of other banks.

Further, of the bank failures that took place between 1930 and 1932, more than 60 percent occurred outside of panic months (Wicker 1996, 1). While some of these failures might have been the residual effect of runs, it seems likely that most of the failed banks had financial problems unrelated to runs. If an otherwise healthy bank survived a panic, it would have survived the aftermath. Last, Calomiris and Mason (1997) examined the financial condition of Chicago banks during the 1932 banking panic in that city. They found that "while depositors did confuse panic survivors with panic failures, the failure of solvent banks did not result from that confusion." Calomiris and Mason note that the Chicago situation may have been special since Chicago banks met the panic by joining forces to protect banks the group viewed as solvent. Such combined efforts would have been more difficult for the many far-flung banks outside of major cities. Nevertheless, through correspondent relationships, support of banks could have transpired even in a sprawling banking system.

Still, one might wonder: if contagion was not the cause of bank failures, then why did bank failures largely come to a halt with the creation of federal deposit insurance in 1934 (as seen in Figure 4)? If, as I argue, a portion of the bank failures of the 1920s and 1930s were the result of a typical shakeout, there is no reason to expect the shakeout to have ended with the creation of the FDIC. One would only expect failures to be ended by the FDIC's creation if they were caused by contagion, a problem overcome by the creation of federal deposit insurance.

Part of the reason for the cessation of failures, and one that is not dependent on the existence of contagion, is the shifting economic backdrop. Those inherently low-profit banks that were created during the rapid growth of the banking industry would have been the first to be driven to insolvency by the extreme macroeconomic problems of the early 1930s. In other words, economic difficulties concentrated the shakeout in the early 1930s. As a result, low-profit banks created during the growth of the industry—banks which might have continued to fail over a number of years after 1933—were quickly expunged by the severe economic times. The severity of the macroeconomic problems is shown by the declines in economic output, which fell by 8.61 percent in 1930, and by 6.42 percent, 13.00 percent, and 1.27 percent in 1931, 1932, and 1933, respectively. But then in 1934 economic output began growing rapidly, increasing by 10.81 percent, and averaged 7.08 percent from 1934 through 1939 so that banks were strong enough to weather the early 1930s the following years.

An additional reason exists—not dependent on contagion—as to why failures ended at the same time the FDIC was formed: Prior to the FDIC's creation, depositors had strong incentive to monitor their banks' health and, in the case of signs of weakness, withdraw deposits. When depositors withdrew funds, government supervisors were forced to review the state of the bank and either close it or allow it to reopen if shown healthy. Once deposits gained FDIC protection, this disciplining mechanism was removed. Without the mechanism, the opportunity for supervisors to forbear arose. Certainly, in the years immediately after the widespread bank failures of the 1920s and early 1930s, it seems that supervisors would tend to err on the side of stability and be reticent to close banks unless the evidence of a bank's problems was quite strong. Consequently, that failures largely ceased at the same time the FDIC was created is not necessarily evidence that contagion was the cause of most bank failures.

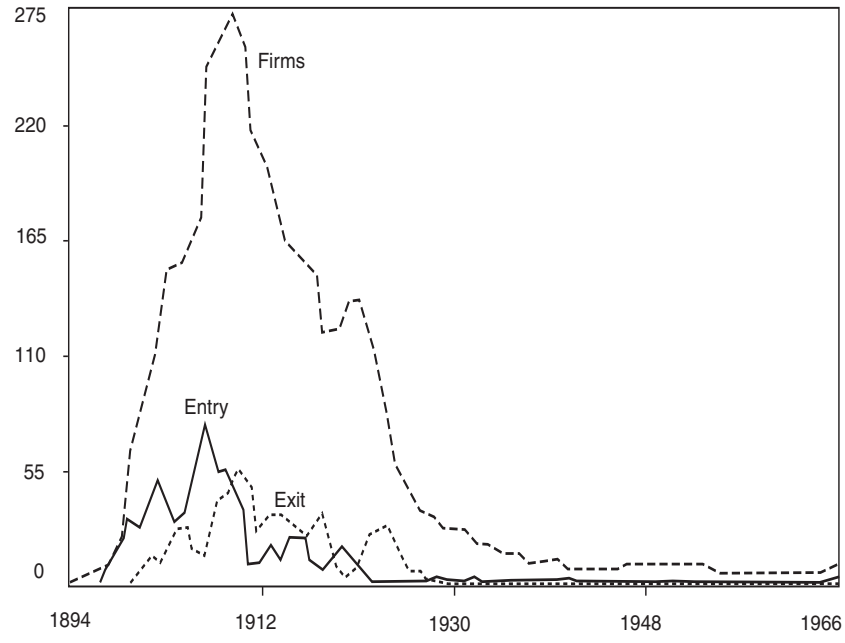
Further, FDIC insurance provided a significant profit-boosting subsidy to the riskiest banks. With the government backing their deposits, in 1934 all banks could suddenly gather deposits at the risk-free interest rate. Those banks that might have failed in the near future—the riskiest banks—benefited most. Unless depositors were completely oblivious to their bank's health, in the absence of deposit insurance, such banks would have had to pay depositors high interest rates. Had the FDIC charged risk-adjusted deposit insurance premia, the riskiest banks would have enjoyed no benefit. But premia were calculated as a simple percentage of deposits. Therefore, troubled banks enjoyed a sudden boost in their profits due to the introduction of FDIC insurance. Failures would naturally be minimized, not because contagion was halted, but instead because FDIC insurance provided a subsidy to those banks most likely to fail.

4. SHAKEOUTS IN OTHER INDUSTRIES

Rapid increase in the number of producers followed by an equally rapid decline is a pattern not only evident in banking, but also in a number of industries. Examples include automobiles, tires, televisions, penicillin, and most recently, telecommunications. While the banking industry failures are blamed on heavy reliance on liabilities that are redeemable upon demand, other factors must have been at work to produce rapid failures in these other industries since they are not reliant upon demand liabilities. Perhaps many of the failures in banking were simply the result of a typical shakeout, driven by factors other than contagion, as in these other industries.

The growth pattern in banking—a significant increase in the number of banks, followed by a rapid decline—is similar to the growth and shakeout that a number of industries experienced, industries for which there are no claims of inherent fragility or contagion-driven runs. Klepper (2002, 37) notes that "After their commercial introduction, the number of producers of automobiles,

Figure 5 Automobiles



Source: Klepper 2002

tires, televisions, and penicillin initially grew and then experienced a sharp decline or shakeout." For example, Figure 5 shows the pattern of growth and decline for automobiles.

More recently, the same pattern was displayed in the telecommunications sector.[4] In the mid-1990s, three factors together implied great opportunities for growth in the industry. First, rapidly growing use of the Internet generated an expanded demand for data transmission services over phone, cable, and long-haul fiber networks. Observers claimed that Internet use was doubling every year, or even several times a year.

Second, an important regulatory change was brought to the telecommunications industry by the Telecommunications Act of 1996. The goal of the Act was to introduce new competition to local telephone service, which was largely monopolized by the regional Bells. It opened local telephone service to competition, for example, by requiring incumbent local service providers

[4] See Couper, Hejkal, and Wolman (2003) for a thorough review of the telecommunications industry's growth in the 1990s and shakeout that began in 2000.

to sell entrants access to the networks that reach homes and businesses. It was thought that entrants would provide local phone service and that selling access would avoid duplication of existing telephone lines to individual homes and businesses. The entrants, known as competitive local exchange carriers (CLECs), rapidly increased in number and earnings. From 30 CLECs in 1996, their number had grown to 711 in 2000. At the same time, revenues of the CLECs rose from $5 billion to $43 billion (Couper et al., 15).

Third, advances in fiber optic technology were occurring rapidly in the mid-1990s, lowering the cost of providing data and voice services to households and businesses. Existing firms, as well as founders of new firms, expected these factors to open huge new markets and responded by rapidly increasing investment in telecommunications equipment and communications lines.

However, as it turned out, by 2000 the promise of the 1996 Act had not been borne out, long-haul fiber communications lines were significantly overbuilt, and the overall economy began to slow. The Act did not lead to nearly as much growth for competitors in local service. The Act's pricing ambiguities meant that rules that would allow entrants to buy access into homes and businesses were slow to be developed by regulatory agencies and were held up by lawsuits. Overcapacity in fiber lines was the result of rapid improvement in the technology for transmitting data over these lines, meaning that fewer were needed than expected. Further, the growth in demand for data communications, while rapid, was not as rapid as expected (Couper et al., 13 and 19). Beginning in mid-2000, the telecom industry began a rapid retrenchment that involved huge numbers of failures of new firms that had only recently appeared to have bright futures. For example, the number of CLECs declined by about 80 percent from 1999 to 2001 (Grossi, 4).

5. EXPLANATION OF THE BANKING SHAKEOUT

Several important industries displayed a pattern of rapid growth followed by widespread failure, similar to that in the banking industry. Yet failures in these industries cannot be blamed on contagion resulting from a reliance on demandable liabilities, implying that something else may also have been at work during the similar boom and bust in banking. Still, these industries differed from banking: Automobiles, tires, televisions, and penicillin were all new industries, and the telecommunications industry was experiencing rapid technological change. Therefore a good bit of instability can be expected as the efficient industrial structure evolved. In contrast, the banking industry of the late 19th and early 20th centuries was not new.

While not new, the banking industry was undergoing rapid change. The industry was a primary source of capital to a nation undergoing rapid technological and industrial change. The expansion of railroads drove down transportation costs. The spread of electronic communication, the telegraph and

telephone, rapidly lowered communications costs. As noted earlier, the popularity of the automobile meant that communities were less isolated from one another, and commerce was better able to flow across local communities. Also, the growth of the corporation as a business structure and the issue of debt by corporations radically changed the financial structure of the business world. Further, branching and capital regulations in banking were being significantly modified.

In an environment of wide-ranging industrial, technological, and regulatory change, it should not be surprising that the banking industry would struggle in ways similar to a new industry. Bank organizers perceived profit opportunities in the changing setting and rapidly formed banks to take advantage of such opportunities. In doing so, they drove down bank size and profitability, as noted earlier. Ultimately they caused the industry to become overbuilt, and failures ensued.

Were bank organizers acting irrationally? Not according to theories economists have advanced to explain the fairly common phenomenon of industry shakeout. Such theories argue that firms can often be expected to expand new investment too far. One such theoretical explanation argues that shakeouts are a result of investments in research and development (R&D) (Klepper 2002, 38).[5] Under this explanation, firms invest in R&D to lower their costs, but the cost advantage grows with size, such that large firms benefit more from R&D. Further, some firms are better at R&D than others.

Firms enter and make R&D investments to acquire the high margins that are expected to accrue from such actions. Instead, rapid entry drives down profit margins. Eventually entry stops, and smaller firms, which have higher per-unit R&D costs, fail. Firms that make less productive use of R&D expenditures also fail, since their costs will be relatively high. In this environment, entry seems to progress too far, since many of the firms that were initially profitable fail when profit margins are driven down. Unanticipated increases in costs or decreases in earnings can also cause failure.

More broadly, overinvestment works as follows. A technological or business process innovation occurs, and firms make investments to take advantage of the new technology. Firms are uncertain concerning how much their output might increase as they implement the new technology but must invest before knowing. Alternatively, firms are uncertain of the extent of demand for their new product, but must invest before knowing. Ultimately firms often overshoot, profits decline, and the least efficient firms fail.

Was such overshooting likely in banking in the late 19th and early 20th centuries? Evolving telecommunications and transportation technology as well as shifting financial arrangements were influencing both the technology

[5] Alternative explanations of boom and shakeouts are offered by Jovanovic and MacDonald (1994) and Barbarino and Jovanovic (2004).

of banking and of the firms to which banks make loans. In this environment, existing banks and potential investors in new banks faced a great deal of uncertainty about the proper scale of the industry, and entry was rapid, relative to GDP growth and population growth. Average profits were driven down in the banking industry, and after 1921, while larger banks remained strong, small bank profits fell significantly. Clearly the demand shocks of the 1920s and the Depression played a major role in the shakeout in banks, but overbuilding appears to have been of significant importance.

6. SUMMARY

As the U.S. economy grew and evolved in the late 19th and early 20th centuries, the banking industry grew even more rapidly, especially in raw numbers of banks, as well as in assets relative to GDP and numbers relative to population. Contemporary analysts maintained that the growth produced an overbuilt industry. Ultimately, failures shrank the number of banks. Failures first became significant in the early 1920s, continued throughout the 1920s, and became even more numerous during the Depression. They largely ended in 1934, at the time of the formation of the FDIC.

A long-held explanation for bank failures during the Depression is contagion, whereby the initial failure of one bank leads to widespread runs on other banks and their failure. According to this explanation, many of the Depression-era failures were inappropriate, meaning that the failed banks were solvent and would have survived without contagion-induced runs. The solution to contagion was deposit insurance provided by the federal government, which put a stop to failures.

Still, given that the cycle of failures began in the early 1920s, long before contagion was evident, one must question contagion as the overriding cause. Instead the banking industry appears to have been experiencing a shakeout, exacerbated by weakened economic conditions during the Depression. While at first blush, the fact that failures stopped virtually simultaneously with the formation of the FDIC implies that contagion must have been at work, other explanations for this simultaneity are just as credible. For example, risky bank profits were certainly boosted by the provision of deposit insurance at premiums that did not reflect bank risk, protecting these banks from failure. Further, the process that drove supervisors to quickly close troubled banks was undercut once deposit insurance was established.

REFERENCES

Barbarino, Alessandro, and Jovanovic, Boyan. 2004. "Shakeouts and Market Crashes." National Bureau of Economic Research Working Paper 10556. Also available online at http://www.nber.org/papers/w10556.pdf (accessed 4 February 2005).

Benston, George J., and George G. Kaufman. 1995. "Is the Banking and Payments System Fragile?" *Journal of Financial Services Research* 9: 209–40.

Calomiris, Charles W., and Joseph R. Mason. 1997. "Contagion and Bank Failures During the Great Depression: The June 1932 Chicago Banking Panic." *American Economic Review* 87 (December): 863–83.

_____, and Eugene N. White. 2000. "The Origins of Federal Deposit Insurance." *U.S. Bank Deregulation in Historical Perspective*. Cambridge: Cambridge University Press, 164–211.

Couper, Elise A., John P. Hejkal, and Alexander L. Wolman. 2003. "Boom and Bust in Telecommunications." Federal Reserve Bank of Richmond *Economic Quarterly* 89 (Fall): 1–24.

Federal Reserve Board. 1933. *Banking Profits, 1890–1931*. Federal Reserve Committee on Branch, Group, and Chain Banking.

Friedman, Milton, and Anna Jacobson Schwartz. 1963. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press (for the National Bureau of Economic Research).

Grossi, Michael. 2004. "A New Ear in Radio Broadcasting: The Coming of Broadband." *Flashpoint, Commentary from Adventis*. Adventis Corporation.

Jovanovic, Boyan, and Glenn MacDonald. 1994. "The Life Cycle of a Competitive Industry." *Journal of Political Economy* 102 (April): 322–47.

Klepper, Steven. 2002. "Firm Survival and the Evolution of Oligopoly." *Rand Journal of Economics* 33 (Spring): 37–61.

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

Temin, Peter. 1976. *Did Monetary Forces Cause the Great Depression?* New York: W. W. Norton and Company, Inc.

Wheelock, David C. 1993. "Government Policy and Banking Market Structure in the 1920s." *The Journal of Economic History* 4 (December): 857–79.

White, Eugene N. 1983. *The Regulation and Reform of the American Banking System, 1900–1929.* Princeton, New Jersey: Princeton University Press.

_____. 1984. "A Reinterpretation of the Banking Crisis of 1930." *Journal of Economic History* 44 (March): 119–38.

Wicker, Elmus. 1996. *The Banking Panics of the Great Depression*. New York: Cambridge University Press.

Banking Markets in a Decade of Mergers: A Preliminary Examination of Five North Carolina Markets

John A. Weinberg

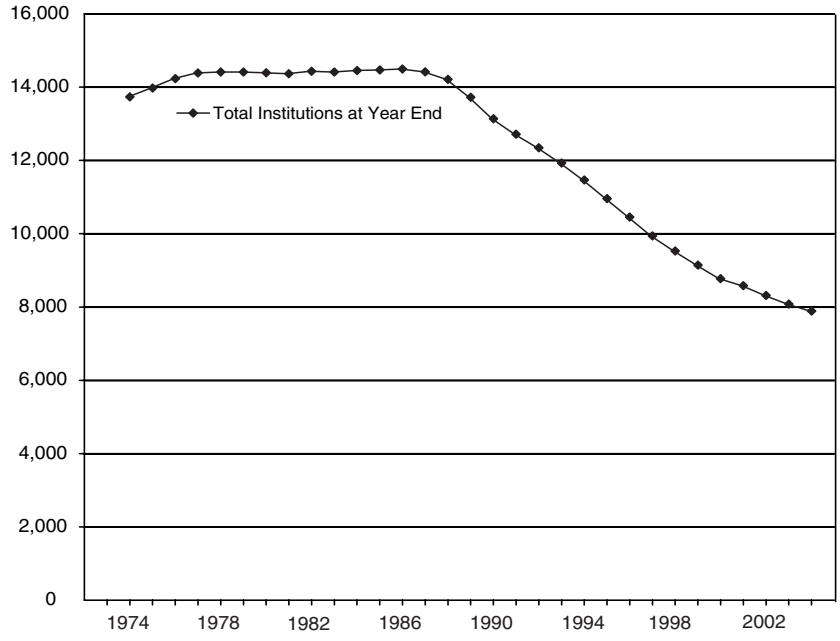
The decade of the 1990s was a period of dramatic consolidation in the banking industry. From 1990 to 2000, the number of banking institutions in the United States fell from 12,212 to 8,252. During this period, the percent of all banking assets held by the largest 1 percent of all banks rose from 54 percent to 70 percent. This consolidation was partly facilitated by the 1994 legislation allowing full interstate branching. Further legislation in 1999 made consolidation across the broader financial services industry easier by loosening restrictions on the combination of commercial and investment banking activities in one company.

On its face, the consolidation of the 1990s appears to be part of a longer wave that began in the mid-1980s. The total number of banks in the United States was notably stable at around 14,000 for a number of years prior to 1986, then began a steady decline. The striking difference in trends is captured in Figure 1.[1] Unsurprisingly, this sustained decline in the number of banks is associated with an historically high level of merger activity, as seen in Figure 2. These figures present a picture of structural change in the banking industry stretching over a period of close to two decades. There is a distinct difference,

Thanks to Patricia Wescott and Shelia Jackson for assistance compiling the data used in this paper. Thanks also to John Walter, Kartik Athreya, and Margarida Duarte for the readings of and comments on an earlier draft, and to Tom Humphrey for his editorial guidance. The views expressed herein are the author's and do not represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] The same figure appears in Ennis (2004), as does a more detailed description of the aggregate trends in failures, mergers, and entry of new banks.

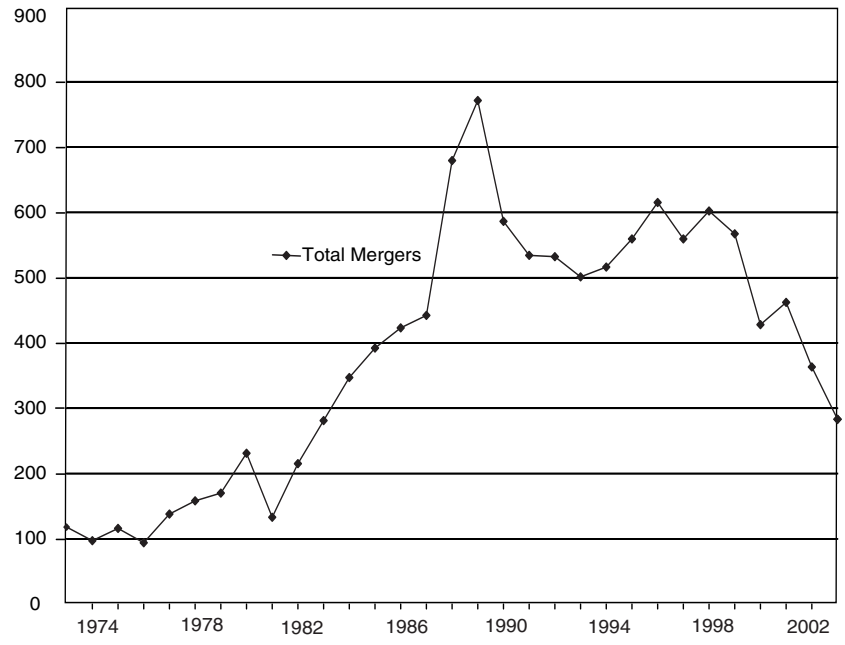
Figure 1 Number of Banks in the United States, 1972–2002



however, between the earlier and later parts of this period. In the 1980s and through the recession of the early 1990s, the banking industry was in a period of crisis, with many undercapitalized or insolvent institutions. Consequently, there was a peak in the number of bank failures in 1988, and much of the merger activity represented the acquisition of weak banks by stronger ones. In contrast to the earlier part of the period, the banking system as a whole has been well capitalized since the mid-1990s, and bank failures have been few in number. Further evidence for differences between the two periods can be seen in the behavior of the concentration of banking assets. As noted above, the percent of assets held by the top 1 percent of banks rose significantly in the 1990s. Before 1990, however, this measure of aggregate concentration changed very little.

Given the difference in the overall condition of the banking industry, the forces driving merger decisions may have been different around the two peaks in merger activity in Figure 2. In the earlier part of the period, both acquirers' strategic considerations and the needs associated with the resolution of weak institutions were important in determining what mergers occurred. In the later part of the period, acquirers' interests were probably predominant. One such consideration is the desire of an acquiring institution to establish or solidify

Figure 2 Bank Mergers, 1972–2002



its position in particular product market segments. That is, a bank with a large share of some of the markets in which it participates might seek to enlarge its share through acquisitions, so as to exercise more control over market prices and enjoy the enhanced profits that come with such market power. Alternatively, mergers might be simply a means of transferring banking resources from less to more efficient users in response to changes in firm-specific or marketwide conditions. This article explores the effect of the merger wave of the 1990s on product markets by examining the evolution of market structure in metropolitan retail banking markets over this period. Specifically, this article explores the changes in market structures in five North Carolina metropolitan areas and asks whether these changes shed light on alternative views of the forces driving mergers.

1. LOCAL BANKING MARKETS

Banks participate in many distinct markets. They provide credit to households and businesses, and credit markets can be further segmented along such lines as the type of household loan (e.g., credit card balances or home-purchase mortgage) or the size of business borrower. Banks also provide deposit

services and the payment services that come with the provision of checkable deposits. Finally, banks provide an increasing array of other financial services, either directly or through holding company affiliates.

Antitrust policy toward bank mergers has long been based on the presumption that bank customers look mainly to nearby institutions for at least some of the financial services they purchase. In particular, attention has centered on services provided to households and small businesses as being most likely to be local in nature. While technological innovations and the development of specialized institutions for some products has certainly made many markets less geographically limited, recent studies suggest that local institutions remain important as providers of certain core banking products to these customers. Such products include consumer and business checkable deposits and unsecured line of credit lending to small businesses. Kwast et al. (1997), in an analysis of the Federal Reserve's Survey of Consumer Finances and National Survey of Small Business Finances, found that these types of customers tend to have a primary banking relationship with a local institution. Examining interest rates on retail deposit accounts, a number of studies, including Heitfield and Prager (2004), find evidence of market segmentation in the form of rate differentials at the level of metropolitan areas. Further, they find such differences as recently as 1999, suggesting that markets for such deposit services continue to be local.

The Federal Reserve's role in bank merger policy, and that of the other federal bank regulatory agencies, derives most significantly from the Bank Merger Act, passed in 1960 and amended in 1966. Under this authority, the Fed examines proposed mergers for possible effects on the competitive structure and behavior of banking markets. Consistent with the evidence on the local nature of markets for some retail banking services, this analysis takes place primarily at the level of local markets. In particular, when there is geographic overlap between the merging institutions' retail operations, the Fed assesses the effects of the merger on the degree of concentration in local markets. For the purposes of this analysis, the Reserve Banks define banking markets in their districts. Many of the defined markets coincide with the metropolitan areas that are used for other statistical purposes.[2] In rural areas, market definitions attempt to link areas according to where people engage in a range of economic activities. Such factors as commuting patterns and the location of major shopping and health care facilities are important for defining markets beyond the well-established metropolitan areas.

The approach to merger policy adopted by the banking regulatory agencies was originally based on the view that banking could be defined as a bundle of services including deposit services and credit to both households and busi-

[2] The Federal Reserve Bank of Richmond's metropolitan area markets are based on the Rand-McNally designations of "regional metropolitan areas," or RMAs.

nesses. The notion of a fixed bundle of services suggested that the relative sizes of two banks would be fairly consistent across services. Hence, a measure of local market concentration in one service, such as deposits, could be used as a broader measure of concentration in the entire bundle of services. Changes in financial services markets, especially in services to large businesses and in some consumer credit products, have weakened the connections among the various services in the traditional bundle. Still, the continuing local nature of markets for retail deposits and small business lending suggests that the approach of assessing concentration in local deposits continues to have some validity in measuring concentration and the possible competitive effects of mergers.[3]

In North Carolina there are 19 banking markets that correspond to metropolitan areas, and 48 rural markets. These markets do not cover the entire state, since a rural area can remain undefined until there is a merger case involving banks in the area. The metropolitan markets range in size from Goldsboro, with a population just over 113,000, to Charlotte, with a population of over 1.3 million.[4] The five markets examined in this article (and their 2000 metropolitan area populations) are Raleigh (797,071), Greensboro (643,430), Durham (426,793), Wilmington (274,478), and Rocky Mount (143, 026).

2. CONCENTRATION AND PRICES

Merger policy is based on the presumption that there is a link between market concentration and the behavior of market prices. The more concentrated a market becomes, it is feared, the more likely are market prices to deviate from the competitive ideal and perhaps approach the pricing of a monopolized market. This concern is founded both on the theory of price-setting in imperfectly competitive markets and on a long history of empirical studies of the relationship between market concentration and prices charged or profits earned by sellers. In banking, a number of studies have found a negative correlation between concentration in local markets and the interest rates paid on retail deposit accounts.

Concentration is typically measured by the Herfindahl-Hirschman Index (HHI), which is the sum of the squares of sellers' market shares. That is, in a market with N sellers, with seller i's share of market sales (or deposits, as market shares are typically measured in banking) denoted by s_i, this concentration

[3] Gilbert and Zaretsky (2003) provide a history and an assessment of the underlying assumptions in the approach to merger policy in banking.

[4] Metropolitan area population numbers are from the 2000 census.

measure is

$$\text{HHI} = \sum_{i=1}^{N} (100 s_i)^2.$$

The greatest possible HHI, for a fully monopolized market, is 10,000, while a market consisting of 100 equal-sized sellers would have an HHI of one. Compared to other potential measures of concentration, for instance, the total share of the top four firms in the market, the HHI has the advantage of being sensitive to both the number of firms and the inequality of market shares among the active firms.

The observation that deposit rates paid to retail customers are negatively correlated with local market concentration is consistent with the hypothesis that mergers are motivated by the desire of acquirers to gain market share and pricing power. Analogously, a large literature finds that, across many industries, seller concentration is positively correlated with prices and seller profits.[5] These same observations, however, are also consistent with a very different theory of the determination of market structure. A critique of the market-power-based hypothesis, originating with Demsetz (1973), argues that the observed correlations between concentration and profits or prices could emerge even if no firms ever exercised any market power. Consider two distinct markets with similar demand conditions. In each market, sellers and potential entrants have access to a production technology with decreasing returns to scale (for large enough output levels) and firm-specific productivity factors that result in cost differences across firms. These factors, which may arise from unique talents of personnel or from locational factors, give some sellers cost advantages that cannot be replicated by competitors, at least in the short run. Now, if in one market, these firm-specific factors tend to be fairly similar across firms (and potential entrants), then in a competitive equilibrium, firms in this market will have relatively equal market shares. Further, with low costs of entry, all firms' profits and prices will be prevented from rising far above normal competitive levels. Contrast this situation with a market in which firm-specific productivity factors vary widely. Now sellers with a cost advantage will enjoy larger shares of market sales. Competitive forces will only drive down the profits of marginal sellers (those with relatively high costs). Low-cost firms will earn higher profits. In short, greater inequality in costs across firms will lead to both greater concentration of total market sales and higher average profits.

What implications does the critique of the market-power hypothesis have regarding the relationship between prices and concentration? The connection here is not as clear as in the case of concentration and profits, and the answer

[5] Gilbert and Zaretsky (2003) provide an excellent review of the literature on banking markets.

is likely to depend on additional characteristics of the markets in question. Plausible models of market behavior might imply either a positive or a negative correlation between concentration and prices. One such specification that implies a positive correlation replaces or supplements the assumption of firm-specific variation in costs with an assumption of firm-specific variation in product characteristics. A seller who is industrious and fortunate enough to produce a product more desirable than its competitors' will enjoy a larger market share and greater profits, and will be able to sell at a higher price. As before, a market with a more unequal distribution of product qualities will be more concentrated and have higher average profits, as well as higher average prices.

In banking, one characteristic that appears to matter for attracting retail deposit customers is location. A bank with a more convenient location may be able to attract deposits while offering a lower interest rate than its competitors. Further, banks with multiple locations may have a competitive advantage, suggesting that size and product quality are somewhat complementary characteristics within a banking market. This dimension of locational advantage, combined with firm-specific differences in the costs of managing multiple-location networks could then result in a distribution of banks according to size, profits, and prices (interest rates). The expected correlations of rates, profits, and concentration would then correspond with those found in cross-market regressions.

So are the observed correlations in banking between local concentration and prices (deposit rates) driven by market power or by competition among heterogeneous sellers? A definitive answer to this question would be difficult to find, and it is likely that both these forces are at work in actual banking markets. But even if purely competitive forces play an important role in driving these cross-market differences, a merger policy based on limiting concentration could still be warranted. Even if large market shares result from the product or cost advantages that some firms have over others, a firm with large market share is more likely to be capable of exercising market power over prices and earning economic profits at the cost of reduced consumer welfare and overall market efficiency. A merger policy that recognizes the possibility of cost efficiencies arising from some large mergers can also recognize that the resulting costs in terms of increased market power could outweigh the gains.

3. MERGERS AND MARKET STRUCTURE

Do mergers lead to increasingly concentrated markets? The trivial answer to this question is yes, in the short run. The immediate effect of a merger between two sellers operating in overlapping markets is that, in the areas of overlap, the total amount of business in the market is divided among a smaller set of competitors than before the merger. Beyond this simple response, the

study of how mergers affect market structure and opportunities for the exercise of market power involves two theoretical questions regarding the behavior of imperfectly competitive markets. First, can two firms in a market increase their pricing power and profits by combining? Second, what are the long-run effects of a merger on market structure, once competitors' responses and the dynamic behavior of the merging firms have been taken into account? The second of these questions inherently refers to dynamic models of market behavior, while the first can be (and has been) addressed in the context of either a static or a dynamic model.

Several papers study the question of whether two firms in an imperfectly competitive market can increase their joint profits by merging. The answer depends on the nature of the strategic interaction among firms. One possibility, pointed out by Salant et al. (1983), is that a reduction in the number of sellers through a merger could cause other rivals to seek to benefit by increasing their own output. These output increases offset the merger's effect on the market price, eliminating any gains for the merging parties. Under alternative models of strategic behavior, other authors have identified conditions under which the merger partners do indeed benefit. Denekere and Davidson (1985), for instance, show that under price competition (with differentiated products), the merger causes other rivals to raise their own prices, leading to gains for everyone, including the merging firms.

Perry and Porter (1985) consider a model of asymmetric competition among sellers with increasing marginal costs that depend on firms' productive capacity. They assume that capacity is tied to physical assets and that there is a fixed amount of such assets available to the sellers in the market.[6] These assumptions have the effect of limiting the competitors' increase in output in response to rising prices. As a result, two firms can find it profitable to merge, reduce their combined outputs, and enjoy the resulting higher prices. Asymmetry in the initial distribution of the productive capacity among the sellers can reinforce this result, making it particularly attractive for two relatively large sellers to merge.

Perry and Porter's analysis can be viewed as a bridge between a static and a dynamic model of market competition. In fact, with freely variable capacity and constant returns to scale, their model is identical to that of Salant et al. If increasing capacity is subject to adjustment costs, and firms make dynamic, strategic investment decisions, the steady state of the model's equilibrium, for a given number of firms, would coincide with the equilibrium of the standard Cournot model. Comparing profits across steady states with different numbers of firms, then, would give the impression that mergers are not profitable, as in Salant et al. But adjustment costs give rise to a transition period, during

[6] Alternatively, they assume that the adjustment costs associated with acquiring and installing new productive capacity are high.

which the merged firms can benefit, as in Perry and Porter. Gowrisankaran (1999) studies mergers in dynamic oligopoly models.

All of these investigations of merger incentives in imperfectly competitive markets share the assumption that changes in market structure do not change the manner in which prices are determined. In all of the foregoing cases, the prices resulting from a given market structure are the equilibrium outcomes of static noncooperative behavior, whether price-setting or quantity-setting. It's possible that dynamic interaction among sellers in a concentrated market could enable sellers to sustain higher prices than those associated with static noncooperative equilibrium. If the feasibility of such tacit collusion depends on market concentration, then the incentives for mergers to gain market power could be considerably strengthened. This type of motivation would tend to favor mergers among sellers that are already fairly large relative to the market.

In many actual banking markets, especially relatively large markets like those examined in the next section, some participants are small enough so that one can reasonably assume that they do not exercise market power. Such markets might best be represented as having some firms with market power and a "competitive fringe" of price-taking firms. Gowrisankaran and Holmes (2000) investigate mergers in a dynamic dominant firm model. For a given initial distribution of productive capacity between the dominant firm and the fringe, they examine the equilibrium path of investment in new capacity and purchase of capacity from the fringe by the dominant firm (mergers). They find that the tendency of the market to become more concentrated over time depends on the initial concentration. Specifically, the greater the dominant firm's initial market share, the more likely the dominant firm will grow by acquiring capacity from the fringe.

In the studies of merger incentives discussed in this section, the driving force is the desire to profit by the exercise of market power. This body of work suggests that such a motivation for mergers is plausible in a variety of market settings, even though entry and growth by smaller firms will eventually erode the monopoly profits acquired. Stigler's (1950) characterization of the merger waves of the early 20th century was consistent with this view. He shows in particular how mergers in many industries created firms with dominant market shares, only to have those dominant positions dwindle over time.

The forgoing discussion focused on studies of merger incentives where the primary motivation was the acquisition or maintenance of market power. But what if market structure is driven by the relative efficiencies of competing firms, rather than by the desire to gain market power? What does this alternative approach imply about the causes and consequences of mergers? In a dynamic version of such a model, like that in Hopenhayn (1992), firms and their market shares grow or decline as their firm-specific characteristics

Table 1 Population and Population Growth 1990–2000

Market	1990 Population	2000 Population	Percent Growth
Raleigh	541,100	797,071	47.3
Greensboro	540,030	643,430	19.1
Durham	344,625	426,793	23.8
Wilmington	200,124	274,478	37.2
Rocky Mount	133,235	143,026	7.3

evolve.[7] A firm that has a productivity advantage at one point in time may see that advantage diminish over time. Such a model predicts market shares that rise or fall but does not distinguish between growth through new investment and growth through acquisition. One can imagine a model in which both internal growth (new investment) and external growth (acquisition) are subject to adjustment costs. The relative costs of the two forms of growth, together with the evolution of firm-specific productivity factors, would then determine the joint pattern of investment and mergers.

4. FIVE NORTH CAROLINA MARKETS

This section examines the behavior of market shares in five North Carolina metropolitan markets from 1991 to 2002. The five markets studied are Durham, Greensboro, Raleigh, Rocky Mount, and Wilmington. These metropolitan areas range in 2000 population from Rocky Mount with 143,026 to Raleigh with 797,021. These two cities, respectively, also had the slowest and fastest population growth of the group between 1990 and 2002. Population and population growth figures for these metropolitan areas are given in Table 1.

This group of markets does not include all the largest banking markets in the state. Most notably, Charlotte and Winston-Salem are not included. Both of these cities housed the headquarters of banks that were among the 10 largest in the United States during this period. Charlotte was the home of Nationsbank, which merged with Bank of America in 1999 to become the second largest bank holding company in the United States, with the combined Bank of America headquartered in Charlotte. Charlotte is also the headquarter city of Wachovia, the fourth largest bank holding company. Winston-Salem was previously the home of Wachovia until it was acquired by First Union (keeping the Wachovia name for the combined institution) in 2001. The status of these cities as headquarters for very large institutions complicates

[7] Erickson and Pakes (1995) provide a dynamic model that includes both imperfect competition and stochastically varying firm-specific productivities. While that model does not address mergers, it could form the basis of a general treatment of merger incentives.

Table 2 Banking Market Characteristics in 1990

Market	Deposits/ Population	HHI	Leading Bank's Share (%)	5 Largest Banks' Share (%)
Raleigh	$9,990	786	15.66	45.02
Greensboro	$11,780	747	15.05	48.13
Durham	$8,800	1286	24.15	63.50
Wilmington	$9,310	1008	19.98	54.13
Rocky Mount	$16,690	1365	25.84	64.89

the interpretation of the deposit data on which market concentration information is based. These data are drawn from the FDIC's Summary of Deposits and are based on banks' reported allocations of their total deposits to their various offices. Large banks that serve large corporate customers may have deposits booked in their headquarter locations that come from more widely dispersed customers. This tendency may have become particularly important after interstate banking powers were expanded by 1994 legislation. The objective of studying market concentration in geographically defined markets is to examine the extent to which relatively few banks dominate a market for local business. For the headquarter cities of large banks, a substantial portion of reported deposits may represent nonlocal business.

Table 2 summarizes some characteristics of these banking markets in 1990. With the exception of Rocky Mount, the markets are similar in terms of deposits relative to population, around $10,000 of deposits per capita. In terms of market concentration, all five markets were either unconcentrated (HHI less than 1,000) or moderately concentrated (HHI between 1,000 and 1,800), as defined by the Department of Justice's merger guidelines. The leading bank's market share ranged from around 15 percent in Greensboro to nearly 26 percent in Rocky Mount. These 1990 market structures fall into three groups. Greensboro and Raleigh appear similar, with HHI less than 800 and the leading firm's share in each market around 15 percent. The "four firm concentration ratios"—the combined market share of the four largest firms in the market—are also similar for Greensboro and Raleigh. These are the two largest metropolitan areas by population, consistent with the general tendency for concentration to be decreasing in city size. The two most concentrated of this group of markets in 1990 were Durham and Rocky Mount. Rocky Mount, the most concentrated, is also the smallest of these metropolitan areas. Wilmington's market concentration characteristics lie between the other pairs of markets.

The differences among the market structures are consistent with evidence that metropolitan areas constitute distinct retail deposit markets. Given that several banks are significant participants in many or all of these cities, one might expect that large, statewide banks would have similar positions in

different cities if the market were integrated on a statewide basis. Instead, in 1990 four different banks had the largest market shares in the five cities. While there are some respects in which the market structures of these cities become more similar over the time period, individual banks' positions remain quite different across the markets.

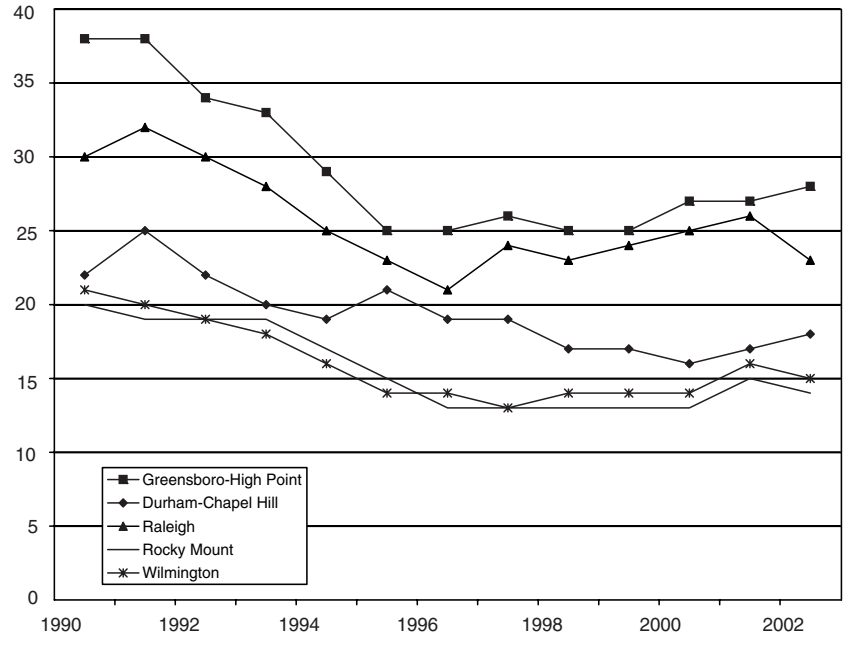
Between 1991 and 2002, there were 45 mergers involving banks that participated in at least one of these markets.[8] Over 80 percent of these were truly *horizontal* mergers, meaning that the merging banks were competitors in at least one market prior to the acquisition. The remaining transactions are better characterized as market extension mergers, in that they involved the purchase of a bank in a market in which the acquirer had no previous presence. In 13 cases, the merging banks had overlapping activities in more than one of the markets. Transactions vary in size from Southern National Bank of North Carolina's 1993 purchase of East Coast Savings Bank, which had a single office in Raleigh with deposits of $6.5 million, to the 2002 merger of First Union and Wachovia, which had overlapping activities in all of five markets, with combined deposits ranging from around $370 million in Rocky Mount to nearly $3 billion in Raleigh.

How did this decade of mergers affect the market structures in these metropolitan areas? Figures 3–6 present information on the evolution of market characteristics from 1990 to 2002. Figure 3 begins with the number of banks operating in this market. In all markets, this period saw a substantial decline in the number of banks. There has also been entry of new banks in all of the markets, with net gains in the number of banks coming toward the end of the period. While all of the markets had fewer banks in 2002 than in 1990, markets did not necessarily receive fewer banking services. As shown in Figure 4, changes in the number of offices (branches) were generally not as dramatic as changes in the number of banks. Rocky Mount experienced a fairly substantial decline in offices (about 25 percent), and Raleigh saw a large increase (about 14 percent). In the other markets the number of offices changed very little (less than 3 percent). It is true that deposits per capita fell slightly in all but one (Rocky Mount) of the markets. While this decline could be due to an overall reduction in supply of deposit services, it is just as likely the result of increasing competition from nonlocal banks or nonbank financial service providers for some segments of customers.

Figures 5 and 6 turn more directly to measures of market concentration. Figure 5 shows the evolution of the HHI for each market. Overall, the markets appear to be more similar at the end of the period than at the beginning, with the two least concentrated markets at the beginning of the period having experienced the largest increases in concentration. But a large part of the

[8] This count includes commercial bank and thrift acquisitions.

Figure 3 Number of Banks, 1990–2002

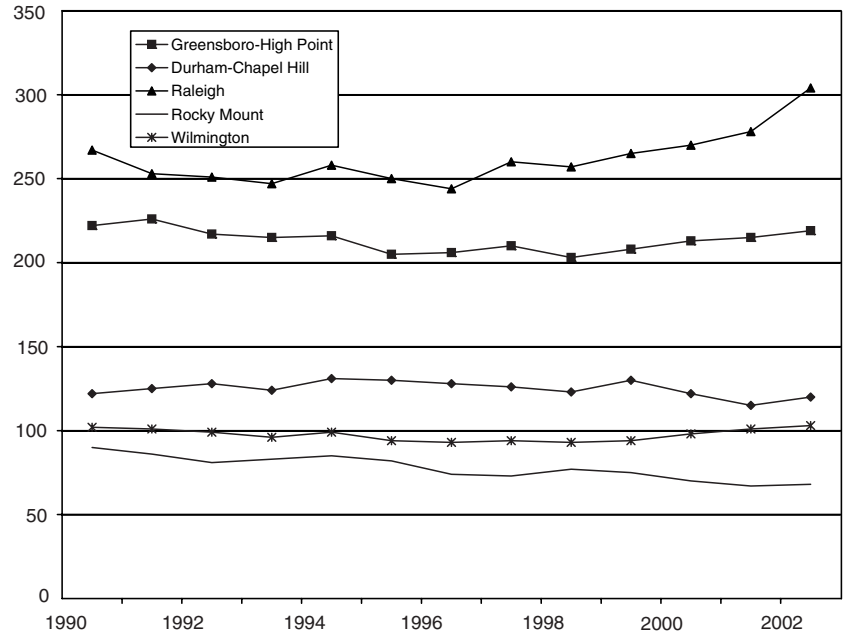


increase in concentration in Raleigh and Greensboro is associated with the First Union-Wachovia merger. These two banks each had market shares in excess of 10 percent in both of these markets at the time of the merger. Differences in concentration across these markets appear to be fairly persistent before the last observation in the data. Markets that start out more concentrated remain more concentrated.

Figure 6 underscores the importance of the First Union-Wachovia merger for the Raleigh and Greensboro markets. This figure shows the market share of the leading bank in each market.[9] In Raleigh and Greensboro, this number changes very little until the last year in the data. A single transaction also plays a large role in the Wilmington market. The relevant merger in this case is BB&T's purchase of United Carolina Bank in 1997. Durham displays a steady increase throughout the period, as does Rocky Mount in the first part of the period, after which the top firm's share declines. In general, the behavior of the top seller's market share seems more idiosyncratic than the broader measure of concentration.

[9] Note that the identity of the leading bank in a market can change over time.

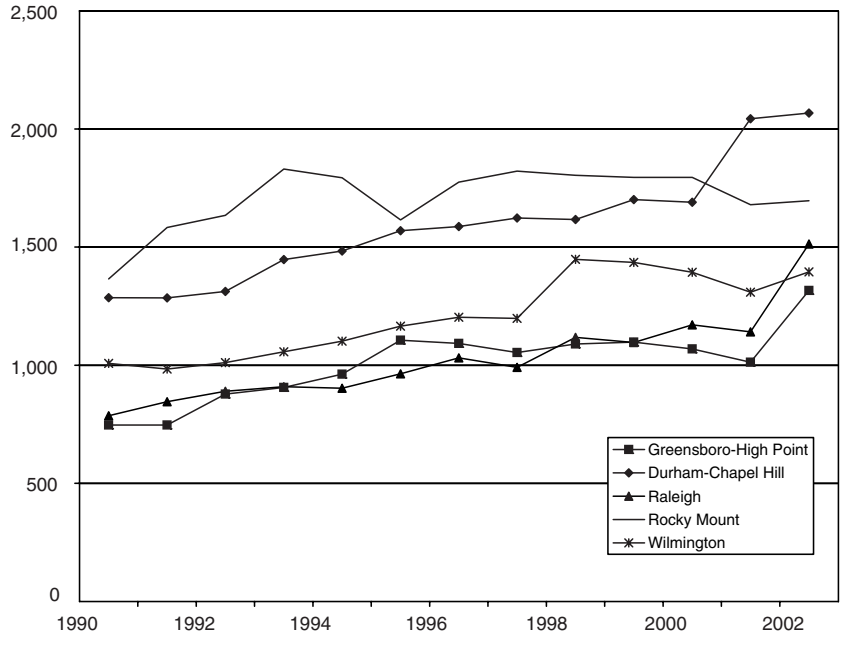
Figure 4 Offices, 1990–2002



How do the observed patterns in these five markets relate to theoretical views on the motivations for and effects of mergers? First, note that the persistence of differences in concentration across markets suggests that, whatever the forces driving consolidation, local factors are important in determining local market structures. There does not appear to be a unique structure toward which these markets are converging. Also, it is important to bear in mind that banking markets in the United States operate under an existing merger policy that limits the degree of concentration that one might expect to observe. Mergers that would create too much concentration might not be allowed, or might be allowed subject to the sale of some of the combined company's branches to a third party. So if there was a tendency, for instance, for all banking markets to eventually become monopolies, that tendency would not show up in the data. Still, these markets are, for the most part, below the level of concentration at which merger policy tends to intervene, and there is not strong evidence that less concentrated markets are consolidating faster than those that are already more concentrated.

While differences between markets seem to persist, it is the case that all of the markets have become more concentrated. Broadly speaking, this rising concentration is consistent with the models of merger incentives in

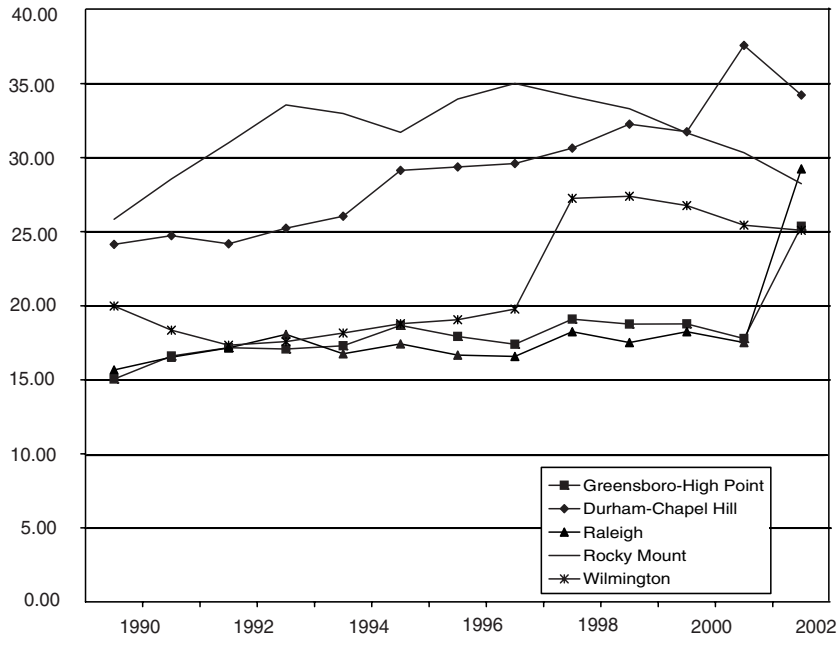
Figure 5 HHI, 1990–2002



imperfectly competitive markets. Some of these models, Perry and Porter (1985) and Holmes and Gowrisankaran (2000) in particular, suggest that the incentive for a market's large participants to grow by acquisition is increasing in the initial level of concentration. This feature does not appear to be present in the cases examined here. Those same models, however, suggest that the increasing incentives may display a threshold effect, changing discretely once concentration becomes big enough. It is possible then, that all of the markets examined are on the same side of the relevant threshold.

The behavior of the leading firms' market shares (Figure 6) contains an interesting mix of patterns. Two of the markets (Rocky Mount and Wilmington) evolve in a way that seems consistent with Stigler's (1950) evidence on early merger waves. A leading firm in a market increases its share through acquisition, and then sees its share decline over time. This pattern suggests an environment in which, due to entry and adjustment costs, gaining market power through acquisition is possible, but only temporarily. In two other markets (Raleigh and Greensboro), the large increase in the leading firm's share comes at the end of the period, so it's not possible to say whether these markets will follow the same pattern. The last market shows a sustained

Figure 6 Share of Largest Bank, 1990–2002



increase in the leading firm's share, which seems consistent with Holmes and Gowrisankaran's model of a dominant firm growing through acquisition.

While the behavior of these markets' structures is consistent with models in which mergers are motivated by the prospect of gains in market power, they are also consistent with the view that mergers are one of the means by which an industry evolves toward an efficient allocation of productive capacity among firms with heterogeneous characteristics. Given the sustained increases in concentration in these markets, one might conclude that the entire period from 1990 to 2002 represents part of a transition from one steady state market structure to another. This view seems consistent with the behavior of the aggregate number of banks, as shown in Figure 1.

5. CONCLUSION

The title of this article includes the phrase "preliminary examination," and the article has attempted to view the evolution of market structures in light of alternative theoretical perspectives on the motives for mergers. The findings suggest that observed behavior of markets is consistent either with a theory based on the acquisition of market power or one based on the efficient allo-

cation of productive capacity in local banking markets. A less preliminary analysis of this topic might proceed in one of two ways. First, one could proceed with a more detailed analysis of a small number of markets. Such a study would consider details of the local economies that might affect the demand for banking services. Important factors may include information about business activities and changes in local labor market conditions. Differences in such demand characteristics are likely to be important for explaining differences in concentration across markets or over time. Accounting for such factors could give one insight into the extent to which changes in market structure represent efficient responses to changes in demand conditions.

A second path to follow would be to develop equilibrium models of industry structure and to take those models to the rich data on banking markets. This is the program set forth by Berry and Pakes (1993) and to which Gowrisankaran (1999) constitutes an important contribution. These authors propose models that incorporate both market power (imperfect competition) and the possibility for efficiency reasons to drive differences in firm size and market share. Such models create the potential for the data to speak more directly to the relative importance of the various forces driving consolidation in markets.

REFERENCES

Bain, Joe. 1951. "The Relation of Profit rate to Industry Concentration: American Manufacturing, 1936–1940." *Quarterly Journal of Economics* 65 (August): 293–324.

Berry, Steven, and Ariel Pakes. 1993. "Some Applications and Limitations of Recent Advances in Empirical Industrial Organization: Merger Analysis." *Papers and Proceedings, American Economic Review* 83 (May): 247–52.

Demsetz, Harold. 1973. "Industry Structure, Market Rivalry and Public Policy." *Journal of Law and Economics* 16 (April): 1–9.

Deneckere, Raymond, and Carl Davidson. 1985. "Incentives to Form Coalitions with Bertrand Competition." *Rand Journal of Economics* 16 (Winter): 473–86.

Ennis, Huberto. 2004. "Some Recent Trends in Commercial Banking." Federal Reserve Bank of Richmond *Economic Quarterly* 90 (Spring): 41–61.

Ericson, Richard, and Ariel Pakes. 1995. "Markov-Perfect Industry Dynamics: a Framework for Empirical Work." *Review of Economic*

Studies 62: 53–82.

Gilbert, Alton, and Adam Zaretsky. 2003. "Banking and Antitrust: Are the Assumptions Still Valid?" *Federal Reserve Bank of St. Louis Review* 85 (November/December): 29–52.

Gowrisankaran, Gautam. 1999. "A Dynamic Model of Endogenous Horizontal Mergers." *Rand Journal of Economics* 30: 56–83.

_____, and Tom Holmes. 2000. "Do Mergers Lead to Monopoly in the Long Run? Results from the Dominant Firm Model." Federal Reserve Bank of Minneapolis Staff Report 264.

Heitfeld, Erik, and Robin Prager. 2004. "The Geographic Scope of Retail Deposit Markets." *Journal of Financial Services Research* 25: 37–55.

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

Kwast, Myron, Martha Starr-McCluer, and John D. Wolken. 1997. "Market Definition and the Analysis of Antitrust in Banking." *Antitrust Bulletin* 42 (Winter): 973–95.

Perry, Martin, and Robert Porter. 1985. "Oligopoly and the Incentive for Horizontal Merger." *American Economic Review* 75 (March): 219–27.

Salant, Stephen, Sheldon Switzer, and Robert Reynolds. 1983. "Losses from Horizontal Merger: The Effects of an Exogenous Change in Industry Structure on Cournot Competition." *Quarterly Journal of Economics* 98 (May): 185–99.

Stigler, George. 1950. "Monopoly and Oligopoly by Merger." *Papers and Proceedings, American Economic Review* 40 (May).