Economic Quarterly—Volume 95, Number 4—Fall 2009—Pages 335–355

Heterogeneity in Sectoral Employment and the Business Cycle

Nadezhda Malysheva and Pierre-Daniel G. Sarte

This paper uses a factor analytic framework to assess the degree to which agents working in different sectors of the U.S. economy are affected by common rather than idiosyncratic shocks. Using Bureau of Labor Statistics (BLS) employment data covering 544 sectors from 1990–2008, we first document that, at the aggregate level, employment is well explained by a relatively small number of factors that are common to all sectors. In particular, these factors account for nearly 95 percent of the variation in aggregate employment growth. This finding is robust across different levels of disaggregation and accords well with Quah and Sargent (1993), who perform a similar analysis using 60 sectors over the period 1948–1989 (but whose methodology differs from ours), as well as with Foerster, Sarte, and Watson (2008), who carry out a similar exercise using data on industrial production.[1]

Interestingly, while common shocks represent the leading source of variation in aggregate employment, the analysis also suggests that this is typically not the case at the individual sector level. In particular, our results indicate that across all goods and services, common shocks explain on average only 31 percent of the variation in sectoral employment. In other words, employment at the sectoral level is driven mostly by idiosyncratic shocks, rather than common shocks, to the different sectors. Put another way, it is not the case that "a rising tide lifts all boats." Moreover, it can be easy to overlook the influence of idiosyncratic shocks since these tend to average out in aggregation.

We wish to thank Kartik Athreya, Sam Henly, Andreas Hornstein, and Thomas Lubik for helpful comments. The views expressed in this article do not necessarily represent those of the Federal Reserve Bank of Richmond, the Board of Governors of the Federal Reserve System, or the Federal Reserve System. All errors are our own.

[1] See also Forni and Reichlin (1998) for an analysis of ouput and productivity in the United States between 1958 and 1986.

Despite the general importance of idiosyncractic shocks in explaining movements in sectoral employment, we nevertheless further document substantial differences in the way that sectoral employment is tied to these shocks. Specifically, we identify sectors where up to 85 percent of the variation in employment is driven by the common shocks associated with aggregate employment variations. Employment in these sectors, therefore, is particularly vulnerable to the business cycle with little in the way of idiosyncratic shocks that might be diversified away. These sectors are typically concentrated in construction and include, for example, residential building.

More generally, our empirical analysis indicates that employment in goods-producing industries tends to more tightly reflect changes in aggregate conditions relative to service-providing industries. However, even within the goods-producing industries, substantial heterogeneity exists in the way that sectoral employment responds to common shocks. For instance, the durable goods and construction industries are significantly more influenced by common shocks than the nondurable goods and mining industries. Among the sectors where employment is least related to aggregate conditions are government, transportation, and the information industry.

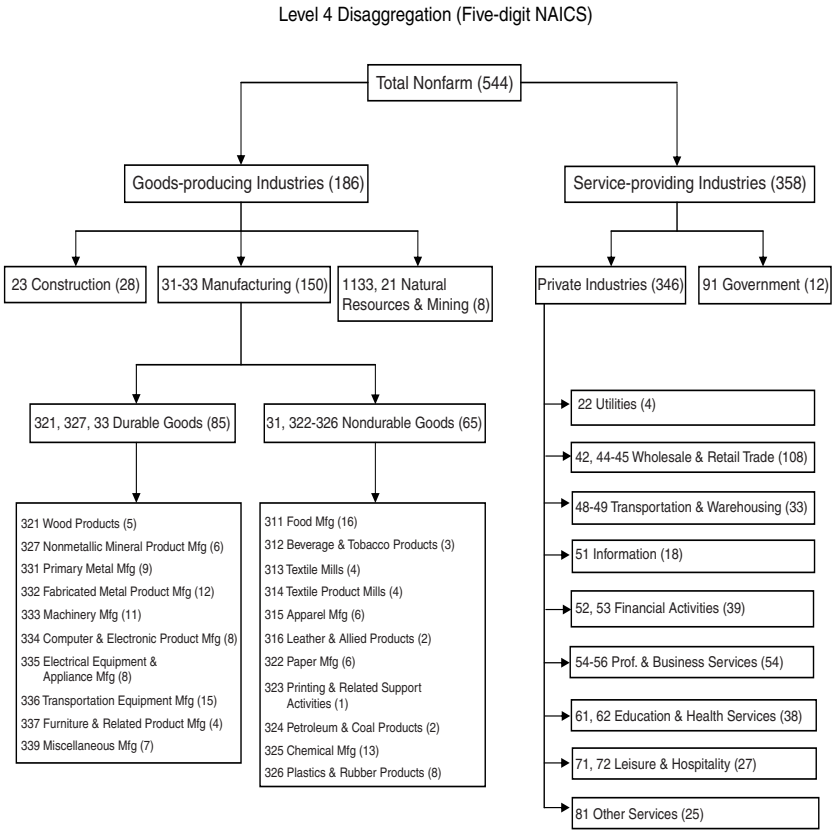
Finally, we present evidence that the factors uncovered in our empirical work play substantially different roles in explaining aggregate and sectoral variations in employment. Although the findings we present are based on a three-factor model, our analysis suggests that one factor is enough to explain roughly 94 percent of the variation in aggregate employment. At the same time, however, that factor appears almost entirely unrelated to employment movements in specific sectors such as in natural resources and mining or education and health services. Interestingly, the reverse is also true in the sense that the analysis identifies factors that significantly help track employment movements in these particular sectors but that play virtually no role in explaining aggregate employment fluctuations.

This article is organized as follows. Section 1 provides an overview of the data. Section 2 describes the factor analysis and discusses key summary statistics used in this article. Section 3 summarizes our findings and Section 4 offers concluding remarks.

1. OVERVIEW OF THE DATA

Our analysis uses data on sectoral employment obtained from the BLS covering the period 1990–2008. The data are available monthly, seasonally adjusted, and disaggregated by sectors according to the North American Industry Classification System (NAICS). Our data cover the period since 1990, the date at which this classification system was introduced. Prior to 1990, BLS employment data were disaggregated using Standard Industry Classification codes, which involve a lower degree of disaggregation and were discontinued

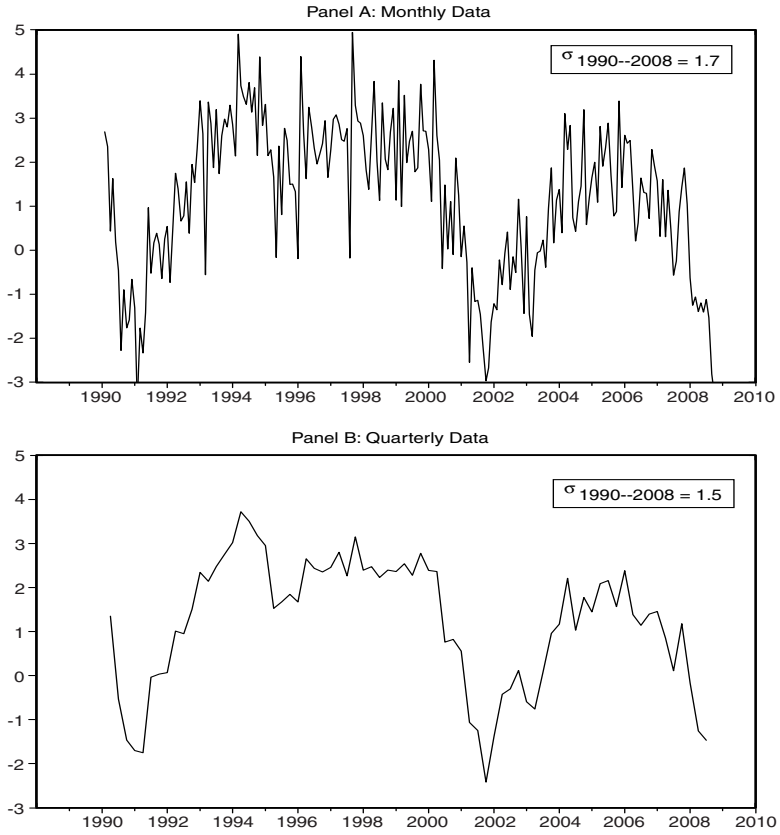
Figure 1 Breakdown of Sectoral Employment Data



as of 2002. For most of the article, we use a five-digit level of disaggregation that corresponds to 544 sectors, although our findings generally apply to other levels of disaggregation as well. The raw data measure the number of employees in different sectors, from which we compute sectoral employment growth rates as well as the relative importance (or shares) of industries in aggregate employment.

When aggregated, the data measure total nonfarm employment. Nonfarm employment is further subdivided into two main groups: goods-producing sectors, comprising 186 sectors at the five-digit level, and service-providing sectors, comprising 358 sectors. The goods-producing sectors are further subdivided into three main categories: construction, with 28 sectors; manufacturing, with 150 sectors; and natural resource and mining, with eight sectors. The manufacturing component of the goods sector contains two main categories: durable goods, comprising 85 sectors, and nondurable goods, with 65 sectors.

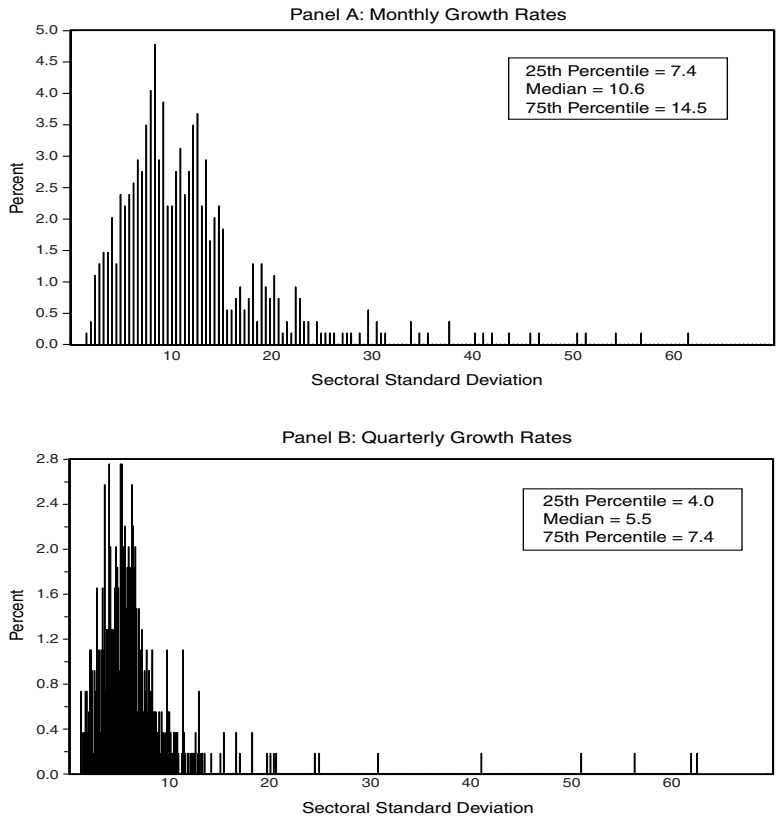
Figure 2 Monthly and Quarterly Employment, All Goods and Services



The service-providing sectors employ more than four times as many workers as the goods-producing sectors. They are made up of two main components: government, with 12 sectors, and a variety of private industries that include 346 sectors spanning wholesale and retail trade, information, financial activities, education and health, as well as many other services. Figure 1 illustrates a breakdown of our sectoral data, along with the number of industries within each broad category of sectors in parenthesis, as well as their corresponding NAICS codes.

Let e_t denote aggregate employment across all goods- and services-producing industries at date t, and let e_{it} denote employment in the ith industry. We construct quarterly values for employment as averages of the months in the quarter. We further denote aggregate employment growth by Δe_t and employment growth in industry, i, by Δe_{it}. At the monthly frequency, we compute Δe_{it} as $1,200 \times \ln(e_{it}/e_{it-1})$ and, at the quarterly frequency, as

Figure 3 Distribution of Standard Deviations of Sectoral Growth Rates (1990–2008)



$400 \times \ln(e_{it}/e_{it-1})$. Aggregate employment growth is computed similarly. Finally, we represent the $N \times 1$ vector of sectoral employment growth rates, where N is the number of sectors under consideration, by $\Delta \mathbf{e}_t$.

Figures 2A and 2B illustrate the behavior of aggregate employment growth at the monthly and quarterly frequencies, respectively, over our sample period. Monthly aggregate employment growth is somewhat more volatile than quarterly employment growth, but in either case the recessions of 1991 and 2001 stand out markedly. At a more disaggregated level, Figures 3A and 3B show the distributions of standard deviations of both monthly and quarterly sectoral employment growth across all 544 sectors. As with aggregate data, quarterly averaging reduces the volatility of sectoral employment. More importantly, it is clear that there exists substantial heterogeneity across sectors in the sense

Table 1 Standard Deviation of Employment Growth Rates

	Monthly Growth Rates	Quarterly Growth Rates
Full Covariance Matrix	1.8	1.5
Diagonal Covariance Matrix	0.7	0.4

Notes: The table reflects percentage points at an annual rate.

that fluctuations in employment are unequivocally more pronounced in some sectors than others.

Let s_i denote the (constant mean) share of sector i's employment in aggregate employment and the corresponding $N \times 1$ vector of sectoral shares be denoted by $\mathbf{s}$. Then, we can express aggregate employment growth as $\Delta e_t = \mathbf{s}'\Delta\mathbf{e}_t$. Furthermore, it follows that the volatility of aggregate employment growth in Figure 2, denoted σ_e^2, is linked to individual sectoral employment growth volatility in Figure 3 through the following equation,

$$\sigma_e^2 = \mathbf{s}'\Sigma_{ee}\mathbf{s}, \tag{1}$$

where Σ_{ee} is the variance-covariance matrix of sectoral employment growth. Thus, we can think of the variation in aggregate employment as driven by two main forces—individual variation in sectoral employment growth (the diagonal elements of Σ_{ee}) and the covariation in employment growth across sectors (the off-diagonal elements of Σ_{ee}).[2]

Table 1 presents the standard deviation of aggregate employment, σ_e^2, computed using the full variance-covariance matrix Σ_{ee} in the first row, and using only its diagonal elements in the second row. As stressed in earlier work involving sectoral data, notably by Shea (2002), it emerges distinctly that the bulk of the variation in aggregate employment is associated with the covariance of sectoral employment growth rates rather than individual sector variations in employment. The average pairwise correlation in sectoral employment is positive at approximately 0.10 in quarterly data and 0.04 in monthly data. Moreover, if one assumed that the co-movement in sectoral employment growth is driven primarily by aggregate shocks, then Table 1 would immediately imply that these shocks represent the principal source of variation in aggregate employment. For example, focusing on quarterly growth rates, the fraction of aggregate employment variability explained by aggregate shocks would amount roughly to $1 - (0.4^2/1.5^2)$ or 0.93. This calculation, of course, represents only an approximation in the sense that the diagonal elements of Σ_{ee} would themselves partly reflect the effects of changes

[2] As in Foerster, Sarte, and Watson (2008), time variation in the shares turns out to be immaterial for the results we discuss in this article.

in aggregate conditions. That said, it does suggest, however, that despite clear differences in employment growth variability at the individual sector level, these differences, for the most part, vanish in aggregation and so become easily overlooked.

2. A FACTOR ANALYSIS OF SECTORAL EMPLOYMENT

As discussed in Stock and Watson (2002), the approximate factor model provides a convenient means by which to capture the covariability of a large number of time series using a relatively few number of factors. In terms of our employment data, this model represents the $N \times 1$ vector of sectoral employment growth rates as

$$\Delta \mathbf{e}_t = \boldsymbol{\lambda} \mathbf{F}_t + \mathbf{u}_t, \tag{2}$$

where $\mathbf{F}_t$ is a $k \times 1$ vector of unobserved factors common to all sectors, $\boldsymbol{\lambda}$ is an $N \times k$ matrix of coefficients referred to as factor loadings, and $\mathbf{u}_t$ is an $N \times 1$ vector of sector-specific idiosyncratic shocks that have mean zero. We denote the number of time series observations in this article by T. Using (1), the variance-covariance matrix of sectoral employment growth is now simply given by

$$\Sigma_{ee} = \boldsymbol{\lambda} \Sigma_{FF} \boldsymbol{\lambda}' + \Sigma_{uu}, \tag{3}$$

where Σ_{FF} and Σ_{uu} are the variance-covariance matrices of $\mathbf{F}_t$ and $\mathbf{u}_t$, respectively.

In classical factor analysis, Σ_{uu} is diagonal so that the idiosyncratic shocks are uncorrelated across sectors. Stock and Watson (2002) weaken this assumption and show that consistent estimation of the factors is robust to weak cross-sectional and temporal dependence in these shocks. Equation (2) can be interpreted as the reduced form solution emerging from a more structural framework (see Foerster, Sarte, and Watson 2008). Given this, features of the economic environment that might cause the "uniquenesses," $\mathbf{u}_t$, to violate the weak cross-sectional dependence assumption include technological considerations, such as input-output (IO) linkages between sectors or production externalities across sectors. In either case, idiosyncratic shocks to one sector may propagate to other sectors via these linkages and give rise to internal co-movement that is ignored in factor analysis. Using sectoral data on U.S. industrial production, Foerster, Sarte, and Watson (2008) show that the internal co-movement stemming from IO linkages in a canonical multisector growth model is, in fact, relatively small. Hence, the factors in that case capture mostly aggregate shocks rather than the propagation of idiosyncratic shocks by way of IO linkages. Thus, for the remainder of this article, we shall interpret $\mathbf{F}_t$ as capturing aggregate sources of variation in sectoral employment.

When N and T are large, as they are in this article, the approximate factor model has proved useful because the factors can simply be estimated by

principle components (Stock and Watson 2002). By way of illustration, the Appendix provides a brief description of the principle component problem and its relationship to the approximate factor model (2). Bai and Ng (2002) further show that penalized least-square criteria can be used to consistently estimate the number of factors, and the estimation error in the estimated factors is sufficiently small that it need not be taken into account in carrying out variance decomposition exercises (Stock and Watson 2002).

Key Summary Statistics

Given equation (2), we shall summarize our findings in mainly two ways. First, we compute the fraction of aggregate employment variability explained by aggregate or common shocks, which we denote by $R^2(\mathbf{F})$. In particular, since $\Delta e_t = \mathbf{s}'\Delta\mathbf{e}_t = \mathbf{s}'\boldsymbol{\lambda}\mathbf{F}_t + \mathbf{s}'\mathbf{u}_t$, we have that

$$R^2(\mathbf{F}) = \frac{\mathbf{s}'\boldsymbol{\lambda}\Sigma_{FF}\boldsymbol{\lambda}'\mathbf{s}}{\sigma_e^2}. \tag{4}$$

For the 544 sectors that make up all goods and services at the five-digit level, $R^2(\mathbf{F})$ then captures the degree to which fluctuations in aggregate employment growth are driven by aggregate rather than sector-specific shocks. Second, we also assess the extent to which aggregate shocks explain employment growth variability in individual sectors. More specifically, denoting a typical equation for sector i in (2) by

$$\Delta e_{it} = \lambda_i \mathbf{F}_t + u_{it}, \tag{5}$$

where λ_i represents the $1 \times k$ vector of factor loadings specific to sector i and u_{it} denotes sector i's idiosyncratic shocks, we compute

$$R_i^2(\mathbf{F}) = \frac{\lambda_i'\Sigma_{FF}\lambda_i}{\sigma_{e_i}^2}, \tag{6}$$

where $\sigma_{e_i}^2$ is the variance of employment growth in sector i.

Note that the analysis yields an entire distribution of $R_i^2(\mathbf{F})$ statistics, one for each sector. Consider the degenerate case where $R_i^2(\mathbf{F}) = 1$ for each i. In this case, employment variations in each sector are completely driven by the shocks common to all sectors and idiosyncratic shocks play no role. Put another way, variations in aggregate employment reflect only aggregate shocks and the fate of each sector is completely tied to these shocks. A direct economic implication, therefore, is that the issue of market incompleteness or insurance considerations (at the sectoral level) tend to become irrelevant as there is no scope for diversifying away idiosyncratic shocks. To the extent that factor loadings differ across sectors, aggregate shocks still affect sectoral employment differentially so that there may remain some opportunity to complete markets. However, in the limit where $\lambda_i = \lambda_j \ \forall i, j$, the standard

Table 2 Decomposition of Variance from the Approximate Factor Model

	Monthly Growth Rates			Quarterly Growth Rates		
	1 Factor	2 Factors	3 Factors	1 Factor	2 Factors	3 Factors
Std. Dev. of Δe_t Implied by Factor Model	1.80	1.80	1.80	1.53	1.53	1.53
$R^2(\mathbf{F})$	0.77	0.80	0.80	0.94	0.95	0.95

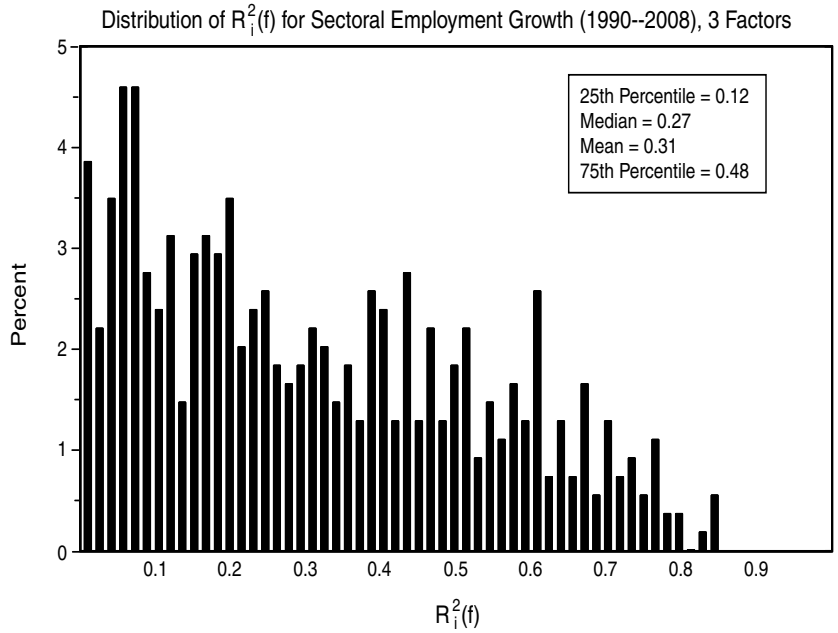
representative agent setup becomes a sufficient framework with which to study business cycles (i.e., without loss of generality). In contrast, when $R_i^2(\mathbf{F}) < 1$ for a subset of sectors, it is no longer true that the fortunes of individual sectors are dictated only by aggregate shocks. Sector-specific shocks help determine sectoral employment outcomes, and the degree of market completeness potentially plays an important part in determining the welfare implications of business cycles.

3. EMPIRICAL FINDINGS

Tables 2 through 4, as well as Figures 4 and 5, summarize the results from computing these key summary statistics using our data on sectoral employment growth rates. We estimated the number of factors using the Bai and Ng (2002) ICP1 and ICP2 estimators, both of which yielded three factors over the full sample period. For robustness, Table 2 shows the factor model's implied standard deviation of aggregate employment (computed using constant shares), as well as the fraction of aggregate employment variability explained by the common factors, $R^2(\mathbf{F})$, using either one, two, or three factors. Most of our discussion will focus on the three-factor model. Two important observations stand out in Table 2. First, the common factors explain essentially all of the variability in quarterly employment growth rates. These common shocks also explain the bulk, or more specifically 80 percent, of fluctuations in monthly growth rates. Second, note that for both monthly and quarterly growth rates, the first factor almost exclusively drives aggregate employment growth, with the second and third factors contributing little additional variability to the aggregate series in relative terms. That is not to say that the absolute variance of the latter factors is small, and we shall see below that these are essential in helping track subsets of the sectors that make up total nonfarm employment.

At a more disaggregated level, Figure 4 illustrates the fraction of quarterly employment growth variability in individual sectors that is attributable to common shocks or, alternatively, the distribution of $R_i^2(\mathbf{F})$. As the figure makes

Figure 4 Contribution of Sector-Specific Shocks to Sectoral Employment



clear, sector-specific shocks play a key role in accounting for employment variations at the sectoral level, with common shocks explaining, on average, only 31 percent of the variability in sectoral employment. In addition, observe that there exists substantial heterogeneity in the way that employment is driven by aggregate and idiosyncratic shocks across sectors. Specifically, the interquartile range suggests $R_i^2(\mathbf{F})$ statistics that are between 0.12 to 0.48, or a 0.36 point gap.

It may seem counterintuitive at first that $R^2(\mathbf{F})$ is close to 1 in Table 2 while the mean or median $R_i^2(\mathbf{F})$ statistic is considerably less than 1 in Figure 4. To see the intuition underlying this result, consider equation (2) when aggregated across sectors:

$$\mathbf{s}'\Delta\mathbf{e}_t = \mathbf{s}'\boldsymbol{\lambda}\mathbf{F}_t + \mathbf{s}'\mathbf{u}_t. \tag{7}$$

When the number of sectors under consideration is large, as in this article, the "uniquenesses" will tend to average out by the law of large numbers. Put another way, since the u_{it}s are weakly correlated across sectors and have mean zero, $\mathbf{s}'\mathbf{u}_t = \sum_{i=1}^{N} s_i u_{it} \to^p 0$ as N becomes large. This result

Table 3 Fraction of Variability in Sectoral Employment Growth Explained by Common Shocks

Sector	$R_i^2(\mathbf{F})$
Residential Building Construction	0.85
Electrical Equipment Manufacturing	0.85
Wood Kitchen Cabinet and Countertop	0.84
Plumbing and HVAC Contractors	0.84
Printing and Related Support Activities	0.80
Other Building Material Dealers	0.80
Wireless Telecommunications Carriers	0.78
Construction Equipment	0.78
Plywood and Engineered Wood Products	0.77
Semiconductors and Electronic Components	0.77
Management of Companies and Enterprises	0.77
Electrical Contractors	0.77
Lumber and Wood	0.77
Metalworking Machinery Manufacturing	0.76
Electric Appliance and Other Electronic Parts	0.76

holds provided that the distribution of sectoral shares is not too skewed so that a few sectors have very large weights (see Gabaix 2005). In contrast, $\mathbf{s}'\boldsymbol{\lambda}\mathbf{F}_t = \mathbf{F}_t \sum_{i=1}^{N} s_i\lambda_i$ does not necessarily go to zero with N since the λ_is are fixed parameters.[3] Therefore, whatever the importance of idiosyncratic shocks in driving individual sectors (i.e., whatever the distribution of $R_i^2(\mathbf{F})$), $R^2(\mathbf{F})$ will generally tend towards 1 in large panels. The rate at which $R^2(\mathbf{F})$ approaches 1 will depend on the particulars of the data-generating process. In this case, with 544 sectors, we find that $R^2(\mathbf{F})$ is around 0.8 in monthly data and 0.95 in quarterly data.

Interestingly, Figure 4 suggests that at the high end of the cross-sector distribution of $R_i^2(\mathbf{F})$ statistics, there exist individual sectors whose variation in employment growth is almost entirely driven by the common shocks that explain aggregate employment, and, thus, that are particularly vulnerable to the business cycle. Table 3 lists the top 15 sectors in which idiosyncratic shocks play the least role in relative terms. Note that all of the sectors listed in Table 3 are goods-producing sectors. In other words, even though service-providing sectors employ more than four times as many workers as the goods-producing sectors, it turns out that it is the latter sectors that are most informative about the state of aggregate employment. In essence, because employment variations in the sectors listed in Table 3 reflect mainly the effects of common shocks, and because movements in aggregate employment growth are associated with

[3] In Foerster, Sarte, and Watson (2008), the factor loadings correspond to reduced-form parameters that can be explicitly tied to the structural parameters of a canonical multi-sector growth model.

Table 4 Sectoral Information Content of Aggregate Employment

Selected Sectors Ranked by $R_i^2(\mathbf{F})$	Fraction of $\Delta\mathbf{e}_t$ Explained by Selected Sectors
Top 5 Sectors	0.88
Top 10 Sectors	0.92
Top 20 Sectors	0.94
Top 30 Sectors	0.96

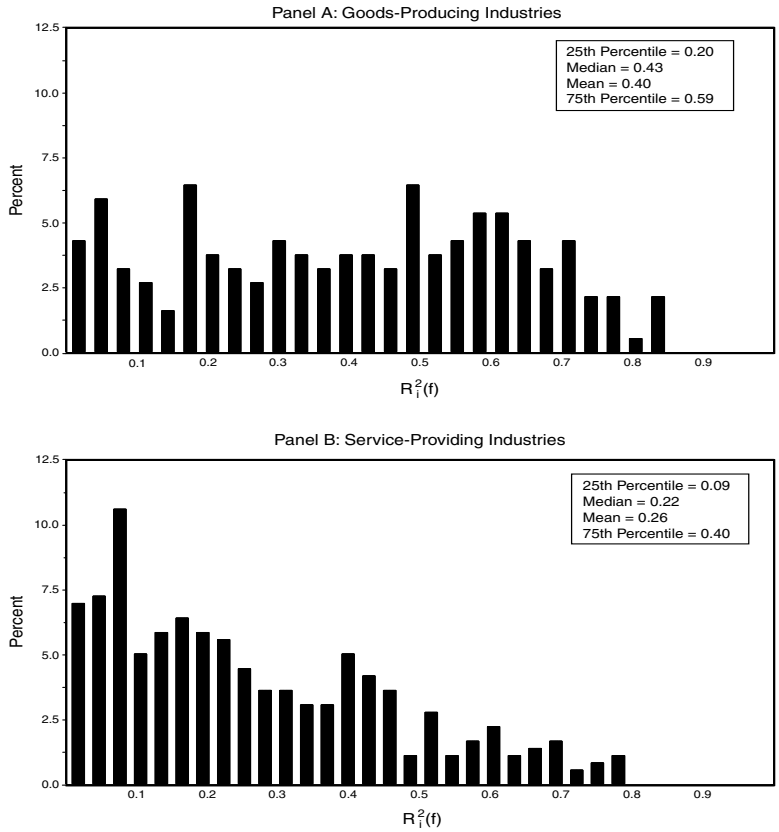
these shocks (Table 2), information regarding aggregate employment tends to be concentrated in these sectors.

This notion of sectoral concentration of information regarding aggregate employment can be formalized further as follows. Consider the problem of tracking movements in aggregate employment using only a subset, M, of the available sectors, say the the five highest ranked sectors in Table 3. This problem pertains, for example, to the design of surveys that are meant to track aggregate employment in real time such as those carried out by the Institute for Supply Management, as well as by various Federal Reserve Banks including the Federal Reserve Bank of Richmond.[4] In particular, the question is: Which sectors are the most informative about the state of aggregate employment and should be included in the surveys? To make some headway toward answering this question, let $\widetilde{\Delta\mathbf{e}}_t$ denote the vector of employment growth rates associated with the M sectors such that $\widetilde{\Delta\mathbf{e}}_t = \mathbf{m}\Delta\mathbf{e}_t$, where $\mathbf{m}$ is an $M \times N$ selection matrix. To help track aggregate employment growth, $\mathbf{s}'\Delta\mathbf{e}_t$, we compute the $M \times 1$ vector of weights, $\mathbf{w}$, attached to the different employment growth series in $\widetilde{\Delta\mathbf{e}}_t$ as the orthogonal projection of $\mathbf{s}'\Delta\mathbf{e}_t$ on $\widetilde{\Delta\mathbf{e}}_t$. That is to say, the weights are optimal in the sense of solving a standard least-square problem, $\mathbf{w} = (\mathbf{m}\Sigma_{ee}\mathbf{m}')^{-1}\mathbf{m}\Sigma_{ee}\mathbf{s}$.

Table 4 reports the fraction of aggregate employment growth explained by the (optimally weighted) employment series related to various sector selections in our data set, $\mathbf{w}'\widetilde{\Delta\mathbf{e}}_t$. Strikingly, using only the sectors associated with the highest five $R_i^2(\mathbf{F})$ statistics in Table 3, this particular filtering already helps us explain 88 percent of the variability in aggregate employment growth. Moreover, virtually all of the variability in aggregate employment growth is accounted for by only considering the 30 highest ranked sectors, according to $R_i^2(\mathbf{F})$, out of 544 sectors. It is apparent, therefore, that information concerning movements in aggregate employment growth is concentrated in a small number of sectors. Contrary to conventional wisdom, these sectors are not necessarily those that have the largest weights in aggregate employment nor the most volatile employment growth series. Because aggregate employment growth

[4] Employment numbers are typically released with a one-month lag and revised up to three months after their initial release. In addition, a revision is carried out annually in March.

Figure 5 Distribution of $R_i^2(F)$ in Goods-Producing and Service-Providing Sectors



is almost exclusively driven by common shocks, the factor analysis proves useful precisely because it allows us to identify the individual sectors whose employment growth also moves most closely with these shocks.

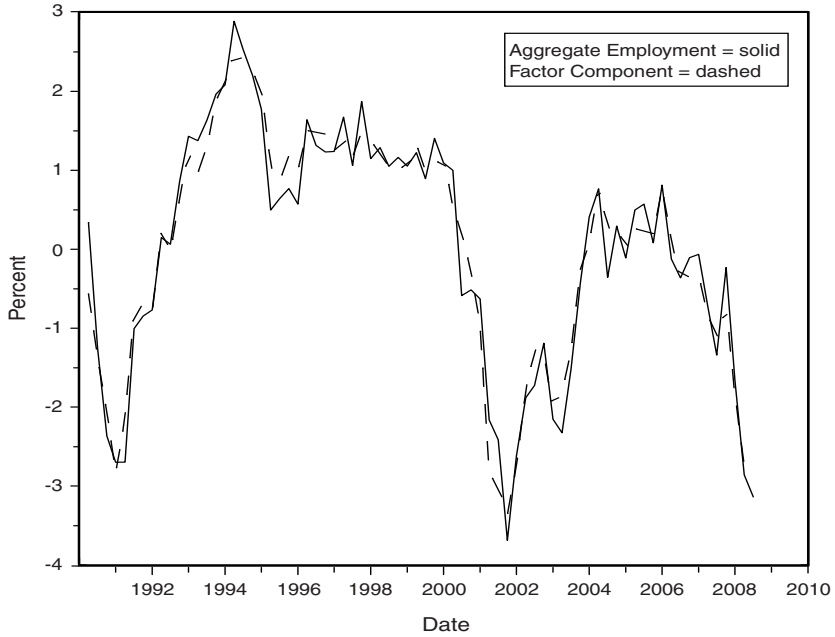
From the exercise we have just carried out, it should be clear that there is much heterogeneity in the way that individual sector employment growth compares to aggregate employment growth over the business cycle. To underscore this point, Figure 5 depicts the breakdown of $R_i^2(\mathbf{F})$ statistics across the main sectors that make up total goods and services separately. Differentiating between goods-producing and service-providing industries, Figure 5 shows that aggregate shocks play a lesser role in driving employment variations in the service sectors relative to the goods-producing sectors. In particular, both the mean and median $R_i^2(\mathbf{F})$ statistics are notably lower in the service-providing

industries than in the goods sectors. That said, it is also the case that there isn't much uniformity within the goods-producing sectors. In particular, we find that employment variations in the durable goods sectors are significantly more subject to common shocks than in the nondurable goods sectors. The median $R_i^2(\mathbf{F})$ statistic is 0.54 in durable goods but only 0.20 in the nondurable goods sectors. In service-providing industries, we find that sector-specific shocks generally play a much greater role in determining employment growth variations. Moreover, the distributions of $R_i^2(\mathbf{F})$ tend to be more similar across service sectors than they are across goods-producing industries. The smallest median $R_i^2(\mathbf{F})$ value across private industries is 0.19, in financial activities, while the largest value is relatively close at 0.29, in the information sector. As indicated above, although employment variations in individual sectors tend to be dominated by sector-specific shocks, these shocks tend to lose their importance in aggregation. To further illustrate this notion, let $\mathbf{s}_j$ denote a vector comprising either the shares corresponding to a particular subsector j of total goods and services, say goods-producing sectors, or zero otherwise. In other words, $\mathbf{s}_j$ effectively selects out employment growth in the different industries making up subsector j. It follows that employment growth in that subsector is given by $\mathbf{s}_j'\Delta\mathbf{e}_t$, and the corresponding factor component in that subsector is $\mathbf{s}_j'\boldsymbol{\lambda}\mathbf{F}_t$. Note that to the degree $\mathbf{s}'\boldsymbol{\lambda}\mathbf{F}_t$ successfully captures the business cycle as it relates to movements in aggregate employment, $\mathbf{s}_j'\boldsymbol{\lambda}\mathbf{F}_t$ captures the analogous concept at a more disaggregated level.

Figures 6 and 7 depict the behavior of $\mathbf{s}_j'\Delta\mathbf{e}_t$ and $\mathbf{s}_j'\boldsymbol{\lambda}\mathbf{F}_t$ for the various sectoral components of our data. Despite the heterogeneity in sectoral employment across sectors as captured by $R_i^2(\mathbf{F})$, the figures suggest that employment growth generally follows movements in the factor component not only at the aggregate level but in subsectors of the economy as well. Of course, at the aggregate level, we have argued that this is to be expected given the results in Table 2 and confirmed in Figure 6. However, we also find that employment growth and the factor component generally move together in goods-producing and service-providing industries separately (Figure 7). In fact, this finding is also true of the main subsectors that make up total goods and services, with the notable exception of government. Perhaps not surprisingly, the latter finding simply reflects the lack of a business cycle component in government services relative to other sectors. Consistent with our earlier findings, our work additionally suggests that employment growth moves less closely with the factor component in service-providing industries than in goods-producing sectors, notably in financial services for instance. On the whole, however, the factor analysis appears to provide a helpful way to track the business cycle as it relates to employment in the broad sectoral components of goods and services.

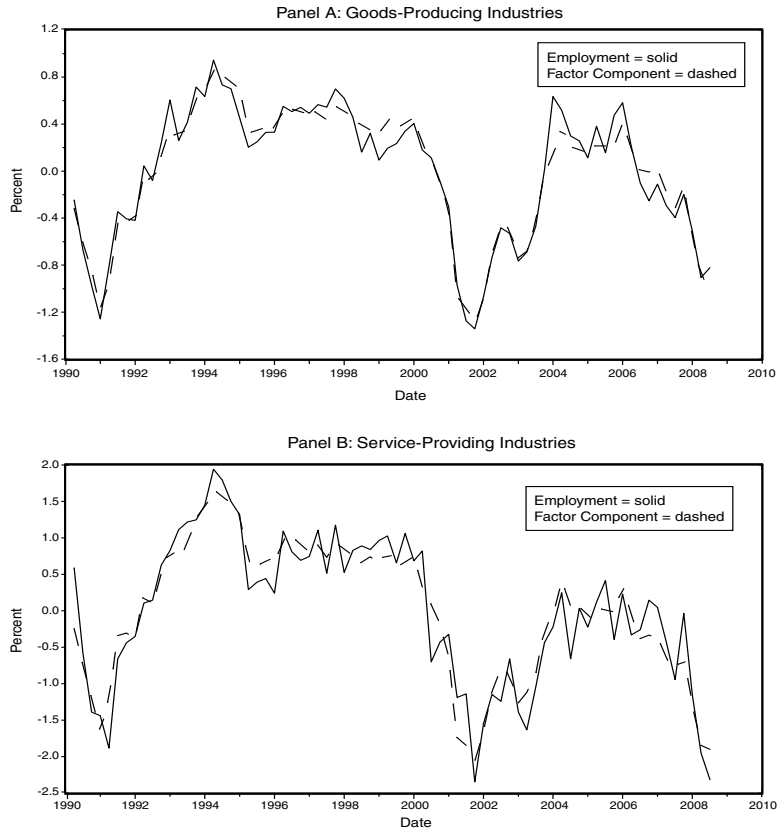
Finally, we note that the factors uncovered in this analysis play substantially different roles in explaining aggregate and sectoral variations in employment. Specifically, even though the first factor alone explains roughly 94

Figure 6 Aggregate Employment Growth and Factor Component



percent of the variation in aggregate employment growth (Table 2), this factor does very little to explain employment growth in particular sectoral components of goods and services. To see this, Figure 8 shows plots of employment growth in natural resources and mining, as well as education and health services, against the factor component using one, two, and three factors. In the first row of Figure 8, we see unambiguously that, despite accounting for the bulk of the variations in aggregate employment, the first factor does very little to capture employment variations in either of the sectors. The correlation between the factor component and employment growth is virtually nil at 0.03 in natural resources and mining and 0.08 in education and health services. In sharp contrast, this correlation jumps to 0.57 in education and health once the second factor is included, and to 0.77 in natural resources and mining once the third factor is included. Note, in particular, that the second factor does little to capture employment growth in natural resources and mining, and it is the third factor alone that helps capture business cycle movements in employment in that sector. In that sense, the Bai and Ng (2002) ICP1 and ICP2 estimators help identify factors that not only explain aggregate employment variations but also account for employment movements at a more disaggregated level.

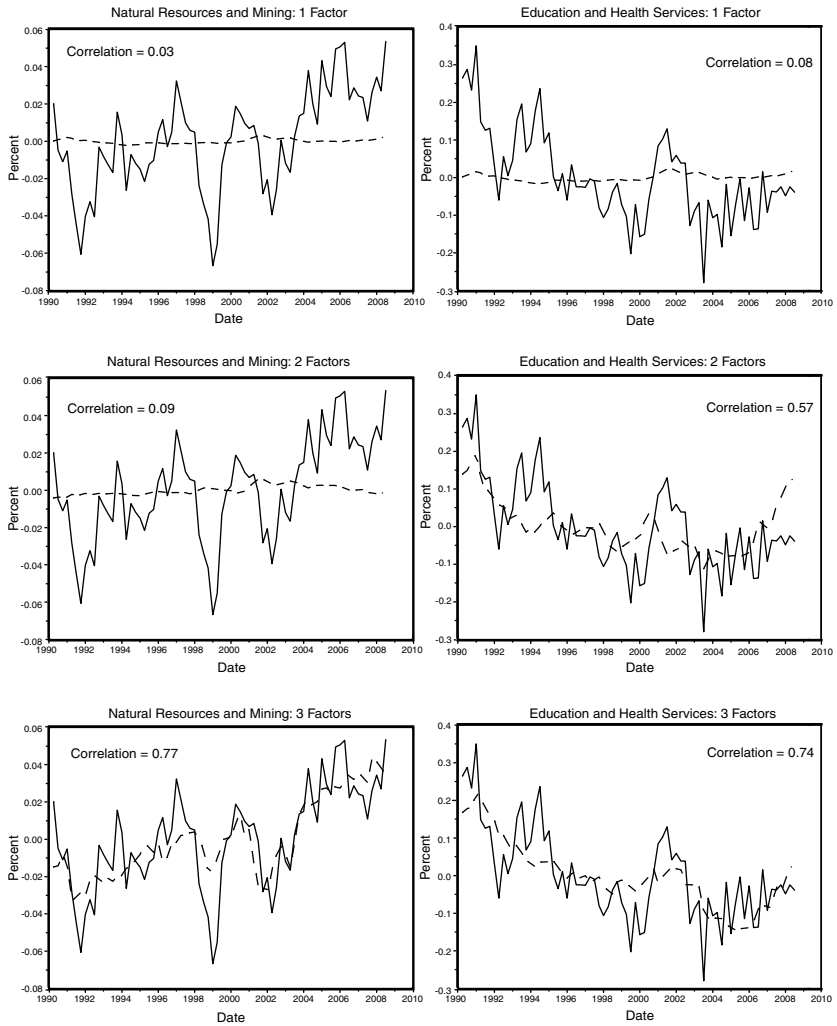
Figure 7 Employment Growth and Factor Component in Goods and Services



4. CONCLUSIONS

In the standard neoclassical one-sector growth model, fluctuations in the representative agent's circumstances are largely determined by shocks to aggregate total factor productivity. This notion is developed, for example, in work going as far back as King, Plosser, and Rebelo (1988). The assumption of a representative agent stands in for a potentially more complicated world populated by heterogenous agents, but where homothetic preferences and complete markets justify focusing on the average agent. Alternatively, we can also think of the representative agent framework as approximating a world in which all agents are essentially identical and affected in the same way by shocks to the economic environment. Under the latter interpretation, a boom in the course of a business cycle characterizes a situation in which "a rising tide lifts all

Figure 8 Contribution of Individual Factors in Explaining Sectoral Employment Growth



boats," and vice versa in the case of a recession. Put another way, idiosyncratic shocks play no role in determining agents' outcomes. More importantly, when individual agents' fortunes are driven mainly by common shocks, the significance of market incompleteness and the importance of insurance considerations tend to vanish since there is no scope for diversifying idiosyncratic shocks away.

Using factor analytic methods, this article documents instead significant differences in employment variations across sectors. In some industries, notably in goods production, variations in employment growth are dominated by aggregate shocks so that these sectors are particularly sensitive to the business cycle. In other industries, in particular some service-providing industries, employment movements are virtually unrelated to aggregate shocks and instead result almost exclusively from sector-specific shocks. The analysis, therefore, suggests that agents working in different sectors of the U.S. economy are affected in very different ways by shocks to the economic environment. Moreover, it underscores the potential importance of market incompleteness and mitigates the usefulness of representative agent models in determining the welfare costs of business cycles.

APPENDIX

This Appendix gives a brief description of the Principle Component (PC) problem based on the discussion in Johnston (1984). See that reference for a more detailed presentation of the problem and its implications.

As described in the main text, suppose we have (demeaned) employment growth observations across N sectors over T time periods summarized in an $N \times T$ matrix, X. In that way, $\Delta \mathbf{e}_t$ in the text is a typical column of X. The nature of the PC problem is to capture the degree of co-movement across these N sectors in a simple and convenient way. To this end, the PC problem transforms the Xs into a new set of variables that will be pairwise uncorrelated and of which the first will have maximum possible variance, the second the maximum possible variance among those uncorrelated with the first, and so on.

Let

$$F_1' = X'\lambda_1$$

denote the first such variable where λ_1 and F_1' are $N \times 1$ and $T \times 1$ vectors, respectively. In other words, F_1' is a linear combination of the elements of X across sectors. The sum of squares of F_1 is

$$F_1 F_1' = \lambda_1' \Sigma_{XX} \lambda_1, \tag{8}$$

where $\Sigma_{XX} = XX'$ represents the variance-covariance matrix (when divided by T) of employment growth rates across sectors. We wish to choose the weights λ_1 to maximize $F_1 F_1'$, but some constraint must evidently be imposed on λ_1 to prevent the sum of squares from being made infinitely large. Thus, a

convenient normalization is to set

$$\lambda_1'\lambda_1 = 1.$$

The PC problem may now be stated as

$$\max_{\lambda_1} \lambda_1'\Sigma_{XX}\lambda_1 + \mu_1(1 - \lambda_1'\lambda_1),$$

where μ_1 is a Lagrange multiplier. Using the fact that Σ_{XX} is a symmetric matrix, the first-order condition associated with this problem is

$$2\Sigma_{xx}\lambda_1 - 2\mu_1\lambda_1 = 0.$$

Thus, it follows that

$$\Sigma_{xx}\lambda_1 = \mu_1\lambda_1.$$

In other words, the weights λ_1 are given by an eigenvector of Σ_{xx} with corresponding eigenvalue μ_1. Observe that when λ_1 is chosen in this way, the sum of squares in (8) reduces to

$$\lambda_1'\Sigma_{XX}\lambda_1 = \lambda_1'\mu_1\lambda_1 = \mu_1.$$

Therefore, our choice of λ_1 must be the eigenvector associated with the largest eigenvalue of Σ_{XX}. The first principle component of X is then F_1.

Now, let us define the next principle component of X as $F_2' = X'\lambda_2$. Similar to the choice of λ_1 we have just described, the problem is to choose the weights λ_2 so as to maximize $\lambda_2'\Sigma_{XX}\lambda_2$ subject to $\lambda_2'\lambda_2 = 1$. In addition, however, because we want the second principle component to capture co-movement that is not already reflected in the first principle component, we impose the further restriction $\lambda_2'\lambda_1 = 0$. This last restriction ensures that F_2 will be uncorrelated with F_1.

The problem associated with the second principle component may then be stated as

$$\max_{\lambda_2} \lambda_2'\Sigma_{XX}\lambda_2 + \mu_2(1 - \lambda_2'\lambda_2) + \phi\lambda_2'\lambda_1.$$

The corresponding first-order condition is

$$2\Sigma_{xx}\lambda_2 - 2\mu_2\lambda_2 + \phi\lambda_1 = 0.$$

Pre-multiplying this last equation by λ_1' gives

$$2\lambda_1'\Sigma_{XX}\lambda_2 - 2\mu_2\lambda_1'\lambda_2 + \phi\lambda_1'\lambda_1 = 0,$$

or

$$\phi = 0,$$

since $\lambda_1'\lambda_1 = 1$, $\lambda_1'\Sigma_{XX} = \mu_1\lambda_1'$, and $\lambda_1'\lambda_2 = 0$. Therefore, we have that the weights λ_2 must satisfy

$$\Sigma_{xx}\lambda_2 = \mu_2\lambda_2,$$

and, in particular, should be chosen as the eigenvector associated with the second largest eigenvalue of Σ_{XX}.

Proceeding in this way, suppose we find the first k principle components of X. We can arrange the weights $\lambda_1, \lambda_2, ..., \lambda_k$ in the $N \times k$ orthogonal matrix

$$\Lambda_k = [\lambda_1, \lambda_2, ..., \lambda_k].$$

Furthermore, the general PC problem may then be described as finding the $T \times k$ matrix of components, $F' = X'\Lambda_k$, such that Λ_k solves

$$\max_{\Lambda_k} \ \Lambda_k' \Sigma_{XX} \Lambda_k \text{ subject to } \Lambda_k' \Lambda_k = I_k. \tag{9}$$

Now, consider the approximate factor model (2) in the text written in matrix form,

$$X = \Lambda_k F + u,$$

where X is $N \times T$, Λ_k is a $N \times k$ matrix of factor loadings, F is a $k \times T$ matrix of latent factors, and u is $N \times T$. One can then show that solving the constrained least-square problem,

$$\min_{\{F_1\}_{t=1}^T, ..., \{F_k\}_{t=1}^T, \Lambda_k} \sum_{t=1}^{T} (X_t - \Lambda_k F_t)'(X_t - \Lambda_k F_t) \text{ subject to } \Lambda_k' \Lambda_k = I_k,$$

is equivalent to solving the general principle component problem (9) we have just described (see Stock and Watson 2002).

REFERENCES

Bai, Jushan, and Serena Ng. 2002. "Determining the Number of Factors in Approximate Factor Models." *Econometrica* 70 (January): 191–221.

Foerster, Andrew, Pierre-Daniel Sarte, and Mark W. Watson. 2008. "Sectoral vs. Aggregate Shocks: A Structural Factor Analysis of Industrial Production," Working Paper 14389. Cambridge, Mass.: National Bureau of Research (October).

Forni, Mario, and Lucrezia Reichlin. 1998. "Let's Get Real: A Factor Analytical Approach to Disaggregated Business Cycle Dynamics." *Review of Economic Studies* 65 (July): 453–73.

Gabaix, Xavier. 2005. "The Granular Origins of Aggregate Fluctuations." Manuscript, Massachusetts Institute of Technology.

Johnston, Jack. 1984. *Econometric Methods*, third edition. New York: McGraw-Hill Book Company.

King, Robert, Charles Plosser, and Sergio Rebelo. 1988. "Production, Growth and Business Cycles: The Basic Neoclassical Model." *Journal of Monetary Economics* 21: 195–232.

Quah, Danny, and Thomas J. Sargent. 1993. "A Dynamic Index Model for Large Cross Sections." In *Business Cycles, Indicators and Forecasting*, edited by J. H. Stock and M. W. Watson. Chicago: University of Chicago Press for the NBER, 285–310.

Shea, John. 2002. "Complementarities and Comovements." *Journal of Money, Credit, and Banking* 34 (May): 412–34.

Stock, J. H., and M. W. Watson. 2002. "Forecasting Using Principal Components from a Large Number of Predictors." *Journal of the American Statistical Association* 97 (December): 1,167–79.

Economic Quarterly—Volume 95, Number 4—Fall 2009—Pages 357–382

Inventories and Optimal Monetary Policy

Thomas A. Lubik and Wing Leong Teo

It has long been recognized that inventory investment plays a large role in explaining fluctuations in real gross domestic product (GDP), although it makes up only a small fraction of it. Blinder and Maccini (1991) document that in a typical recession in the United States, the fall in inventory investment accounts for 87 percent of the decline in output despite being only one half of 1 percent of real GDP. A lot of research has been trying to explain how this seemingly insignificant component of GDP has such a disproportionate role in business cycle fluctuations.[1] However, surprisingly few studies have focused on the conduct of monetary policy when firms can invest in inventories. In this article we attempt to fill this gap by investigating how inventory investment affects the design of optimal monetary policy.

We employ the simple New Keynesian model that has become the benchmark for analyzing monetary policy from both a normative and a positive perspective. We introduce inventories into the model by assuming that the inventory stock facilitates sales, as suggested in Bils and Kahn (2000). We first establish that the dynamics, and therefore the monetary transmission mechanism, differ between the models with and without inventories for a given behavior of the monetary authority. Monetary policy is then endogenized by assuming that policymakers solve an optimal monetary policy problem.

First, we compute the optimal Ramsey policy. A Ramsey planner maximizes the welfare of the agents in the economy by taking into account the

We are grateful to Andreas Hornstein, Pierre Sarte, Alex Wolman, and Nadezhda Malysheva, whose comments greatly improved the paper. Lubik is a senior economist at the Federal Reserve Bank of Richmond. Teo is an assistant professor at National Taiwan University. Lubik wishes to thank the Department of Economics at the University of Adelaide, where parts of this research were conducted, for their hospitality. The views expressed in this paper are those of the authors and should not necessarily be interpreted as those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mails: thomas.lubik@rich.frb.org; wlteo@ntu.edu.tw.

[1] See Ramey and West (1999) and Khan (2003) for extensive surveys of the literature.

private sector's optimality conditions. In doing so, the planner chooses a socially optimal allocation. While this does not necessarily bear any relationship to the typical conduct of monetary policymakers, it provides a useful benchmark. Subsequently, we study optimal policy when the planner is constrained to implement simple rules. That is, we specify a set of rules that lets the policy instrument (the nominal interest rate) respond to target variables such as the inflation rate and output. The policymaker chooses the respective response coefficients that maximize welfare. Optimal rules of this kind may be preferable to Ramsey plans from an actual policymaker's perspective since they can be operationalized and are easier to communicate to the public.

Our most interesting but surprising finding is that Ramsey-optimal monetary policy deviates from full inflation stabilization in our model with inventories. This stands in contrast to the standard New Keynesian model. In the New Keynesian model, perfectly stable inflation is optimal since movements in prices represent deadweight costs to the economy. Introducing inventories potentially modifies that basic calculus for the following reasons. First, we assume that a firm's inventory holdings are relevant for its sales only in relative terms, that is, when they deviate from the aggregate inventory stock. This presents an externality, which a Ramsey planner may want to address. Second, inventories change the economy's propagation mechanism as they allow firms to smooth sales over time with concomitant effects on consumption; that is, output and consumption need no longer coincide, which has a similar effect as capital in that it provides future consumption opportunities. Changes in prices serve as the equilibrating mechanism for the competing goals of reducing consumption volatility and avoiding price adjustment costs. The inventory specification therefore contains something akin to an inflation-output tradeoff. Consequently, the optimal policy no longer fully stabilizes inflation. The second important finding concerns the efficacy of implementing simple rules. Similar to most of the optimal policy literature, we show that simple rules can come exceedingly close to the socially optimal Ramsey policy in welfare terms.

Our article relates to two literatures. First, the amount of research on optimal monetary policy in the New Keynesian framework is very large already, and we do not have much to contribute conceptually to the modeling of optimal policy. Schmitt-Grohé and Uribe (2007) is a recent important and comprehensive contribution. A main conclusion from this literature is that optimal monetary policy will choose to almost perfectly stabilize inflation. In environments with various nominal and real distortions, this policy prescription becomes slightly modified, but nevertheless perseveres. We thus contribute to the optimal policy literature by demonstrating that the results carry over to a framework with another, previously unconsidered modification to the basic framework in the form of inventories.

The study of inventory investment has a long pedigree, to which we cannot do full justice here. Much of the earlier literature, as surveyed in Blinder and Maccini (1991), was concerned with identifying the determinants of inventory investment, such as aggregate demand and expectations thereof, or the opportunity costs of holding inventories. Most work in this area was largely empirical using semi-structural economic models, with West (1986) being a prime example.[2] Almost in parallel to this more explicitly empirical literature, inventories were introduced into real business cycle models. The seminal article by Kydland and Prescott (1982) introduces inventories directly into the production function. More recent contributions include Christiano (1988), Fisher and Hornstein (2000), and Khan and Thomas (2007). The latter two articles especially build a theory of a firm's inventory behavior on the microfoundation of an S-s environment. The focus of these articles is on the business cycle properties of inventories, in particular the high volatility of inventory investment relative to GDP and the countercylicality of the inventory-sales ratio, both of which are difficult to match in typical inventory models. In an important article, Bils and Kahn (2000) demonstrate that time-varying and countercyclical markups are crucial for capturing this co-movement pattern.

This insight lends itself to considering inventory investment within a New Keynesian framework since it features interplay between marginal cost, inflation, and monetary policy, which might therefore be a source of inventory fluctuations.[3] Recently, several articles have introduced inventories into New Keynesian models. Jung and Yun (2005) and Boileau and Letendre (2008) both study the effects of monetary policy from a positive perspective. The former combines Calvo-type price setting in a monopolistically competitive environment with the approach to inventories as introduced by Bils and Kahn (2000). The use of the Calvo approach to modeling nominal rigidity allows these authors to discuss the importance of strategic complementarities in price setting. Boileau and Letendre (2008), on the other hand, compare various approaches to introducing inventories in a sticky-price model. This article is differentiated from those contributions by its focus on the implications of inventories as a transmission mechanism for optimal monetary policy.

The rest of the article is organized as follows. In the next section we develop our New Keynesian model with inventories. Section 2 analyzes the differences between the standard New Keynesian model and our specification with inventories. We calibrate both models and compare their implications for business cycle fluctuations. We present the results of our policy exercises in

[2] A more recent example of applying structural econometric techniques to partial equilibrium inventory models is Maccini and Pagan (2008).

[3] Incidentally, Maccini, Moore, and Schaller (2004) find that an inventory model with regime switches in interest rates is quite successful in explaining inventory behavior despite much previous empirical evidence to the contrary. The key to this result is the exogenous shift in interest rate regimes, which lines up with breaks in U.S. monetary policy.

Section 3, which also includes a robustness analysis with respect to changes in the parameterization. Section 4 concludes with a brief discussion of the main results and suggestions for future research.

1. THE MODEL

We model inventories in the manner of Bils and Kahn (2000) as a mechanism for facilitating sales. When firms face unexpected demand, they can simply draw down their stock of previously produced goods and do not have to engage in potentially more costly production. This inventory specification is embedded in an otherwise standard New Keynesian environment. There are three types of agents: monopolistically competitive firms, a representative household, and the government. Firms face price adjustment costs and use labor for the production of finished goods, which can be sold to households or added to the inventory. Households provide labor services to the firms and engage in intertemporal consumption smoothing. The government implements monetary policy.

Firms

The production side of the model consists of a continuum of monopolistically competitive firms, indexed by $i \in [0, 1]$. The production function of a firm i is given by

$$y_t(i) = z_t h_t(i), \tag{1}$$

where $y_t(i)$ is output of firm i, $h_t(i)$ is labor hours used by firm i, and z_t is aggregate productivity. We assume that it evolves according to the exogenous stochastic process

$$\ln z_t = \rho_z \ln z_{t-1} + \varepsilon_{zt}, \tag{2}$$

where ε_{zt} is an i.i.d. innovation.

We introduce inventories into the model by assuming that they facilitate sales as suggested by Bils and Kahn (2000).[4] In their partial equilibrium framework, they posit a downward-sloping demand function for a firm's product that shifts with the level of inventory available. As shown by Jung and Yun (2005), this idea can be captured in a New Keynesian setting with monopolistically competitive firms by introducing inventories directly into the

[4] This approach is consistent with a stockout avoidance motive. Wen (2005) shows that it explains the fluctuations of inventories at different cyclical frequencies better than alternative theories.

Dixit-Stiglitz aggregator of differentiated products:

$$s_t = \left(\int_0^1 \left(\frac{a_t(i)}{a_t} \right)^{\frac{\mu}{\theta}} s_t(i)^{(\theta-1)/\theta} \, di \right)^{\theta/(\theta-1)}, \tag{3}$$

where s_t are aggregate sales; $s_t(i)$ are firm-specific sales; a_t and $a_t(i)$ are, respectively, the aggregate and firm-specific stocks of goods available for sales; $\theta > 1$ is the elasticity of substitution between differentiated goods; and $\mu > 0$ is the elasticity of demand with respect to the relative stock of goods. Holding inventories helps firms to generate greater sales at a given price since they can rely on the stock of previously produced goods when, say, demand increases. Note, however, that a firm's inventory matters only to the extent that it exceeds the aggregate level. In a symmetric equilibrium, having inventories does not help a firm to make more sales, but it affects the firm's optimality condition for inventory smoothing.

Cost minimization implies the following demand function for sales of good i:

$$s_t(i) = \left(\frac{a_t(i)}{a_t} \right)^{\mu} \left(\frac{P_t(i)}{P_t} \right)^{-\theta} s_t, \tag{4}$$

where $P_t(i)$ is the price of good i, and P_t is the price index for aggregate sales s_t:

$$P_t = \left(\int_0^1 \left(\frac{a_t(i)}{a_t} \right)^{\mu} P_t(i)^{1-\theta} \, di \right)^{1/(1-\theta)}. \tag{5}$$

A firm's sales are thus increasing in its relative inventory holdings and decreasing in its relative price. The inventory term can alternatively be interpreted as a taste shifter, which firms invest in to capture additional demand (see Kryvtsov and Midrigan 2009). Finally, the stock of goods available for sales $a_t(i)$ evolves according to

$$a_t(i) = y_t(i) + (1-\delta)\left(a_{t-1}(i) - s_{t-1}(i)\right), \tag{6}$$

where $\delta \in (0, 1)$ is the rate of depreciation of the inventory stock. It can also be interpreted as the cost of carrying the inventory over the period.

Each firm faces quadratic costs for adjusting its price relative to the steady state gross inflation rate π: $\frac{\phi}{2}\left(\frac{P_t(i)}{\pi P_{t-1}(i)} - 1\right)^2 s_t$, with $\phi > 0$, and $\pi \geq 1$, the steady state gross inflation rate. Note that the costs are measured in units of aggregate sales instead of output since s_t is the relevant demand variable in the model with inventories. Firm i's intertemporal profit function is then given by

$$E_t \sum_{\tau=0}^{\infty} \rho_{t,t+\tau} \left[\frac{P_{t+\tau}(i)\, s_{t+\tau}(i)}{P_{t+\tau}} - \frac{W_{t+\tau} h_{t+\tau}(i)}{P_{t+\tau}} - \frac{\phi}{2}\left(\frac{P_{t+\tau}(i)}{\pi P_{t+\tau-1}(i)} - 1 \right)^2 s_{t+\tau} \right], \tag{7}$$

where W_t is the nominal wage and $\rho_{t,t+\tau}$ is the aggregate discount factor that a firm uses to evaluate profit streams.

Firm i chooses its price, $P_t(i)$, labor input, $h_t(i)$, and stock of goods available for sales, $a_t(i)$, to maximize its expected intertemporal profit (7), subject to the production function (1), the demand function (4), and the law of motion for $a_t(i)$ (6). The first order conditions are

$$\phi\left(\frac{P_t(i)}{\pi P_{t-1}(i)}-1\right)\frac{s_t}{\pi P_{t-1}(i)}=(1-\theta)\frac{s_t(i)}{P_t}$$
$$+E_t\rho_{t,t+1}\left[\phi\left(\frac{P_{t+1}(i)}{\pi P_t(i)}-1\right)\frac{s_{t+1}P_{t+1}(i)}{\pi P_t^2(i)}+(1-\delta)\theta\frac{s_t(i)}{P_t(i)}mc_{t+1}(i)\right] \quad (8)$$

$$\frac{W_t}{P_t}=z_t mc_t(i), \quad (9)$$

and

$$mc_t(i)=\mu\frac{P_t(i)}{P_t}\frac{s_t(i)}{a_t(i)}+(1-\delta)\left(1-\mu\frac{s_t(i)}{a_t(i)}\right)E_t\rho_{t,t+1}mc_{t+1}(i), \quad (10)$$

where $mc_t(i)$ is the Lagrange multiplier associated with the demand constraint (4). It can also be interpreted as real marginal cost.

Equation (8) is the optimal price-setting condition in our model with inventories. It resembles the typical optimal price-setting condition in a New Keynesian model with convex costs for price adjustment (e.g., Krause and Lubik 2007), except that marginal cost now enters the optimal pricing condition in expectations because of the presence of inventories. In this model, the behavior of marginal cost, mc, can be interpreted from two different directions. As captured by Equation (9), it is the ratio of the real wage to the marginal product of labor, which in the standard model is equal to the cost of producing an additional unit of output. Alternatively, it is the cost of generating an additional unit of goods available for sale, which can either come out of current production or out of (previously) foregone sales. This in turn reduces the stock of goods available for sales in future periods, which would eventually have to be replenished through future production. This intertemporal tradeoff between current and future marginal cost is captured by Equation (10).

Household

We assume that there is a representative household in the economy. It maximizes expected intertemporal utility, which is defined over aggregate

consumption,[5] c_t, and labor hours, h_t:

$$E_0 \sum_{t=0}^{\infty} \beta^t \left[\zeta_t \ln c_t - \frac{h_t^{1+\eta}}{1+\eta} \right], \tag{11}$$

where $\eta \geq 0$ is the inverse of the Frisch labor supply elasticity.

ζ_t is a preference shock and is assumed to follow the exogenous AR(1) process

$$\ln \zeta_t = \rho_\zeta \ln \zeta_{t-1} + \varepsilon_{\zeta,t}, \tag{12}$$

where $0 < \rho_\zeta < 1$ and $\varepsilon_{\zeta,t}$ is an i.i.d. innovation.

The household supplies labor hours to firms at the nominal wage rate, W_t, and earns dividend income, D_t, (which is paid out of firms' profits) from owning the firms. It can purchase one-period discount bonds, B_t, at a price of $1/R_t$, where R_t is the gross nominal interest rate. Its budget constraint is

$$P_t c_t + B_t/R_t \leq B_{t-1} + W_t h_t + D_t. \tag{13}$$

The first-order conditions for the representative household's utility maximization problem are

$$h_t^\eta = \frac{\zeta_t}{c_t} \frac{W_t}{P_t}, \text{ and} \tag{14}$$

$$\frac{\zeta_t}{c_t} = \beta R_t E_t \left(\frac{\zeta_{t+1}}{c_{t+1}} \frac{P_t}{P_{t+1}} \right). \tag{15}$$

Equation (14) equates the real wage, valued in terms of the marginal utility of consumption, to the disutility of labor hours. Equation (15) is the consumption-based Euler equation for bond holdings.

Government and Market Clearing

In order to close the model, we also need to specify the behavior of the monetary authority. The main focus of the paper is the optimal monetary policy in the New Keynesian model with inventories. In the next section, however, we briefly compare our specification to the standard model without inventories in order to assess whether introducing inventories significantly changes the model dynamics. We do this conditional for a simple, exogenous interest rate feedback rule that has been used extensively in the literature:

$$\widetilde{R}_t = \rho \widetilde{R}_{t-1} + \psi_1 \widetilde{\pi}_t + \psi_2 \widetilde{y}_t + \varepsilon_{R,t}, \tag{16}$$

[5] Consumption can be thought of as a Dixit-Stiglitz aggregate, as is typical in New Keynesian models. We abstract from this here for ease of exposition.

where a tilde over a variable denotes its log deviation from its deterministic steady state. ψ_1 and ψ_2 are monetary policy coefficients and $0 < \rho < 1$ is the interest smoothing parameter. $\varepsilon_{R,t}$ is a zero mean innovation with constant variance; it is often interpreted as a monetary policy implementation error. Finally, we impose a symmetric equilibrium, so that the firm-specific indices, i, can be dropped. In addition, we assume that bonds are in zero net supply, $B_t = 0$. Market clearing in the goods market requires that consumption, together with the cost for price adjustment, equals aggregate sales:

$$s_t = c_t + \frac{\phi}{2}\left(\frac{\pi_t}{\pi} - 1\right)^2 s_t. \tag{17}$$

2. ANALYZING THE EFFECTS OF MONETARY POLICY

The main focus of this article is how the introduction of inventories into an otherwise standard New Keynesian framework changes the optimal design of monetary policy. However, we begin by briefly comparing the behavior of the model with and without inventories to assess the changes in the dynamic behavior of output and inflation, given the exogenous policy rule (16). The standard New Keynesian model differs from our model with inventories in the following respects. First, there is no explicit intertemporal tradeoff in terms of marginal cost as in equation (10). This implies, secondly, that the driving term in the Phillips curve (8) is current marginal cost, as defined by equation (9). Finally, in the standard model, consumption, output, sales, and goods available of sales are first-order equivalent. We note, however, that the standard specification is not nested in the model with inventories; that is, the equation system for the latter does not reduce to the former for a specific parameterization.

Calibration

The time period corresponds to a quarter. We set the discount factor, β, to 0.99. Since price adjustment costs are incurred only for deviations from steady-state inflation, its value is irrelevant for first-order approximations of the model's equation system but plays a role when we perform the optimal policy analysis. We therefore set $\pi = 1.0086$ to be consistent with the average post-war, quarter-over-quarter inflation rate. In the baseline calibration, we choose a fairly elastic labor supply and set $\eta = 1$, which is a common value in the literature and corresponds to quadratic disutility of hours worked. We impose a steady-state markup of 10 percent, which implies $\theta = 11$. The price adjustment cost parameter is then calibrated so that $\eta(\theta - 1)/\phi = 0.1$, as in Ireland (2004). This is a typical value for the coefficient on marginal

cost in the standard New Keynesian Phillips curve.[6] The parameters of the monetary policy rule are chosen to be broadly consistent with the empirical Taylor rule literature for a unique equilibrium. That is, ψ_1 and ψ_2 are set to 0.45 and 0, respectively, while the smoothing parameter is set to $\rho = 0.7$. This choice corresponds to an inflation coefficient of $0.45/0.3 = 1.5$ that obeys the Taylor principle. We specify the policy rule in this manner since it allows us to analyze later the effects of inertial and super-inertial rules with $\rho \geq 1$.

The persistence of the technology shock and the preference shock are both set to $\rho_z = \rho_\zeta = 0.95$. The standard deviation of the productivity innovation is then chosen so as to match the standard deviation of HP-filtered U.S. GDP of 1.61 percent. This yields a value of $\sigma_z = 0.005$. We set the standard deviation of the preference shocks at three times the value of the former, which is consistent with empirical estimates from a variety of studies (e.g., Ireland 2004). In the same manner, we choose a standard deviation of the monetary policy shock of 0.003. The parameters related to inventories, μ and δ, are calibrated following Jung and Yun (2005); specifically the elasticity of demand with respect to the stock of goods available for sales is $\mu = 0.37$, while the depreciation rate of the inventory stock is $\delta = 0.01$.

Do Inventories Make a Difference?

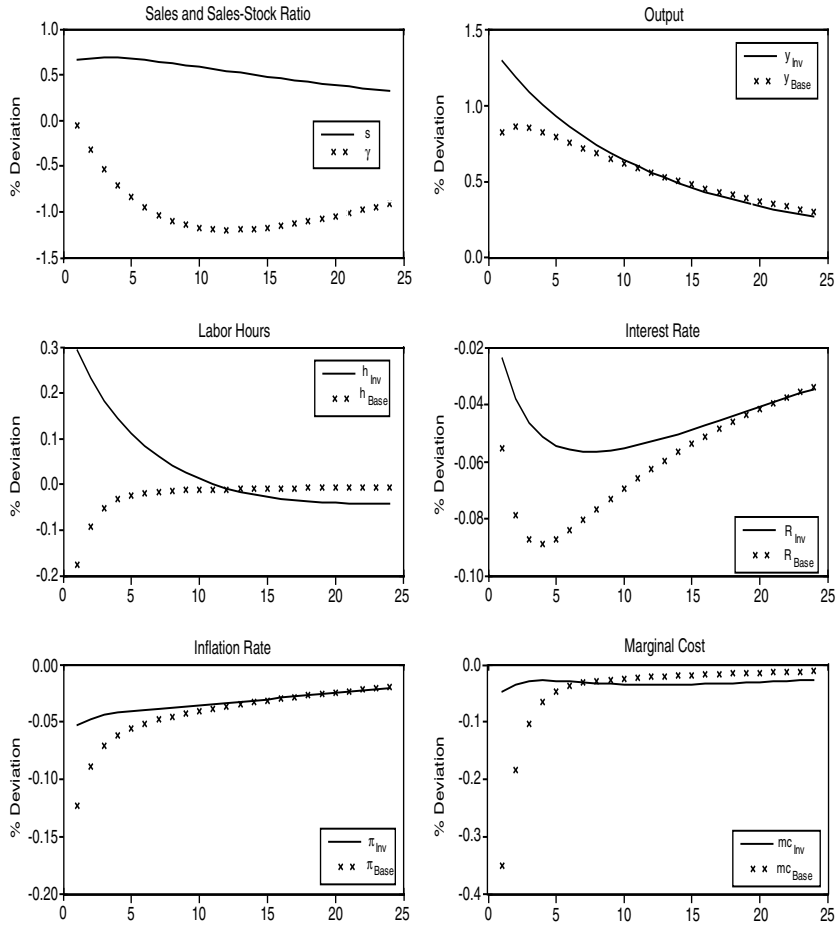
To get an idea how the introduction of inventories changes the model dynamics, we compare the responses of some key variables to technology, preference, and monetary policy shocks for the specification with and without inventories. The impulse responses are found in Figures 1–3, respectively. In the figures, the label "Base" refers to the responses under the specification without inventories, while "Inv" indicates the inventory specification. The key qualitative difference between the two models is the behavior of labor hours. In response to a persistent technology shock, labor increases in the model with inventories, while it falls in the standard New Keynesian model before quickly returning to the steady state.[7] In the New Keynesian model, firms can increase production even when economizing on labor because of the higher productivity level. There is further downward pressure on labor since the productivity shock raises the real wage. Higher output is reflected in a drop in prices, which are drawn out over time due to the adjustment costs, and marginal cost falls strongly.

The presence of inventories, however, changes this basic calculus as firms can use inventories to take advantage of current low marginal cost. With

[6] This value is also consistent with an average price duration of about four quarters in the Calvo model of staggered price adjustment.

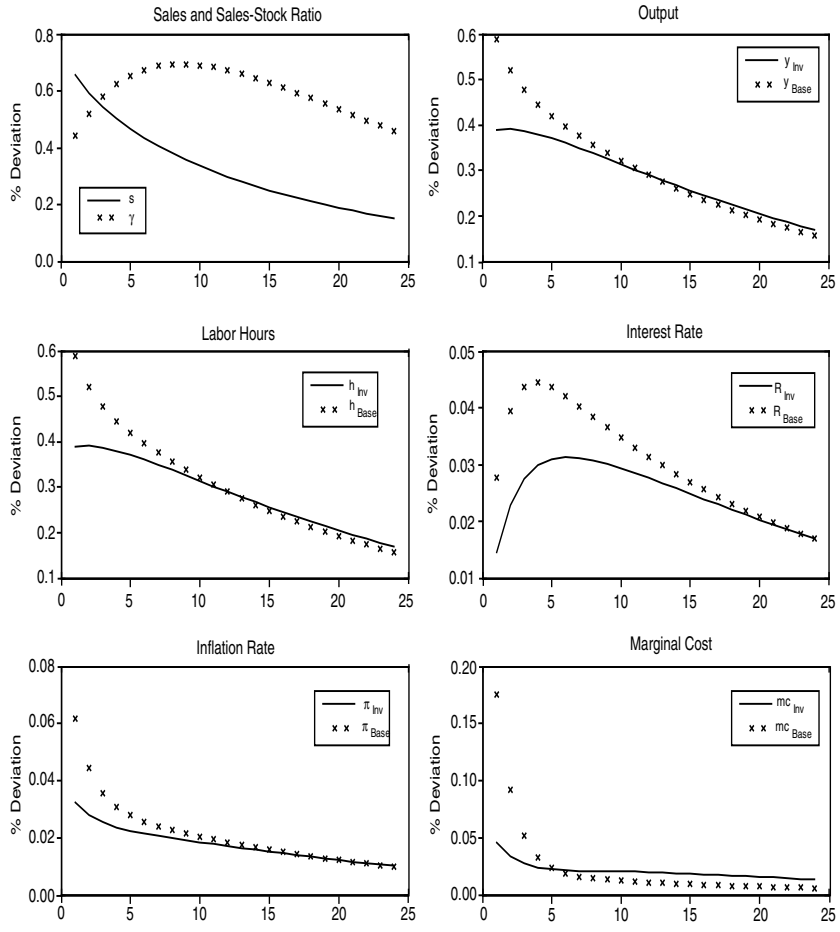
[7] Chang, Hornstein, and Sarte (2009) also emphasize that in the presence of nominal rigidities labor hours can increase in response to a persistent technology shock when firms hold inventories.

Figure 1 Impulse Response Functions to Productivity Shock



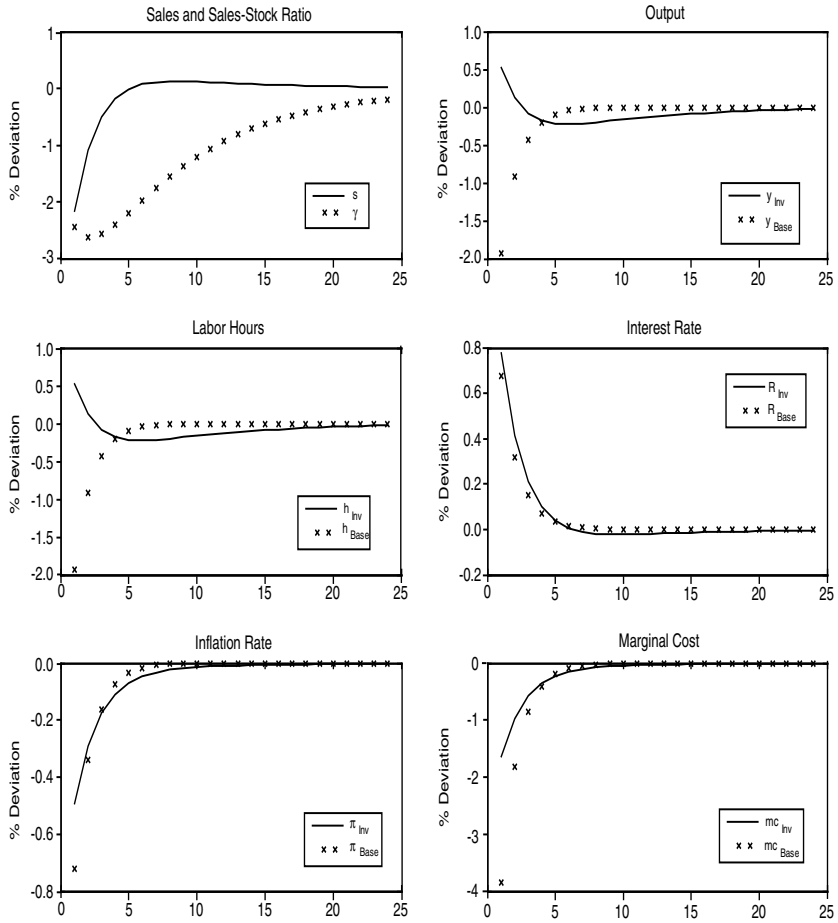
inventory accumulation firms need not sell the additional output immediately, which prompts them to increase labor input. Consequently, output rises by more than in the standard model and the excess production is put in inventory. The stock of goods available for sales thus rises, whereas the sales-to-stock ratio, $\gamma_t \equiv s_t/a_t$, falls. This is also reflected in the (albeit small) fall in marginal cost, which is, however, persistent and drawn out. In other words, firms use inventories to take advantage of current and future low marginal cost. Inflation moves in the same direction as in the standard model, but is much smoother, as the increased output does not have to be priced immediately. This behavior is just the flip side of the smoothing of marginal cost.

Figure 2 Impulse Response Functions to Preference Shock



In response to a preference shock, hours move in the same direction in both models. However, the response with inventories is smaller since firms can satisfy the additional demand out of their inventory holdings, which therefore does not drive up marginal cost as much. Compared to the standard model, firms do not have to resort to increases in price or labor input to satisfy the additional demand. Inventories are thus a way of smoothing revenue over time, which is also consistent with a smoother response of inflation. The dynamics following a contractionary policy shock are qualitatively similar to those of technology shocks in terms of co-movement. Sales in the inventory model fall, but output and hours increase to take advantage of the falling marginal

Figure 3 Impulse Response Functions to Monetary Policy Shock



cost. All series are again noticeably smoother when compared to the standard model.

We now briefly discuss some business cycle implications of the inventory model.[8] Table 1 shows selected statistics for key variables. A notable stylized fact in U.S. data is that production is more volatile than sales. We find that our inventory model replicates this observation in the case of productivity shocks, that is, output is 30 percent more volatile. This implies that

[8] This aspect is discussed more extensively in Boileau and Letendre (2008) and Lubik and Teo (2009).

Table 1 Business Cycle Statistics

Moments	Technology	Preference	Policy	All Shocks
Standard Deviation (%)				
Output	1.61	1.93	0.23	2.52
Sales	1.18	2.37	0.74	2.80
Hours	0.25	1.93	0.23	2.02
Correlation				
(Sales, $\frac{\text{Sales}}{\text{Inventory}}$)	−0.85	0.87	0.51	0.49
($\frac{\text{Sales}}{\text{Stock}}$, Marginal Cost	0.95	0.90	0.72	0.49

consumption, which is equal to sales in our linearized setting, is also less volatile than GDP. The introduction of inventories is thus akin to the modeling of capital and investment in breaking the tight link between output and consumption embedded in the standard New Keynesian model. However, the model has counterfactual implications for the co-movement of inventory variables. Sales are highly negatively correlated with the sales-inventory ratio, whereas in the data the two series co-move slightly positively and are at best close to uncorrelated. This finding can be overturned when either preference or policy shocks are used, both of which imply a strong positive co-movement. However, in the case of policy shocks, sales are counterfactually more volatile than output. When all shocks are considered together, we find that co-movement between the inventory variables are positive, but not unreasonably so, while sales are slightly more volatile than output.

The model also has implications for inflation dynamics. Most notably, inflation is less volatile in the inventory specification than in the standard model. In the New Keynesian model, inflation is driven by marginal cost; hence, the standard model predicts that the two variables are highly correlated. In the data, however, proxies for marginal cost, such as unit labor cost or the labor share, co-move only weakly with inflation. This has been a challenge for empirical studies of the New Keynesian Phillips curve. Our model with inventories may, however, improve the performance of the Phillips curve in two aspects. First, marginal cost smoothing translates into a smoother and thus more persistent inflation path; second, the form and the nature of the driving process in the Phillips curve equation changes, as is evident from equations (8) and (10). The latter equation predicts a relationship between marginal cost and the sales-to-stock ratio, γ, which changes the channel by which marginal cost affects inflation dynamics.[9]

[9] This is further and more formally empirically investigated in Lubik and Teo (2009), who suggest that the inventory channel does not contribute much to explain observed inflation behavior.

We can tentatively conclude that a New Keynesian model with inventories presents a modified set of tradeoffs for an optimizing policymaker. In the standard model optimal policy is such that both consumption and the labor supply should be smoothed and price adjustment costs minimized. In the inventory model, these objectives are still relevant since they affect utility in the same manner, but the channel through which this can be achieved is different. Inventories allow for a smoother adjustment path of inflation, which should help contain the effects of price stickiness, while the consumption behavior depends on the nature of the shocks. We now turn to an analysis of optimal policy with inventories.

3. OPTIMAL MONETARY POLICY

The goal of an optimizing policymaker is to maximize a welfare function subject to the constraints imposed by the economic environment and subject to assumptions about whether the policymaker can commit or not to the chosen action. In this article, we assume that the optimizing monetary authority maximizes the intertemporal utility function of the household subject to the optimal behavior chosen by the private sector and the economy's feasibility constraints. Furthermore, we assume that the policymaker can credibly commit to the chosen path of action and does not re-optimize along the way. We consider two cases. For our benchmark, we assume that the monetary authority implements the Ramsey-optimal policy.[10] We then contrast the Ramsey policy with an optimal policy that is chosen for a generic set of linear rules of the type used in the simulation analysis above.

We can alternatively interpret the policymaker's actions as minimizing the distortions in the model economy. In a typical New Keynesian setup like ours, there are two distortions. The first is the suboptimal level of output generated by the presence of monopolistically competitive firms. The second distortion arises from the presence of nominal price stickiness, as captured by the quadratic price adjustment cost function, which is a deadweight loss to the economy. In the standard model, the optimal policy perfectly stabilizes inflation at the steady-state level. Introducing inventories can change this basic calculus in our model, as the sales-relevant terms are relative inventory holdings that present an externality for a Ramsey planner. We will now investigate whether this additional wedge matters quantitatively for optimal policy.

Welfare Criterion

We use expected lifetime utility of the representative household at time zero, V_0^a, as the welfare measure to evaluate a particular monetary policy

[10] See Khan, King, and Wolman (2003), Levin et al. (2006), and Schmitt-Grohé and Uribe (2007) for wide-ranging and detailed discussions of this concept in New Keynesian models.

regime, a:

$$V_0^a \equiv E_0 \sum_{t=0}^{\infty} \beta^t \left[\zeta_t \ln C_t^a - \frac{\left(h_t^a\right)^{1+\eta}}{1+\eta} \right]. \tag{18}$$

As in Schmitt-Grohé and Uribe (2007), we compute the expected lifetime utility conditional on the initial state being the deterministic steady state for given sequences of optimal choices of the endogenous variables and exogenous shocks. Our welfare measure is in the spirit of Lucas (1987) and expresses welfare as a percentage Θ of steady-state consumption that the household is willing to forgo to be as well off under the steady state as under a given monetary policy regime, a. Θ can then be computed implicitly from

$$\sum_{t=0}^{\infty} \beta^t \left[\zeta \ln \left[\left(1 - \frac{\Theta}{100} \right) c \right] - \frac{h^{1+\eta}}{1+\eta} \right] = V_0^a, \tag{19}$$

where variables without time subscripts denote the steady state of the corresponding variables.[11] Note that a higher value of Θ corresponds to lower welfare. That is, the household would be willing to give up Θ *percent* of steady-state consumption to implement a policy that delivers the same level of welfare as the economy in the absence of any shocks. This also captures the notion that business cycles are costly because they imply fluctuations that a consumption-smoothing and risk-averse agent would prefer not to have.

Optimal Policy

We compute the Ramsey policy by formulating a Lagrangian problem in which the government maximizes the welfare function (18) of the representative household subject to the private sector's first-order conditions and the market-clearing conditions of the economy. The optimality conditions of this Ramsey policy problem can then be obtained by differentiating the Lagrangian problem with respect to each of the endogenous variables and setting the derivatives to zero. This is done numerically by using the Matlab procedures developed by Levin and Lopez-Salido (2004). The welfare function is then approximated around the distorted, non-Pareto-optimal steady state. The source of steady-state distortion is the inefficient level of output due to the presence of monopolistically competitive firms.

In our second optimal policy case, we follow Schmitt-Grohé and Uribe (2007) and consider optimal, simple, and implementable interest rate rules.

[11] We assume that the policymaker chooses the same steady-state inflation rate for all monetary policies that we consider. The steady state of all variables will thus be the same for all policies.

Specifically, we consider rules of the following type:

$$\widetilde{R}_t = \rho \widetilde{R}_{t-1} + \psi_1 E_t \widetilde{\pi}_{t+i} + \psi_2 E_t \widetilde{y}_{t+i}, i = -1, 0, 1. \tag{20}$$

The subscript i indicates that we consider forward-looking ($i = 1$), contemporaneous ($i = 0$), and backward-looking rules ($i = -1$). Following the suggestion in Schmitt-Grohé and Uribe (2007), we focus on values of the policy parameters ρ, ψ_1, and ψ_2 that are in the interval [0, 3]. Note that this rule also allows for the possibility that the interest rate is super-inertial; that is, we assume ρ can be larger than 1. In order to find the constrained-optimal interest rate rule, we search for combinations of the policy coefficients that maximize the welfare criterion. As in Schmitt-Grohé and Uribe (2007), we impose two additional restrictions on the interest rate rule: (i) the rule has to be consistent with a locally unique rational expectations equilibrium; (ii) the interest rate rule cannot violate $2\sigma_R < R$, where σ_R is the unconditional standard deviation of the gross interest rate while R is its steady-state value. The second restriction is meant to approximate the zero bound constraint on the nominal interest rate.[12]

Ramsey-Optimal Policy

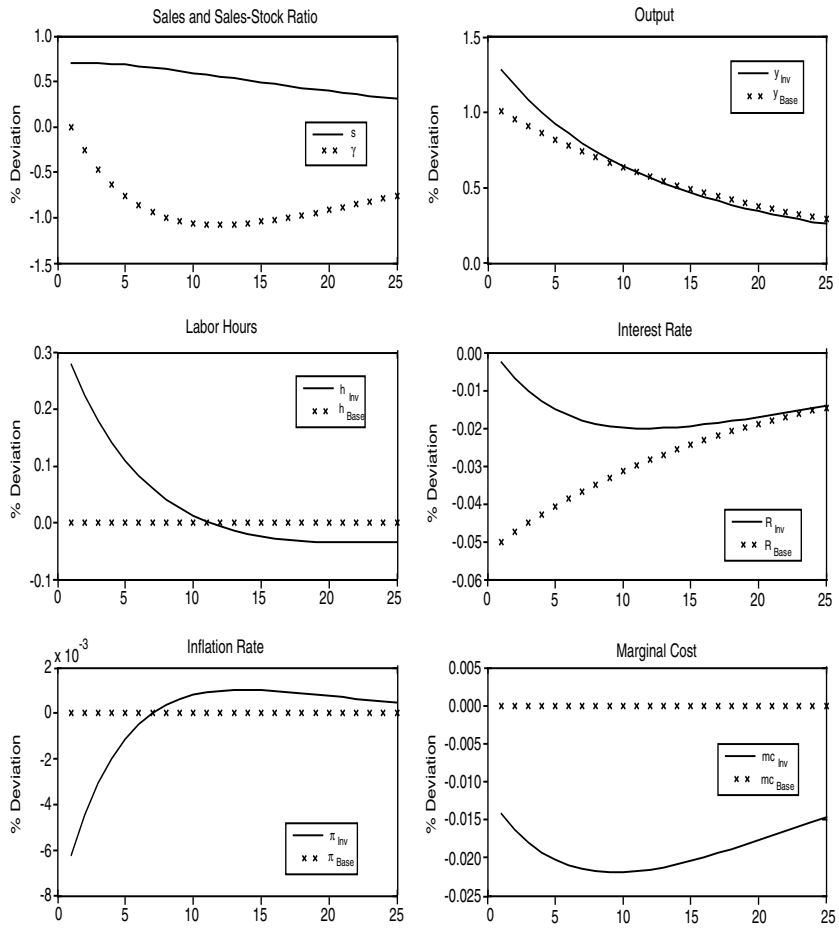
A key feature of the standard New Keynesian setup is that Ramsey-optimal policy completely stabilizes inflation. Price movements represent a deadweight loss to the economy because of the existence of adjustment costs.[13] An optimizing planner would, therefore, attempt to remove this distortion. This insight is borne out by the impulse response functions for the standard model without inventories in Figure 4. Inflation does not respond to the technology shock, nor do labor hours or marginal cost as per the New Keynesian Phillips curve. The path of output simply reflects the effect of increased and persistent productivity. The Ramsey planner takes advantage of the temporarily high productivity and allocates it straight to consumption without feedback to higher labor input or prices. The planner could have reduced labor supply to smooth the time path of consumption. However, this would have a level effect on utility due to lower consumption, positive price adjustment cost via the feedback from lower wages to marginal cost, and increased volatility in hours. The solution to this tradeoff is thus to bear the brunt of higher consumption volatility.

The possibility of inventory investment, however, changes this rationale (see Figure 4). In response to a technology shock, output increases by more

[12] If R is normally distributed, $2\sigma_R < R$ implies that there is a 95 percent chance that R will not hit the zero bound.

[13] In a framework with Calvo price setting, the deadweight loss comes in the form of relative price distortions across firms, which lead to the misallocation of resources.

Figure 4 Impulse Response Functions to Productivity Shock: Ramsey Policy



compared to the model without inventories, while consumption, which is first-order equivalent to sales, rises less. Ramsey-optimal policy can induce a smoother consumption profile by allowing firms to accumulate inventories. Similarly, the planner takes advantage of higher productivity in that he induces the household to supply more labor hours. Inflation is now no longer completely stabilized as the lower increase in consumption leads to an initial decline in inflation. Inventories thus serve as a savings vehicle that allows the planner to smooth out the impact of shocks. The planner incurs price adjustment costs and disutility from initially high labor input. The benefit is a smoother and more prolonged consumption path than would be possible

Table 2 Welfare Costs and Standard Deviations under Ramsey-Optimal Policy

	Technology	Preference	All Shocks
	Panel A: Model without Inventories		
Welfare Cost (Θ)	0.0000	−0.0521	−0.0521
Standard Deviation (%)			
Output	1.60	2.40	2.89
Inflation	0.00	0.00	0.00
Consumption	1.60	2.40	2.89
Labor	0.00	2.40	2.40
	Panel B: Model with Inventories		
Welfare Cost (Θ)	0.000	−0.0529	−0.0529
Standard Deviation (%)			
Output	1.73	2.28	2.86
Inflation	0.02	0.04	0.04
Consumption	1.45	2.60	2.97
Labor	0.24	2.28	2.29
	Panel C: Full Inflation Stabilization		
Welfare Cost (Θ)	0.000	−0.0528	−0.0528
Standard Deviation (%)			
Output	1.73	2.29	2.87
Inflation	0.00	0.00	0.00
Consumption	1.45	2.61	2.99
Labor	0.24	2.29	2.30

without inventories. The model with inventories therefore restores something akin to an output-inflation tradeoff in the New Keynesian framework.

The quantitative differences between the two specifications are small, however. Table 2 reports the welfare costs and standard deviations of selected variables for the two versions of the model under Ramsey-optimal policy. The welfare costs of business cycles in the standard model are vanishingly small when only technology shocks are considered and undistinguishable from the specification with inventories. The standard deviation of inflation is zero for the model without inventories while it is slightly higher for the model with inventories. This is consistent with the evidence from the impulse responses and highlights the differences between the two model specifications. Note also that consumption is less volatile in the model with inventories than in the standard model, which reflects the increased degree of consumption smoothing in the former.[14]

[14] This is consistent with the simulation results reported in Schmitt-Grohé and Uribe (2007) in a model with capital. They also find that full inflation stabilization is no longer optimal since

Figure 5 Impulse Response Functions to Preference Shock: Ramsey Policy

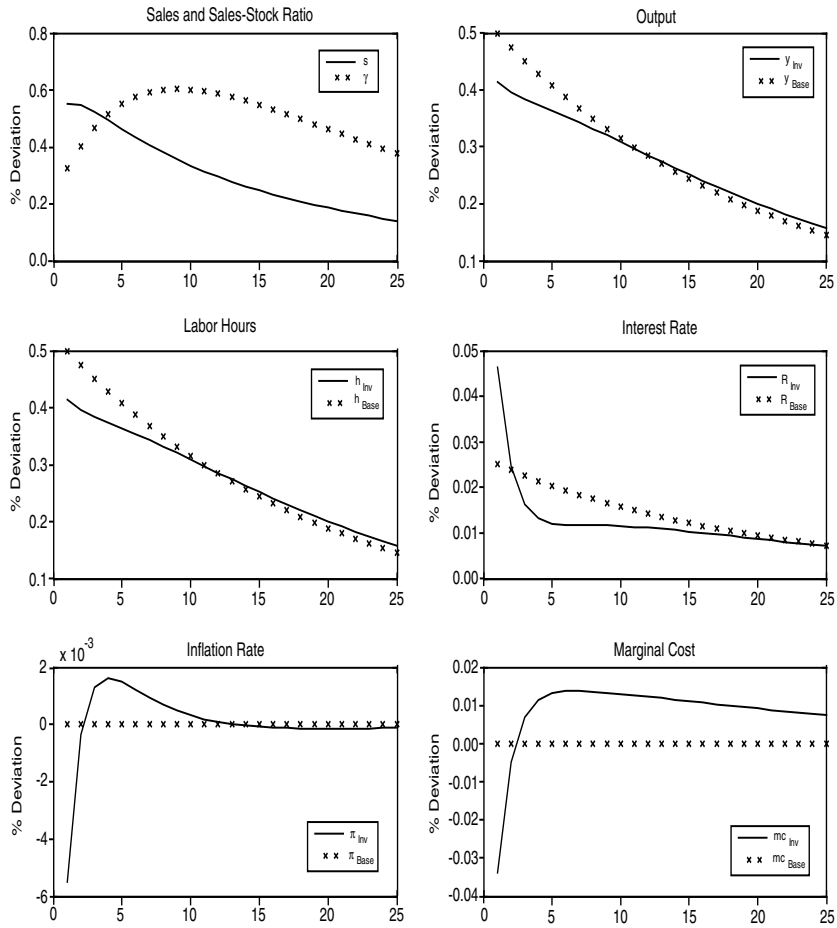


Figure 5 depicts the impulse responses to the preference shock under Ramsey-optimal policy. Inflation and marginal cost are fully stabilized in the standard model, which the planner achieves through a higher nominal interest rate that reduces consumption demand in the face of the preference shock. At the same time, the planner lets labor input go up to meet some of the additional demand. In contrast, Ramsey policy for the inventory model

investment in capital provides a mechanism for smoothing consumption, just as inventory holdings do in our model.

can allow consumption to increase by more since firms can draw on their stock of goods for sale. Consequently, output and labor increase by less for the inventory model. Similarly to the case of the technology shock, optimal policy does not induce complete inflation stabilization as it uses the inventory channel to smooth consumption. This is confirmed by the simulation results in Table 2, which show the Ramsey planner trading off volatility between inflation, consumption, and labor when compared to the standard model.

Interestingly, eliminating business cycles and imposing the steady-state allocation is costly for the planner in the presence of preference shocks that multiply consumption. This is evidenced by the negative entries for the welfare cost in both model specifications. In other words, agents would be willing to pay the planner 0.05 percent of their steady-state consumption *not* to eliminate preference-driven fluctuations. This stems from the fact that, although fluctuations per se are costly in welfare terms for risk-averse agents, they can also induce co-movement between the shocks and other variables that have a level effect on utility. Specifically, preference shocks co-move positively with consumption due to an increase in demand. This positive co-movement is reflected in a positive covariance between these two variables. In our second-order approximation to the welfare functions, this overturns the negative contribution to welfare from consumption volatility.

When we consider both shocks together, the differences between the two specifications are not large in welfare terms and with respect to the implications for second moments. Inflation and consumption are more volatile in the inventory version, while labor is less volatile compared to standard specification. We also compare Ramsey-optimal policy with inventories to a policy of fully stabilizing inflation only (as opposed to using the utility-based welfare criterion from above). Panel C of Table 2 shows that the latter is very close to the Ramsey policy. The welfare difference between the two policies is small—less than 0.001 percentage points of steady-state consumption. The effects of inventories can be seen in the slightly higher volatility of consumption and labor under the full inflation stabilization policy. Inventory investment allows the planner to smooth consumption more compared to the standard model, and the mechanism is a change in prices. Although price stability is feasible, the planner chooses to incur an adjustment cost to reduce the volatility of consumption and labor.

Optimal Policy with a Simple and Implementable Rule

Ramsey-optimal policy provides a convenient benchmark for welfare analysis in economic models. However, from the point of view of a policymaker, pursuing a Ramsey policy may be difficult to communicate to the public. It may also not be operational in the sense that the instruments used to implement

the Ramsey policy may not be available to the policymaker. For instance, in a market economy the government cannot simply choose allocations as a Ramsey plan might imply. The literature has therefore focused on finding simple and implementable rules that come close to the welfare outcomes implied by Ramsey policies (see Schmitt-Grohé and Uribe 2007).

Therefore, we investigate the implications for optimal policy conditional on the simple rule (20). Panel A of Table 3 shows the constrained-optimal interest rate rules for the model without inventories with all shocks considered simultaneously. The rule that delivers the highest welfare is a contemporaneous rule, with a smoothing parameter $\rho = 1$ and reaction coefficients on inflation ψ_1 and output ψ_2 of 3 and 0, respectively.[15] This is broadly consistent with the results of Schmitt-Grohé and Uribe (2007), where the constrained-optimal interest rate rule also features interest smoothing and a muted response to output. Without interest rate smoothing the welfare cost of implementing this policy increases, which is exclusively due to a higher volatility of inflation.

On the other hand, the difference between the constrained-optimal contemporaneous rule and the Ramsey policy is small—less than 0.001 percentage points. This confirms the general consensus in the literature that simple rules can come extremely close to Ramsey-optimal policies in welfare terms. The characteristics of constrained-optimal backward-looking and forward-looking rules are similar to the contemporaneous rule, i.e., they also feature full interest smoothing and no output response. The welfare difference between the constrained-optimal contemporaneous rule and the other two rules are also small.

Turning to the model with inventories, we report the results for the constrained-optimal rules in Panel B of Table 3. All rules with interest smoothing deliver virtually identical results but strictly dominate any rule without smoothing. As before, the coefficient on output is zero, while the policymakers implement a strong inflation response. The main difference to the Ramsey outcome is that inflation is slightly less volatile, while output is more volatile. This again confirms the findings in other articles that a policy rule with a fully inertial interest rate and a hawkish inflation response delivers almost Ramsey-optimal outcomes.

Sensitivity Analysis

We now investigate the robustness of our optimal monetary policy results to alternative parameter values. The results of alternative calibrations are reported in Table 4, where we only document results for the rule that comes closest to the Ramsey benchmark. In the robustness analysis, we change

[15] The reader may recall that we restricted the policy coefficients to lie within the interval $[0, 3]$.

Table 3 Optimal Policy with a Simple Rule

	ρ	ψ_1	ψ_2	**Welfare Cost** (Θ)	σ_π	σ_y
			Panel A: Model without Inventories			
Ramsey Policy				−0.0521	0.00	2.89
Optimized Rules						
Contemporaneous ($i = 0$)						
Smoothing	1.00	3.00	0.00	−0.0520	0.04	2.89
No Smoothing	0.00	3.00	0.00	−0.0499	0.28	2.89
Backward ($i = -1$)						
Smoothing	1.00	3.00	0.00	−0.0520	0.05	2.89
No Smoothing	0.00	3.00	0.00	−0.0501	0.27	2.90
Forward ($i = 1$)						
Smoothing	1.00	3.00	0.00	−0.0518	0.08	2.90
No Smoothing	0.00	3.00	0.00	−0.0496	0.30	2.90
			Panel B: Model with Inventories			
Ramsey Policy				−0.0529	0.04	2.86
Optimized Rules						
Contemporaneous ($i = 0$)						
Smoothing	1.00	3.00	0.00	−0.0528	0.01	2.87
No Smoothing	0.00	3.00	0.00	−0.0518	0.20	2.87
Backward ($i = -1$)						
Smoothing	1.00	3.00	0.00	−0.0528	0.02	2.87
No Smoothing	0.00	3.00	0.00	−0.0518	0.19	2.87
Forward ($i = 1$)						
Smoothing	1.00	3.00	0.00	−0.0528	0.02	2.87
No Smoothing	0.00	3.00	0.00	−0.0517	0.20	2.87

one parameter at a time while holding all other parameters at their benchmark values. The overall impression is that in all alternative calibrations the optimal simple rule comes close to the Ramsey policy, and that the relative welfare rankings for the individual rules established in the benchmark calibration are unaffected. Specifically, inertial rules tend to dominate rules with a lower degree of smoothing.

We first look at the implications of alternative values for the two parameters related to inventories: the elasticity of demand with respect to the stock of goods available for sale, μ, and the depreciation rate of the inventory stock, δ. As in Jung and Yun (2005), we consider the alternative value $\mu = 0.8$. Since sales now respond more elastically to the stock of goods available for sale, the inventory channel becomes more valuable as a consumption-smoothing device and inflation becomes more volatile under a Ramsey policy. The best simple rule has contemporaneous timing and comes very close to the Ramsey policy in terms of welfare. The optimal rule is inertial and strongly reacts to inflation only. The volatility of inflation is lower than under the Ramsey policy and closer to that of the optimally simple rule with the benchmark calibration. This

Table 4 Optimal Policy for the Model with Inventories: Alternative Calibration

	ρ	ψ_1	ψ_2	**Welfare Cost** (Θ)	σ_π	σ_y
			Panel A: $\mu = 0.8$			
Ramsey Policy				−0.0508	0.05	2.88
Contemporaneous ($i = 0$)	1.0	3.0	0.0	−0.0507	0.01	2.89
			Panel B: $\delta = 0.05$			
Ramsey Policy				−0.0557	0.09	2.85
Contemporaneous ($i = 0$)	1.0	3.0	0.0	−0.0553	0.02	2.86
			Panel C: $\eta = 5$			
Ramsey Policy				−0.0193	0.03	1.79
Contemporaneous ($i = 0$)	1.0	3.0	0.0	−0.0190	0.01	1.80
			Panel D: $\theta = 21$			
Ramsey Policy				−0.0539	0.05	2.85
Contemporaneous ($i = 0$)	1.0	3.0	0.0	−0.0537	0.02	2.86

suggests that the response coefficients of the optimal rule are insensitive to changes in elasticity parameter μ, and that the Ramsey planner can exploit the changes in the transmission mechanism in a way that the simple rule misses. The quantitative differences are small, however.

In the next experiment, we increase the depreciation rate of the inventory stock to $\delta = 0.05$. It is at this value that Lubik and Teo (2009) find that the inclusion of inventories has a marked effect on inflation dynamics in the New Keynesian Phillips curve. Panel B of Table 4 shows that the preferred rule is again contemporaneous, but the differences between the alternatives are very small. Interestingly, Ramsey policy leads to a volatility of inflation that is almost an order of magnitude higher than in the benchmark case, which is consistent with the findings in Lubik and Teo (2009).

The benchmark calibration imposed a very elastic labor supply with $\eta = 1$. The results of making the labor supply much more inelastic by setting $\eta = 5$ are depicted in Panel C of the table. For this value, the differences to the benchmark are most pronounced. In particular, the volatility of output declines substantially across the board, which is explained by the difficulty with which firms change their labor input. The best simple rule is contemporaneous, but the differences to the other rules are vanishingly small. Optimal policy again puts strong weight on inflation, with the optimal rule being inertial. Another difference to the benchmark parameterization is that the welfare cost of no interest smoothing is also much bigger for $\eta = 5$.[16] Finally, we also report results for calibration with a lower steady-state markup of 5 percent, which

[16] The welfare cost of no interest smoothing is 0.0088 for $\eta = 5$, while it is 0.0021 for the benchmark parameterization.

corresponds to a value of $\theta = 21$. The qualitative and quantitative results are mostly similar to the benchmark results.

In summary, the results from the benchmark calibration are broadly robust. Under a Ramsey policy full inflation stabilization is not optimal, while the best optimal simple rule exhibits inertial behavior on interest smoothing and a strong inflation response. The welfare differences between alternative calibrations are very small, with the exception of changes in the labor supply elasticity. A less elastic labor supply reduces the importance of the inventory channel to smooth consumption by making it more difficult to adjust employment and output in the face of exogenous shocks.

4. CONCLUSION

We introduce inventories into an otherwise standard New Keynesian model that is commonly used for monetary policy analysis. Inventories are motivated as a way to generate sales for firms. This changes the transmission mechanism of the model, which has implications for the conduct of optimal monetary policy. We emphasize two main findings in the article. First, we show that full inflation stabilization is no longer the Ramsey-optimal policy in the simple New Keynesian model with inventories. While the optimal planner still attempts to reduce inflation volatility to zero since it is a deadweight loss for the economy, the possibility of inventory investment opens up a tradeoff. In our model, production no longer needs to be consumed immediately, but can be put into inventory to satisfy future demand. An optimizing policymaker therefore has an additional channel for welfare-improving consumption smoothing, which comes at the cost of changing prices and deviations from full inflation stabilization. Our second finding confirms the general impression from the literature that simple and implementable optimal rules come close to replicating Ramsey policies in welfare terms.

This article contributes to a growing literature on inventories within the broader New Keynesian framework. However, evidence on the usefulness of including inventories to improve the model's business cycle transmission mechanism is mixed, as we have shown above. Future research may therefore delve deeper into the empirical performance of the New Keynesian inventory model, in particular on how modeling inventories affect inflation dynamics. Jung and Yun (2005) and Lubik and Teo (2009) proceed along these lines. A second issue concerns the way inventories are introduced into the model. An alternative to our setup is to add inventories to the production structure so that instead of smoothing sales, firms can smooth output. Finally, it would be interesting to estimate both model specifications with structural methods and compare their overall fit more formally.

REFERENCES

Bils, Mark, and James A. Kahn. 2000. "What Inventory Behavior Tells Us About Business Cycles." *American Economic Review* 90 (June): 458–81.

Blinder, Alan S., and Louis J. Maccini. 1991. "Taking Stock: A Critical Assessment of Recent Research on Inventories." *Journal of Economic Perspectives* 5 (Winter): 73–96.

Boileau, Martin, and Marc-André Letendre. 2008. "Inventories, Sticky Prices, and the Persistence of Output and Inflation." Manuscript.

Chang, Yongsung, Andreas Hornstein, and Pierre-Daniel Sarte. 2009. "On the Employment Effects of Productivity Shocks: The Role of Inventories, Demand Elasticity and Sticky Prices." *Journal of Monetary Economics* 56 (April): 328–43.

Christiano, Lawrence J. 1988. "Why Does Inventory Investment Fluctuate So Much?" *Journal of Monetary Economics* 21(2/3): 247–80.

Fisher, Jonas D. M., and Andreas Hornstein. 2000. "(S,s) Inventory Policies in General Equilibrium." *Review of Economic Studies* 67 (January): 117–45.

Ireland, Peter N. 2004. "Technology Shocks in the New Keynesian Model." *Review of Economics and Statistics* 86(4): 923–36.

Jung, YongSeung, and Tack Yun. 2005. "Monetary Policy Shocks, Inventory Dynamics and Price-setting Behavior." Manuscript.

Khan, Aubhik. 2003. "The Role of Inventories in the Business Cycle." Federal Reserve Bank of Philadelphia *Business Review* Q3: 38–45.

Khan, Aubhik, and Julia K. Thomas. 2007. "Inventories and the Business Cycle: An Equilibrium Analysis of (S,s) Policies." *American Economic Review* 97: 1,165–88.

Khan, Aubhik, Robert G. King, and Alexander L. Wolman. 2003. "Optimal Monetary Policy." *Review of Economic Studies* 70 (October): 825–60.

Krause, Michael U., and Thomas A. Lubik. 2007. "The (Ir)relevance of Real Wage Rigidity in the New Keynesian Model with Search Frictions." *Journal of Monetary Economics* 54 (April): 706–27.

Kryvtsov, Oleksiy, and Virgiliu Midrigan. 2009. "Inventories and Real Rigidities in New Keynesian Business Cycle Models." Manuscript.

Kydland, Finn E., and Edward C. Prescott. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica* 50 (November): 1,345–70.

Levin, Andrew T., and David Lopez-Salido. 2004. "Optimal Monetary Policy with Endogenous Capital Accumulation." Manuscript.

Lucas, Robert. 1987. *Models of Business Cycles*. Yrjö Johansson Lectures Series. London: Blackwell.

Lubik, Thomas A., and Wing Leong Teo. 2009. "Inventories, Inflation Dynamics and the New Keynesian Phillips Curve." Manuscript.

Maccini, Louis J., and Adrian Pagan. 2008. "Inventories, Fluctuations and Business Cycles." Manuscript.

Maccini, Louis J., Bartholomew Moore, and Huntley Schaller. 2004. "The Interest Rate, Learning, and Inventory Investment." *American Economic Review* 94 (December): 1,303–27.

Ramey, Valerie A., and Kenneth D. West. 1999. "Inventories." In *Handbook of Macroeconomics*, Volume 1, edited by John B. Taylor and Michael Woodford. pp. 863–923.

Schmitt-Grohé, Stephanie, and Martín Uribe. 2007. "Optimal, Simple and Implementable Monetary and Fiscal Rules." *Journal of Monetary Economics* 54: 1,702–25.

Wen, Yi. 2005. "Understanding the Inventory Cycle." *Journal of Monetary Economics* 52 (November): 1,533–55.

West, Kenneth D. 1986. "A Variance Bounds Test of the Linear Quadratic Inventory Model." *Journal of Political Economy* 94 (April): 374–401.

Economic Quarterly—Volume 95, Number 4—Fall 2009—Pages 383–418

Dynamic Provisioning: A Countercyclical Tool for Loan Loss Reserves

Eliana Balla and Andrew McKenna

The methodology to recognize loan losses set forth by the Financial Accounting Standards Board (FASB) and the International Accounting Standards Board (IASB) is referred to as the incurred loss model and defined as the identification of inherent losses in a loan or portfolio of loans. Inherent credit losses, under current accounting standards in countries following FASB and IASB, are event driven and should only be recognized upon an event's occurrence.[1] This has tended to mean that reserves for loan losses on a bank's balance sheet need to grow significantly during an economic downturn, a time associated with increased credit impairment and default events. Critics of the incurred loss model have pointed to it as one of the causes of the severity of strain many financial institutions experienced at the onset of the financial crisis of 2007–2009. As rapid provisioning to increase loan loss reserves made headlines, discussions of international regulatory banking reform included the method of dynamic provisioning as a potential alternative to the incurred loss approach (see, for example, Cohen 2009). Dynamic provisioning is a statistical method for loan loss provisioning that relies on historical data for various asset classes to determine the level of provisioning that should occur on a quarterly basis in addition to any provisions that are

The authors would like to thank Teresita Obermann, Jesús Saurina, and Tricia Squillante for their help in researching the details of the Spanish provisioning system, and Stacy Coleman, Borys Grochulski, Sabrina Pellerin, Mike Riddle, Diane Rose, David Schwartz, and John Walter for their helpful comments. We are also grateful to David Gearhart and Susan Maxey for their research assistance. Any errors are our own. The views expressed in this article are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. The authors can be reached at Eliana.Balla@rich.frb.org and Andrew.McKenna@rich.frb.org.

[1] FASB and IASB, March 2009 meeting. Information for Observers. Project: Loan Loss Provisioning.

event driven.[2] The primary goal of dynamic provisioning is the incremental building of reserves during good economic times to be used to absorb losses experienced during economic downturns.

We begin this paper with a discussion of the current approach to loan loss reserves (LLR) in the United States. We argue that, to a social planner who cares both about avoiding bank failures and the efficiency of bank lending, the current accounting and regulatory approach for LLR may be suboptimal on both fronts. First, by taking provisions after the economic downturn has set in, a bank faces higher insolvency risk. When a banker or regulator determines that a bank has inadequate LLR, the bank will have to build the reserves in an unfavorable economic environment. Also, inadequate reserves imply that regulatory capital ratios have been overstated, placing the bank at a higher risk for resolution by the Federal Deposit Insurance Corporation (FDIC). Second, as most banks tend to increase LLR during the economic downturn, the current approach may be procyclical; that is, it may amplify the business cycle. We aim to highlight some of the potential inefficiencies under the incurred loss approach by contrasting it to dynamic provisioning. Dynamic provisioning was instituted in Spain in 2000 in response to some of the same problems we highlight in the United States. We present a conceptual framework to compare loan loss provisioning under the incurred loss framework and dynamic provisioning. Then we simulate dynamic provisioning with U.S. data to present an empirical comparison. In the remainder of this section, we offer a brief summary of our main arguments and the conclusions from the simulation exercise.

In accounting terms, the LLR account, also known as the allowance for loan and lease losses (ALLL), is a contra-asset account used to reduce the value of total loans and leases on a bank's balance sheet by the amount of losses that bank managers anticipate in the most likely future state of the world.[3] LLR incorporate both statistical estimates and subjective assessments. Provisioning is the act of building the LLR account through a provision expense item on the income statement. While we present the intuition behind LLR in this section, the Appendix to the paper describes their important accounting features in basic terms.

Interest margin income from loans is a smooth flow whereas a loan default or impairment event causes a lumpy drop in the stock of bank assets. This introduces volatility to banks' balance sheets. By themselves, a large number

[2] Dynamic provisioning is also known as statistical provisioning and countercyclical provisioning.

[3] See, for example, Ahmed, Takeda, and Thomas (1999). See Benston and Wall (2005) for a treatment of fair value accounting as it pertains to loan losses. The key to Benston and Wall's arguments is that if loans could be reported reliably at fair value, where fair value is value in use, there would be no need for a loan loss provision or reserves. But a market for the full transfer of credit risk does not exist and loans cannot be reported reliably at fair value.

of loans may be insufficient to smooth these fluctuations out due to the correlation between the risks in the portfolio of bank loans. Some defaults are to be expected in a typical portfolio of bank loans. In order to avoid excess volatility of bank capital levels, banks can build a buffer stock of reserves against expected losses. Intuitively, LLR should serve to absorb expected loan losses while bank capital serves to absorb unexpected losses.[4] The key difference between a conventional economic definition of expected losses and incurred losses is that, unlike expected losses, incurred losses cannot incorporate information from expected future changes into economic variables that affect credit defaults. Incurred losses are entirely based on historical information.[5] If expected losses are greater than the loan loss reserve, regulatory capital ratios overstate the bank capital available to protect against insolvency risk.

To understand the argument that current loan loss accounting standards may have procyclical effects, we have to think about the LLR through the economic cycle. U.S. banking data show that LLR tend to be much lower during good economic times relative to bad economic times. An event-driven approach to LLR does not account for a booming economy resulting in banks relaxing their underwriting standards and taking greater risks.[6] Most bad loans will only reveal themselves in a recession. In that sense, the current approach may magnify the economic boom. By delaying provisioning for loan losses until the economic downturn has set in, the current approach may also magnify the bust. Reserves have to be built at a time when bank funds are otherwise strained, potentially furthering the credit crunch.[7] Therefore, even though banks should want to build "excess" LLR voluntarily during the boom years (it is efficient to do so from their perspective and it would have the benefit of offsetting cyclicality), the accounting guidelines pose a constraint.

---

[4] See Laeven and Majnoni (2003, Appendix A) for a detailed description of the conceptual relationship between LLR, provisions, capital, and earnings.

[5] Typically, expected loss is the mean of a loss distribution measured over a one-year horizon (expected loss is loss given default times the probability of default times the exposure at default); see Davis and Williams (2004). One way to separate the two concepts is by stating that no expected economic impacts are taken into account in LLR methodology. A bank manager cannot, for example, consider the increases in default risk due to future increases in unemployment.

[6] Independently of any LLR effects, stylized facts and a burgeoning literature suggest that bank lending behavior is procyclical. Many explanations have been presented. The classical principal-agent problem between shareholders and managers may lead to procyclical banking if managers' objectives are related to credit growth. Two of the more recent theories are "herd behavior" and "institutional memory hypothesis." Rajan (1994) suggests that credit mistakes are judged more leniently if they are common to the whole industry (herd behavior). Berger and Udell (2003) suggest that, as the time between the current period and the last crisis increases, experienced loan officers retire or genuinely forget about the lending errors of the last crisis and become more likely to make "bad" loans (the institutional memory hypothesis). Our argument here is that LLR effects may add to this otherwise present procyclicality of bank lending.

[7] See Hancock and Wilcox (1998) and the sources cited therein for a discussion of the literature on the credit crunch. Also see Eisenbeis (1998) for a critique of Hancock and Wilcox (1998).

"Excess" reserves are associated with managing earnings, which is viewed as undesirable by the accounting profession. Wall and Koch (2000) offer a review of the theoretical and empirical evidence on earnings management via loan loss accounting. The evidence they summarize suggests that banks both have an incentive to and, in general, are using loan loss accounting to manage reported earnings. From the perspective of the accounting profession, using LLR to manage reported earnings is in conflict with the goals of transparency of a bank's balance sheet as of the date of the financial statement. We take it as a given that the goals and concerns of the accounting standard setters are valid. We simply highlight the resulting tradeoffs.

We illustrate the tradeoffs by pointing to an alternative system of reserving for loan losses—dynamic provisioning. Dynamic provisioning is, at its core, a deliberate method to build LLR in good economic times to absorb loan losses during an economic downturn, without putting undue pressure on earnings and capital. Spanish regulators instituted dynamic provisioning in 2000 explicitly to combat their banking system's procyclicality.[8] In maintaining a focus on the use of historical data in its approach to loan loss provisioning, the Bank of Spain (the regulator of Spanish banks) has been able to adopt dynamic provisioning in compliance with IASB standards. We describe the Spanish method in some detail and present data on Spanish reserves (relative to contemporaneous credit quality) against the United States and other Western European economies in 2006, before the beginning of the crisis. According to these data, the Spanish policy was effective in building relatively higher reserves and thus worthy of further study.

We compare the incurred loss and dynamic provisioning approaches. Through a basic example we illustrate that the key difference is not the level of provisioning but the timing of the provisioning. By taking provisions early when economic conditions are good, banks will avoid using capital in an economic downturn when it is more expensive, thereby reducing the probability of failure from capital deficiencies. Moreover, a goal of dynamic provisioning is to ensure that the balance sheet accurately reflects the true value of assets to banks. If income is not reduced to provision for assets that are not collectable, then managers may be pressured to provide greater dividends to investors based on the income that is reported in the period.

As a next step in our analysis, we conduct an empirical simulation to illustrate that a dynamic provisioning framework (akin to the one implemented in Spain) could have allowed for a build-up of reserves during the boom years in the United States. The results demonstrate that the alternate framework would have smoothed bank income through the cycle and provisioning levels

[8] In the current provisioning system, outside of Spain, loan loss provisions are generally countercyclical but their effect is thought to be procyclical. We refer to "procyclicality" as the amplification of otherwise normal business fluctuations.

would have been significantly lower during the financial crisis of 2007–2009. Note that, in contrast to accountants, bank regulators would not take issue with LLR resulting in income smoothing because the regulators' primary concern is the adequacy of the reserves to sustain loan losses.

The remainder of the article proceeds as follows. Section 1 describes current rules for LLR in the United States, as well as the issues confronting the current system, particularly as identified during the financial crisis of 2007–2009. Section 2 provides a conceptual framework for comparing the incurred loss and the dynamic provisioning approaches to LLR. Section 3 describes the approach as implemented by the Bank of Spain. Section 4 builds a simulation of dynamic provisioning with historical U.S. data. Section 5 concludes.

1. THE CURRENT ACCOUNTING AND REGULATORY FRAMEWORK FOR LOAN LOSS PROVISIONING IN THE UNITED STATES

Bank regulators view adequate LLR as a "safety and soundness" issue because a deficit in LLR implies that a bank's capital ratios overstate its ability to absorb unexpected losses. As a result of their important relationship to bank capital and financial reporting transparency, rules governing LLR have been revisited many times by bank regulators and accounting standard setters. Two crucial revision points relate to the new regulatory capital rules in the Basel Capital Accord (signed in 1988) as enacted by the Federal Deposit Insurance Corporation Improvement Act of 1991 (FDICIA)[9] and the landmark case of the SunTrust Bank earnings restatement that occured in 1998.[10] Changes in capital rules may have reduced bank manager incentives to keep large reserve buffers, while the implementation of accounting rules following the SunTrust case may have made it more difficult to justify building a reserve buffer during good economic times. This section documents current rules that govern LLR, the U.S. data from the last three cycles, and the importance of LLR both for bank solvency and the procyclicality of bank lending.

Incurred Loss Accounting

Provisioning for loan losses in the United States is accounted for under Financial Accounting Standard (FAS) Statement 5, Accounting for Contingencies, and FAS 114, Accounting by Creditors for Impairment of a Loan—an amendment of FAS Statements 5 and 15. Impaired loans evaluated under FAS 114,

[9] See Walter (1991) for extensive coverage of LLR leading to the 1991 changes. Ahmed, Taekda, and Thomas (1999) study how FDICIA (1991) changes affected the relationship between loan loss provisioning, capital, and earnings.

[10] See Wall and Koch (2000) for an extensive summary of the theoretical and empirical evidence on bank loan loss accounting and LLR philosophies.

which provides guidance on estimating losses on loans individually evaluated, must be valued based on the present value of cash flows discounted at the loan's effective interest rate, the loan's observable market price, or the fair value of the loan's collateral if it is collateral-dependent.[11] Loans individually evaluated under FAS 114 that are not found to be impaired are transferred to homogenous groups of loans that share common risk characteristics, which are evaluated under FAS 5. FAS 5 provides for accrual of losses by a charge to the income statement based on estimated losses if two conditions are met: (1) information available prior to the issuance of the financial statements indicates that it is probable that an asset has been impaired or a liability has been incurred at the date of the financial statement, and (2) the amount of the loss can be reasonably estimated.[12]

Both FAS 114 and 5 allow banks to include environmental or qualitative factors in consideration of loan impairment analysis. Examples of these factors include, but are not limited to, underwriting standards, credit concentration, staff experience, local and national economic trends, and business conditions. In addition, FAS 5 allows for the use of loss history in impairment analysis.[13] These elements provide bankers with flexibility in determining the level of provisions taken against incurred losses when they are well substantiated by relevant data and/or documentation required by supervisors and accountants. This paper includes illustrative examples to support our explanation of the technical aspects of accounting for loan losses. For consistency, when we refer to identified loan losses, we mean the accounting conditions for taking a provision were met. In practice, banks identify losses by categorizing loans based on their payment status (i.e., current, 30 days past due, 60 days past due, etc.) and the severity of delinquency (which can vary by asset class) and assess whether a provision should be taken on loans they expect to experience a loss, if the loss is probable and estimable.[14]

The U. S. Data

The adequacy of LLR to cover loan losses is generally measured against the level of non-performing loans (the ratio of the two is known as the coverage ratio), meaning loans that are seriously delinquent by being 90 or more days past due or in non-accrual status. Figure 1 shows LLR and non-performing

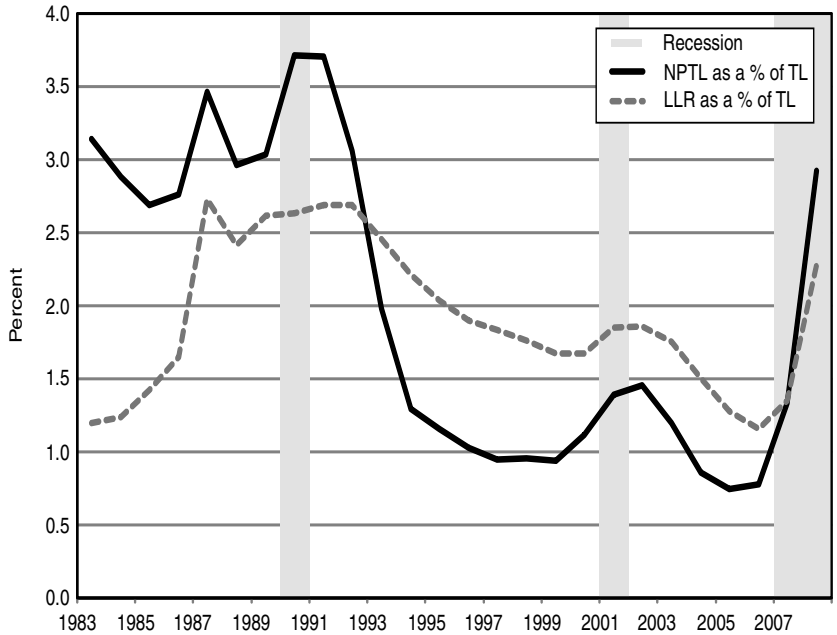
[11] Financial Accounting Standards Board, Summary of Statement No. 114: Accounting by Creditors for Impairment of a Loan—An amendment of FASB Statements No. 5 and 15 (Issued 5/93).

[12] Financial Accounting Standards Board, Summary of Statement No. 5: Accounting for Contingencies (Issued 3/75).

[13] SR 06-17: Interagency Policy Statement on the ALLL, December 13, 2006. SR 01-17: Policy Statement on ALLL Methodologies and Documentation for Banks and Savings Institutions, July 2, 2001. SR 99-22: Joint Interagency Letter on the Loan Loss Allowance, July 26, 1999 IPS.

[14] See Walter (1991) for an explanation of how banks identify and categorize defaults.

Figure 1 Loan Loss Reserves Versus Non-Performing Loans Ratios: U. S. Bank Aggregates in Levels, 1983–2008

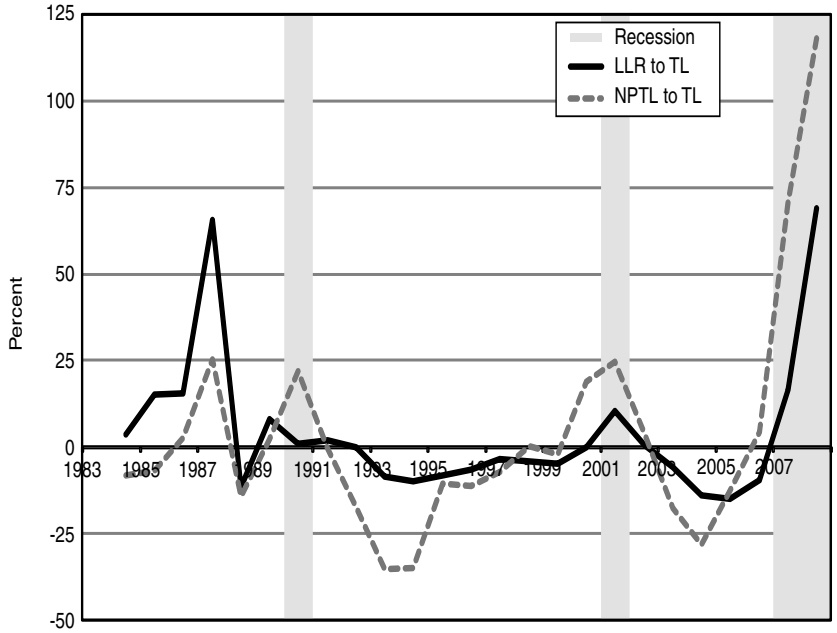


Notes: LLR = loan loss reserves; NPTL = non-performing total loans; TL = total loans.
Source: Call Reports.

loans, both scaled by total loans, between 1983 and 2008. The data come from the Commercial Bank Consolidated Reports of Condition and Income Reports (Call Reports) and they are constructed by combining all U.S. banks into an "aggregate" balance sheet. We show the aggregate level of reserves in the U.S. banking system at any point in time, the aggregate level of non-performing loans, and so on.

Figure 1 depicts the cyclicality of LLR. The first cycle is different from the subsequent two. Reserves are lower than non-performing loans during the banking crisis of the late 1980s and early 1990s. Because of the major regulatory changes that took full effect in 1992, we look more closely at the last two cycles. At the height of the boom, in 2005, we saw some of the lowest reserves relative to total loans on record. Note that this is not surprising given that non-performing loans and reserves move together. At any point between 1992 and 2006, there were more reserves in the system than there were non-performing loans. In 2005, banks had historically high coverage ratios, not

Figure 2 Loan Loss Reserves Versus Non-Performing Loans Ratios: U. S. Bank Aggregates Year Over Year Percentage Change, 1983–2008



Notes: LLR = loan loss reserves; NPTL = non-performing total loans; TL = total loans.
Source: Call Reports.

because reserves were high (indeed they were at a historical low), but because non-performing loans were so low. The period 2006–2008 demonstrated how reserves needed to be built in an economic downturn to keep pace with the rapid deterioration in credit quality. Like credit problems, reserves grow just ahead of a recession and continue to grow after the recession has ended. In the first quarter of 2006, overall U.S. reserves were 1.1 percent of total loans. In the first quarter of 2009, they were built up to 2.7 percent of total loans.

Trends in reserve adjustments against changes in non-performing loans, shown in Figure 2, illustrate that modifications to LLR tend to lag credit problems, and reserves increase more slowly than non-performing loans during economic busts and fall more slowly in booms. Both Figures 1 and 2 indicate that some build-up of reserves relative to non-performing loans existed in the

U.S. banking system but the cushion shrank in the 2000s boom relative to the 1990s boom.[15]

Loan Loss Provisioning and Bank Solvency

The Basel Accord set current rules for LLR that prescribe the use of impairment and estimated loss methodology.[16] FDICIA enacted these changes into law. LLR were no longer counted as a component of Tier 1 capital but were counted toward Tier 2 capital, up to 1.25 percent of the bank's risk-weighted assets. Laeven and Majnoni (2003) have argued that "... from the perspective of compliance with regulatory capital requirements, it became much more effective for U.S. banks to allocate income to retained earnings (entirely included in Tier 1 capital) than to loan loss reserves (only partially included in Tier 2 capital)" (Laeven and Majnoni 2003, 194).

In the new regulatory regime of Basel I, banking regulators remained concerned with the roles that the loan losses and banks' reserve for losses play in insolvency risk. Comptroller of the Currency John Dugan, the regulator of U.S. national banks, has stated that "... banking supervisors love the loan loss reserve. When used as intended, it allows banks to recognize an estimated loss on a loan or portfolio of loans when the loss becomes likely, well before the amount of the loss can be determined with precision and is actually charged off. That means banks can be realistic about recognizing and dealing with credit problems early, when times are good, by building up a large 'war chest' of loan loss reserves. Later, when the loan losses crystallize, the fortified reserve can absorb the losses without impairing capital, keeping the bank safe, sound, and able to continue extending credit" (Dugan 2009). But accounting guidelines, as enforced in the late 1990s and 2000s, may have limited the ability of LLR to function in the way summarized by Comptroller Dugan.

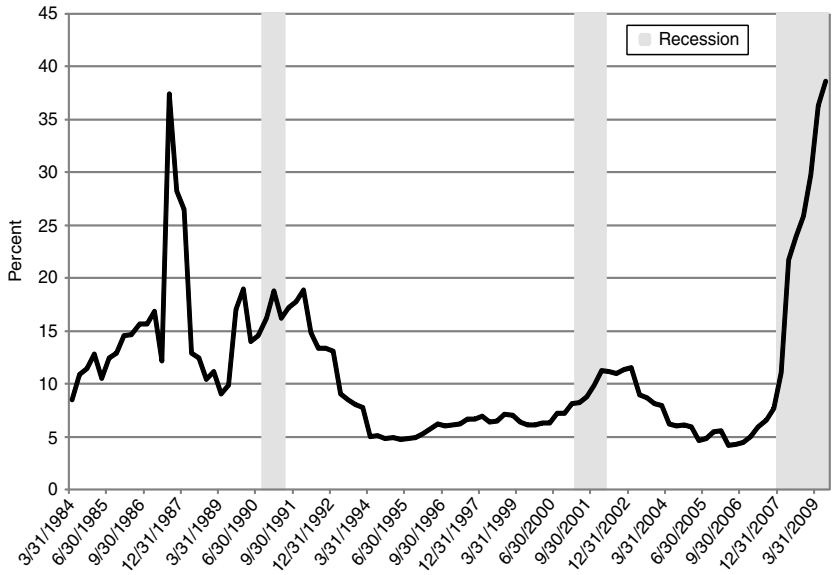
In the mid-1990s, the Securities and Exchange Commission (SEC) was increasingly concerned that U.S. banks may be overstating their LLR, potentially using this account to manage reported earnings. In 1998, following an SEC inquiry, SunTrust Bank agreed to restate prior years' financial statements, reducing its provisions in each of the years 1994–1996 and resulting in a cumulative reduction of $100 million to its LLR (see Wall and Koch 2000). Analysts of the U.S. banking industry viewed the SunTrust restatement as a permanent strengthening of the existing accounting constraint on a bank's LLR policy.

---

[15] Note that non-performing loans are only shown as a limited approximation of incurred losses. There is no regulatory guidance that advocates a 100 percent reserve coverage for non-performing loans. Nonetheless, it is a helpful standard simplification to present the data in this way.

[16] Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework—Comprehensive Version. http://www.bis.org/publ/bcbs128.htm.

Figure 3 Loan Loss Provisions as a Percentage of Net Operating Revenue: U. S. Bank Aggregates, 1984–2009



Source: Call Reports.

Bankers desire flexibility in recognition of the subjective aspects in determining appropriate reserves. Bank regulators desire flexibility in recognition of the importance of LLR for bank safety and soundness. Accounting standard setters stress the need for transparency and comparability across banks' financial statements. We take it as a given that the goals of the accounting standard setters and their concern over earnings management are valid. We simply highlight the resulting tradeoffs.

Figure 3 illustrates the importance of provision expense relative to bank income. First, the size of provisions relative to earnings helps us understand their importance to bank managers, accountants, and bank regulators. Second, the period 2007–2009 illustrates nicely the inverse relationship between earnings and provisions in a recession. Banks had to sharply increase provisions in recognition of pending losses, which for many banks more than offset earnings and reduced capital.

Loan Loss Provisioning and Procyclicality

By entering the current economic downturn with low LLR, the banking sector may have unintentionally exacerbated the cycle. In a speech in March 2009, Ben Bernanke, the Chairman of the Board of Governors of the Federal Reserve stated that there is "considerable uncertainty regarding the appropriate levels of loan loss reserves over the cycle. As a result, further review of accounting standards governing. . . loan loss provisioning would be useful, and might result in modifications to the accounting rules that reduce their procyclical effects without compromising the goals of disclosure and transparency" (Bernanke 2009).

The cyclicality of loan loss provisioning is well documented with cross-country data. During periods of economic expansion, provisions fall (as a percentage of loans) and, conversely, they rise during downturns. Figure 1 illustrated the cyclicality of loan loss provisioning with U.S. data. As with bank regulatory capital, the concern is that with an approach in which banks have to rapidly raise reserves during bad times, the bad times could get prolonged.[17] Laeven and Majnoni (2003) and Bouvatier and Lepetit (2008) document the procyclicality of loan loss provisions with cross-country data. Banks delay provisioning for bad loans until economic downturns have already begun, amplifying the impact of the economic cycle on banks' income and capital.

Section 1 documented the current framework around LLR in the United States, the U.S. data from the last three cycles, and the importance of LLR both for bank solvency and the procyclicality of bank lending. In response to the recent experience where many banks had to increase their LLR abruptly and drastically, various U.S. and international regulators have expressed the desire to revisit LLR policies. The Financial Stability Forum's Working Group on Provisioning (2009) has recommended that accounting standard setters give due consideration to alternative approaches to recognizing and measuring loan losses. One approach that has garnered attention is dynamic provisioning.

2. INCURRED LOSS ACCOUNTING VERSUS DYNAMIC PROVISIONING: A CONCEPTUAL FRAMEWORK

In this section, we discuss and compare dynamic provisioning with the incurred loss methodology using two simplified examples that will set the stage for a more complicated simulation completed in a subsequent section. And, while

[17] For simplicity, we are not addressing in this article all the links between LLR and regulatory capital, nor the similarities between the cyclical effects of LLR and regulatory bank capital. On the latter, we refer the reader to a large literature ranging from Bernanke and Lown (1991) to Peek and Rosengren (1995) to Pennacchi (2005), who use U.S. data to analyze the effects of capital requirements on banks and the economy.

we will review at a high level the technical nuances of accounting, a more detailed discussion of accounting basics is included in the Appendix.

The IASB and the British Bankers Association (BBA) provide more flexibility than the FASB for firms to include forward-looking elements based on expectations of total losses over the life of loans, rather than losses already realized, but are still similar to the U.S. standards. The BBA Statement of Recommended Practice on advances indicates specific provisions should be made to cover the difference between the carrying value and the ultimate realizable value, made when a firm determines that information suggests impairment. General provisions[18] take into account past experience and assumptions about economic conditions, but are generally small because the Basel Accord of 1988 limits general provisions to 1.25 percent of risk-weighted assets and such provisions are not tax deductible (Mann and Michael 2002).

The IASB, in standard 39, requires assessment at each balance sheet date to determine whether there is objective evidence that an asset or group of assets is impaired—the difference between the asset's carrying amount and the present value of estimated cash flows discounted at the original interest rate—based on criteria such as the financial condition of the issuer, breach of contract, probability of bankruptcy, and historical pattern of collections of accounts receivable.[19]

The incurred loss model for accounting for loan losses is divorced from prudential goals of maintaining the safety and soundness of financial institutions in that the FASB and IASB frameworks do not aim to influence the decisions investors make toward a specific objective, such as financial stability.[20] Instead, the goal of financial reporting is to provide financial statement users with the most accurate information about identified losses in the loans and leases portfolio. The dynamic provisioning model, conversely, has as its primary objective the enhancement of the safety and soundness of banks. Its fundamental premise is that when loans are made the probability that default will occur on a loan is greater than zero. In accordance with this philosophy, the dynamic provisioning approach provides a model-based mechanism for a bank to build a stock of provisions in good times so it will not face insolvency due to charge-offs and provisions in bad times. In a simple exercise, presented in Tables 1 and 2, we demonstrate how the application of the incurred loss model would differ from that of dynamic provisioning.[21]

We will make a number of simplifications to help facilitate an understanding of this topic. These assumptions, if not applied, wouldn't substantively

[18] The Bank of Spain refers to statistical provisions as general provisions. See Saurina (2009a).

[19] Deloitte: Summaries of International Financial Reporting Standards. http://www.iasplus.com/standard/ias39.htm.

[20] The Financial Crisis Advisory Group. Public Advisory Meeting Agenda, February 13, 2009.

[21] The illustration builds on an example provided by Mann and Michael (2002).

Table 1 Incurred Loss Model

Assets	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**
Total Loans and Leases	1,000	1,000	1,000	1,000	1,000
Expected Losses	40	40	70	100	100
Stock of Specific Provisions	0	0	30	100	100
Yearly Charge-offs	0	0	5	60	35
Total Charge-offs	0	0	5	65	100
Total Stock of Provisions Net of Stock of Charge-offs	0	0	25	35	0
Loans Net of Charge-offs	1,000	1,000	995	935	900
Income Statement					
Profit before Provision	30	30	30	30	30
Specific Provision	0	0	30	70	0
Profit after Provision	30	30	0	−40	30
Shareholders' Equity					
Shareholders Equity, Beginning	44	50	56	56	16
Other Expenses and Dividends	24	24	0	0	24
Retained Earnings	6	6	0	−40	6
Shareholders' Equity, End	50	56	56	16	22
Equity to Assets	5.00%	5.60%	5.63%	1.71%	2.44%
Solvency: Equity to Assets $\geq$ 2%	Solvent	Solvent	Solvent	Insolvent	Solvent

Notes: Expected losses represent bank managers' expectations for losses over the five-year period in the listed year. Specific provisions are those identified as probable and estimable. Yearly charge-offs are those taken in the listed year, while the total is the sum of charge-offs over the five-year period. Equity to assets is shareholders' equity, end divided by loans net of charge-offs.

change the results of our example, but instead, on average, only reinforce our conclusions. First, the example bank in question will make all its loans in Year 1 and will not grow its portfolio. This prevents us from having to account for changes in bank managers' preference for varying risks of loans over time as the economic environment changes. Second, we assume that, over the five-year cycle, the bank's pre-provision profits don't vary. If we introduced profit declines as conditions worsened, which would likely occur since the bank's net interest margin (its only source of income in the example) would decline due to charge-offs, the result would be reduced pre-provision profits and the reduction of profit due to provisions would be greater. Third, we assume we know the ex post level of risk in the portfolio of loans. In our example, losses at the end of the five-year cycle will equal 10 percent of the loan portfolio. For illustrative purposes we make the assumption that bank managers believe, based on the bank's historical loss experience, total ex post losses will equal 4 percent in Year 1 when loans are issued. As a loan approaches maturity in Year 5, managers can, with greater certainty, estimate the true extent of ex

Table 2 Dynamic Provisioning Model

Assets	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**
Total Loans and Leases	1,000	1,000	1,000	1,000	1,000
Expected Losses	40	40	70	100	100
Stock of Specific Provisions	0	0	30	100	100
Stock of Statistical Provisions	20	40	40	0	0
Total Stock of Provisions	20	40	70	100	100
Yearly Charge-offs	0	0	5	60	35
Total Charge-offs	0	0	5	65	100
Total Stock of Provisions Net of Stock of Charge-offs	0	0	65	35	0
Loans Net of Charge-offs	1,000	1,000	995	935	900
Income Statement					
Profit before Provision	30	30	30	30	30
Provisions	20	20	30	30	0
Profit after Provision	10	10	0	0	30
Shareholders' Equity					
Shareholders' Equity, Beginning	44	46	48	48	48
Other Expenses and Dividends	8	8	0	0	24
Retained Earnings	2	2	0	0	6
Shareholders' Equity, End	46	48	48	48	54
Equity to Assets	4.60%	4.80%	4.82%	5.13%	6.00%
Solvency: Equity to Assets $\geq$ 2%	Solvent	Solvent	Solvent	Solvent	Solvent

Notes: Expected losses represent bank managers' expectations for losses over the five-year period in the listed year. Specific provisions are those identified as probable and estimable. Statistical provisions are those taken under the dynamic provisioning model based on the historical data used to estimate the annual statistical provision. Yearly charge-offs are those taken in the listed year, while the total is the sum of charge-offs over the five-year period. Equity to assets is shareholders' equity, end divided by loans net of charge-offs.

post losses. To show how provisioning levels are determined over time, we allowed managers' expectations about expected losses to change in each year to justify provisions each period (the expected losses entry in Tables 1 and 2). Lastly, since the primary intent of dynamic provisioning is to better prepare financial institutions to absorb loan losses, we make assumptions about shareholders' equity. First, we assume shareholders' equity is all tangible equity and total loans and leases equal total assets. Accordingly, shareholders' equity is determined by adding retained earnings, after dividends and expenses, to shareholders' equity. To determine yearly retained earnings we assume a constant ratio of dividend and other expenses equal to 80 percent of after-provision profit. Under current prompt corrective action (PCA) standards used by banking regulators, the Tier I capital ratio, total capital ratio, and leverage ratio are used to determine when banks need to be resolved by the FDIC. In this

Table 3 Capital Guidelines

	Leverage Ratio
Well Capitalized	5 percent
Adequately Capitalized	4 percent
Undercapitalized	< 4 percent
Significantly Undercapitalized	< 3 percent
Critically Undercapitalized	< 2 percent

example, since shareholders' equity most closely resembles tangible equity, we will focus on the leverage ratio, which in the example is equal to equity divided by assets, as the primary measure of solvency. In Year 1 we assume the bank starts with $44 shareholders' equity and $1,000 total assets, leaving it adequately capitalized at 4.4 percent. The PCA guidelines listed in Table 3 apply to the leverage ratio.

We begin our discussion with the incurred loss model. As previously mentioned, bank managers expect that, given the average risk and performance of loans and leases issued, the banks' total losses will amount to 4 percent of the face value of loans at the end of Year 5; this is reflected in Year 1, the first column in Tables 1 and 2, listed in the row entitled Expected Losses. However, since data do not exist that support identification of losses—evidence suggesting the 4 percent of losses is probable and estimable—bank managers cannot take provisions until such data exist. In Year 3, the first losses of $30 are identified in the portfolio and managers' expectation about total losses increases to 7 percent (or $70) based on new data. However, since only $30 in losses have been identified, only a provision of that amount can be taken. Additionally, since the $30 pre-provision profit was wiped out by the provision, the bank's managers can't modify other expenses and dividends to increase capital to prepare for the increase in losses. In Year 4, charge-offs increase to $60 for the year and bank managers believe losses will amount to $100, or 10 percent of loans and leases. The bank's profit of $30 is eliminated by a provision of $70 required by accountants, and a net loss of $40 (shown as negative retained earnings in Table 1) requires the bank to reduce its capital by that amount. When a net loss occurs, banks must use capital to equate assets with liabilities and shareholders' equity on the balance sheet. The reduction in capital leaves the bank with only $16, resulting in a leverage ratio of 1.71 percent. By PCA standards, this leaves the bank critically undercapitalized and it will be resolved by banking regulators.

Table 2 presents the example bank under dynamic provisioning. Under the dynamic provisioning model, bank managers would take a different approach to provision for loan losses focusing on incrementally building up a fund (referred to as the statistical fund) to protect the bank against losses expected but not yet identified in the loans and leases portfolio. The model driving

the building process for the statistical fund would rely on historical default data for the types of loans and leases issued by the bank rather than models estimating expected losses; this is an important distinction between the dynamic provisioning approach and a strict expected loss approach. In Year 1, when loans and leases of $1,000 were made, bank managers expected a total of 4 percent of the loans and leases portfolio to default based on its historical data, as reflected in the stock of statistical provisions built. Accordingly, the managers establish a plan to build a fund for the bank to absorb those losses over a two-year period. In Years 1 and 2 the bank provisions $20 to reach the $40 fund desired. This has the effect of reducing profits in those years to $10 instead of $30, which in turn reduces the amount of other expenses and dividends the bank's managers have and the amount of retained earnings available to build capital. However, since the fund is capital set aside to absorb losses the bank is anticipating based on historical data, the Year 2 capital ratio including the fund is 8.8 percent, compared to 5.6 percent under the incurred loss model in the same year. In Year 3, charge-offs increase to 0.5 percent and the bank's managers modify their expectation of total losses expected on the loans and leases portfolio to 7.0 percent.

As with the incurred loss model, the $30 identified in specific provisions would also be taken under the dynamic provisioning model. Bank managers, in segmenting specific from statistical provisions, are not only preparing for losses they expect to incur at some point over the life of the loan, but they are also signaling to users of financial statements the difference between the expectation for losses as suggested by the historical data on defaults for the assets held and those actually identified in the portfolio as probable and estimable. The combined fund of $40 and the specific provision of $30 allow the bank to have total provisions equal to $70, preparing the bank to fully absorb the expected losses. And while this reduces profit to $0, the bank's solvency is well protected. In Year 4, charge-offs increase sharply to $60 and the bank's managers expect losses to increase to $100. Bank managers shift the $40 statistical fund to specific provisions and add $30. After annual charge-offs of $60, the bank has $35 remaining at the end of Year 4 in its stock of provisions to absorb the remaining $35 in identified losses. Again, the bank's provisioning reduces its profit to $0, but in a difficult economic environment when bank losses are sharply increasing, dividends are being halted throughout the industry and the bank's managers are actively attempting to reduce expenses, the primary concern is solvency. The bank under the incurred loss model becomes insolvent in Year 4, while the bank under the dynamic provisioning model is actually able to increase its leverage ratio, although due to a reduction in balance sheet assets from charge-offs. In Year 5, under the assumption of no recoveries, the remaining charge-offs of $35 occur and the stock of provisions is depleted. The remaining loans and leases are paid down and the bank's leverage ratio increases again.

This basic example illustrates one of the primary points of importance for dynamic provisioning: the key difference is not the level of provisioning but the timing of the provisioning. By taking provisions early when economic conditions are good, banks may be able to avoid using their capital in an economic downturn when it is more expensive, thereby reducing the probability of failure from capital deficiencies. Moreover, an objective of dynamic provisioning is to ensure that the balance sheet accurately reflects the true value of assets to banks. If income is not reduced to provision for assets that are not collectable, then managers may be pressured to provide greater dividends to investors based on the income that is reported in the period.

3. DYNAMIC PROVISIONING AND THE SPANISH EXPERIENCE

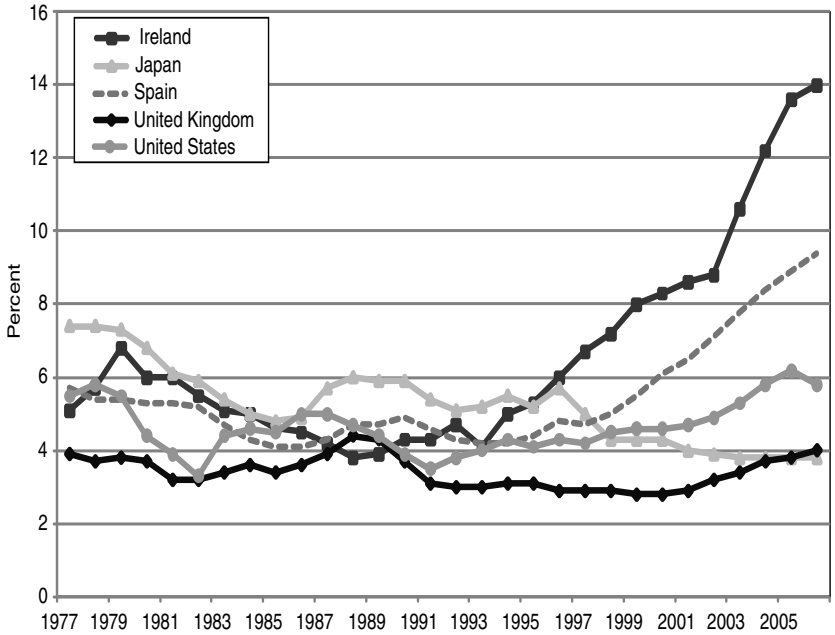
In the wake of the financial crisis of 2007–2009, as various banking policymakers revisit loan loss provisioning rules, there have been calls to study the Spanish experience. The Spanish provisioning system has been credited by most market observers as positioning its banks to avoid the strain that their international peers experienced in 2007. This section reviews the historical developments that led to Spain's adoption of dynamic provisioning. Subsequently, we provide methodological details of the original dynamic provisioning approach implemented in 2000 and the modifications made in 2004. Lastly, we briefly discuss the position of Spanish banks at the end of 2006, just prior to the onset of the financial crisis, which will allow readers to gauge the efficacy of the policy; 2006 is the year of emphasis for our analysis because the primary intent of dynamic provisioning is to prepare banks to absorb losses, which, if effective, should have been accomplished by year's end in 2006.

Why is Spain a Relevant Example?

Spain makes for an interesting case in reference to an international financial crisis precipitated by a widespread housing boom.[22] Spain's housing sector boom greatly outpaced the United States', Japan's, and the United Kingdom's, as seen in Figure 4. Spanish loan loss provisions historically demonstrated high procyclicality with the business cycle. From 1991–1999, Spain's correlation between loan loss provisioning levels and GDP was −0.97, the highest in the Organisation for Economic Co-operation and Development (OECD). Similarly, in 1999, Spain had the lowest level of loan loss provisions (to total loans) of any OECD country (Saurina 2009a). In response to pronounced procyclicality, the Bank of Spain in 2000 implemented a countercyclical

[22] For a discussion of the housing boom in Spain, see Garcia-Herrero and Fernández de Lis (2008).

Figure 4 Gross Fixed Capital Formation in Housing as a Percentage of GDP, 1977–2006



Source: OECD data.

method of loan loss provisioning that allowed banks to build a reserve in good times to cover losses in bad times.[23, 24]

The Original Model

In 1989, the Bank of Spain was authorized to establish the accounting practices of the banks it supervises, helping to resolve the conflict between the objectives of accountants and regulators over issues such as loan loss provisioning. The Bank of Spain has historically viewed loan loss provisioning as a necessary policy to accomplish its goals of prudential supervision. In a

[23] Subsequent to the Asian Financial Crisis, many emerging Asian economies instituted loan loss provisioning policies (some used discretionary measures) that increased provisioning in good times, leaving the banking system well prepared to absorb losses associated with an economic downturn. For a detailed description of these measures, see Angklomkliew, George, and Packer (2009).

[24] For a brief summary of the Spanish method, see Saurina (2009b).

Table 4 Risk Categories Under Standard Approach to Statistical Provisioning: The 2000 Methodology

Category	Description
Without Risk (0.0%)	Risks involving the public sector
Low Risk (0.1%)	Mortgages with outstanding risk below 80 percent of the property value, as well as risks with firms whose long-term debts are rated at least "A"
Medium-Low Risk (0.4%)	Financial leases and other collateralized risks (different from the former in point 2)
Medium Risk (0.6%)	Risks not mentioned in other points
Medium-High Risk (1.0%)	Personal credits to financial purchases of durable consumer goods
High Risk (1.5%)	Credit card balances, current account overdrafts, and credit account excesses

regulation adopted in 2000 upon a foundation set out in Circular 4/1991, the Bank of Spain instituted dynamic provisioning (Poveta 2000). The Bank of Spain places significant emphasis on the growth years in a credit cycle and, given the enhanced procyclicality recognized in Spain (see Fernández de Lis, Pagés, and Saurina 2000), the goal of statistical provisioning is to compensate for the underpricing of risk that takes place during those years. Statistical provisioning intends to anticipate the next economic cycle, although it is not meant to be a pure expected loss model because it is backward-looking (Fernández de Lis, Pagés, and Saurina 2000). The statistical fund, built by quarterly provisions recognized on the income statement, was meant to complement the insolvency fund, but, instead of covering incurred losses, was built from estimates of latent losses on homogenous asset groups.

The regulation established two methods for computing the quarterly statistical provision. Banks can create their own internal model using their own loss experience, provided that the data used spans at least one economic cycle and is verified by the bank supervisor. Conversely, banks can use the standard approach outlined by the Bank of Spain, based on a parameter measuring the risk of institutions' portfolio of loans and leases. For this analysis, our focus will be on the standard approach for two reasons. First, the majority of institutions in Spain use the standard method and, second, the internal models are not available publicly and therefore cannot be the subject of analysis.

The standard approach was developed on the assumption that asset risk is homogenous. The Bank of Spain, in adopting the statistical provision in July 2000, created six risk categories of coefficients by assets for banks to use to take a quarterly provision. The coefficients, shown in Table 4, are based on historical data of the average net specific provision over the period

1986–1998, meant to reflect one economic cycle in Spain.[25] Similar to loan loss provision methods under International Financial Reporting Standards (IFRS) and generally accepted accounting principles (GAAP), the statistical provision is recognized on the income statement and thus has the effect of reducing profits when the difference between the statistical and specific provision is positive. When the statistical provision is negative it reduces the statistical fund, which cannot be negative, and increases profits. The statistical provision is not a tax-deductible expense. To limit the size of the statistical fund, which is a function of the type of assets a bank holds and the duration of economic growth, the fund was capped at 300 percent of the coefficient times the exposure.

The Current Model

In 2004, the Bank of Spain made several modifications to its approach to the statistical provision to conform to the IFRS guidance adopted by the European Union, becoming compulsory for Spanish banks in 2005.[26] The following equation sets out the new approach:

$$\text{General provision}_t = \sum_{i=1}^{6} \alpha_i \Delta C_{it} + \sum_{i=1}^{6} \left(\beta_i - \frac{\text{Specific provision}_{it}}{C_{it}} \right) C_{it}.$$

The new model retains a general risk parameter, α, meant to capture the different risks across homogenous categories of assets. For each asset class, α is the average estimate of credit impairment in a cyclically neutral year; and, although it is meant to anticipate the next cycle, it is not intended to predict it. Two components of the α parameter—that it is backward-looking and unbiased (cyclically neutral)—are crucial in differentiating the policy from an expected loss approach, which would attempt to gauge the characteristics of the next economic cycle through forecasting methods or incorporating expectations about future economic performance. Rather, the use of cyclically unbiased historical data allows provisions to be taken based on the past experience of assets. This feature removes the potential for conflict between banking regulators and banks because there is no discretion over the expectations of

[25] Fernández de Lis, Pagés, and Saurina (2000) indicate that the data incorporate banks' credit risk measurement and management improvements without specifically providing details.

[26] The Spanish experience could be considered unique in that the central bank is also the accounting standard setter for banks. Safety and soundness prudential decisions are not made in an accounting vacuum as illustrated by the need for changes to the Spanish model in 2004. The ongoing debate in many countries sets those who argue that a countercyclical reserve account could be made transparent in financial reporting (and sophisticated investors would differentiate the statistical build-up) against the accounting standard setters' reluctance to turn bank financial reports into prudential regulatory reports. Any LLR reform would involve compromise between bank regulators and accounting standard setters.

Table 5 Risk Categories Under Standard Approach to Statistical Provisioning: The 2004 Methodology

Category	Description
Negligible Risk ($\alpha = 0\%$, $\beta = 0\%$)	Cash and public sector exposures (both loans and securities)
Low Risk ($\alpha = 0.6\%$, $\beta = 0.11\%$)	Mortgages with a loan-to-value ratio below 80 percent and exposure to corporations with a rating of "A" or higher
Medium-Low Risk ($\alpha = 1.5\%$, $\beta = 0.44\%$)	Mortgages with a loan-to-value ratio above 80 percent and other collateralized loans not previously mentioned
Medium Risk ($\alpha = 1.8\%$, $\beta = 0.65\%$)	Other loans, including corporate exposures that are non-rated or have a rating below "A" and exposures to small- and medium-size firms
Medium-High Risk ($\alpha = 2.0\%$, $\beta = 1.1\%$)	Consumer durables financing
High Risk ($\alpha = 2.5\%$, $\beta = 1.64\%$)	Credit card exposures and overdrafts

economic performance that may lead to over- or underprovisioning. It is also beneficial for investors because it provides transparency of provisioning.

In contrast to the original model, the 2004 approach includes a β parameter that interacts with the specific provision (see Table 5). β is the historical average specific provision for each of the homogenous groups. The interaction between β and the specific provision measures the speed with which non-specific provisions become specific for each asset class. C_{it} represents the stock of asset i at time t. The limit of the statistical fund was modified to 1.25 times latent exposure.[27]

Spanish Banks in 2006

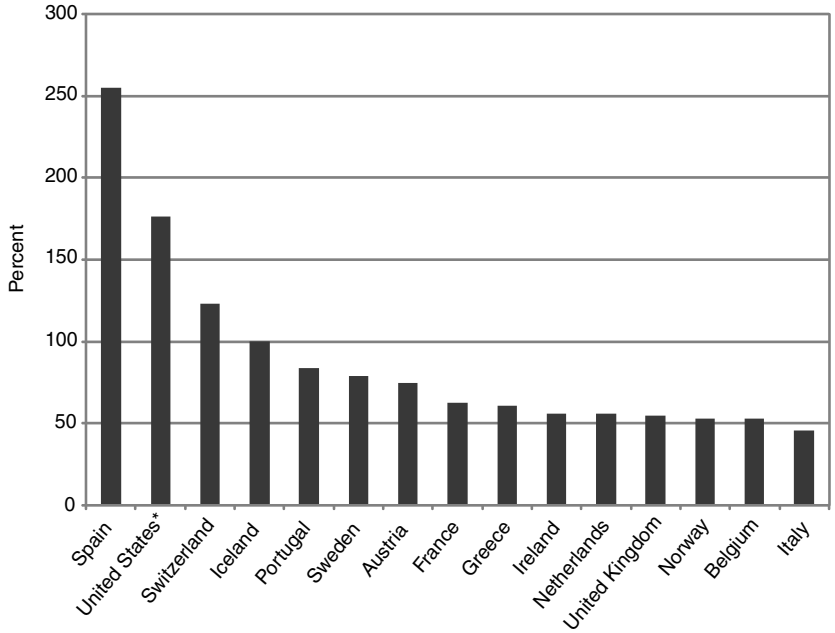
What difference did dynamic provisioning make for Spain?[28] In 2006, the Spanish banking system had by far the highest coverage ratio (the ability of the LLR to cover non-performing loans) among Western European countries at 255 percent. At the same time, the U.S. aggregate coverage ratio was 176 percent.[29] Figure 5 shows the coverage ratios for Spain and the United States and also Spain's European peers as of 2006. To reiterate, 2006 was important

[27] The Bank of Spain defines latent loss generally as the probability of default times loss given default. See Saurina (2009a).

[28] As part of the Financial Sector Assessment Program, the International Monetary Fund (2006) states that the Central Bank of Spain has pioneered a rigorous provisioning system that enhances the safety and soundness of Spanish banks. While we will briefly discuss the condition of Spanish banks in 2006, see Saurina (2009d) for a more detailed review.

[29] The data were obtained from Banco de España (2008), U.S. Call Reports, Saurina (2009c) and IMF cited in Catan and House (2008).

Figure 5 Loan Loss Reserves as a Percentage of Non-Performing Assets



Notes: *The coverage ratio for the United States is the aggregate LLR for all commercial banks as a percentage of the aggregate non-performing loans. It was computed using Call Report data. All other countries' data come from the International Monetary Fund.

for measurement purposes because the primary intent of dynamic provisioning is the timing of provisions. At the onset of the financial crisis and global recession, dynamic provisioning worked as expected in Spain, allowing large Spanish banks like Santander and BBVA to enter the crisis with substantial reserve cushions relative to non-Spanish peers. Many observers, including G20 Finance Ministers, have singled out the Spanish dynamic provisioning system as a contributor to that banking sector's soundness entering the financial crisis of 2007–2009. It is beyond the scope of this article to assess what, if any, fraction of banking sector stability in Spain is attributable to dynamic provisioning versus other Spain-specific factors. Additional factors unique to Spain may have a bearing were this policy to be adopted more widely.

4. DYNAMIC PROVISIONING: A SIMULATION WITH U. S. DATA

One way to illustrate potential inefficiencies of the current U.S. loan loss provisioning framework is to simulate U.S. bank loan loss provisioning over the last two business cycles under an alternate provisioning framework. This section illustrates that a dynamic provisioning framework (akin to that implemented in Spain) could have allowed for a build-up of reserves during the boom years. The results demonstrate that the alternate framework would have smoothed provisions and bank income through the cycle. The severe drop in bank income associated with the actual steep rise in loan loss provisioning during the financial crisis of 2007–2009 would have been substantially reduced. With positive net income in its place, banks could have increased their capital through internal means and thus reduced the need for assistance from the U.S. government. Note that no precise conclusions can be reached as to the magnitude of these effects from the exercise. Any conclusions based on the simulation are subject to the Lucas Critique in that a change in LLR methodology is likely to influence bank lending behavior (Lucas 1976). Lending and other related variables would therefore take on values different from those actually observed and used in the simulation. Nonetheless, the simulation is a useful illustration of the relationship between loan loss provisioning and bank income under the two different frameworks.

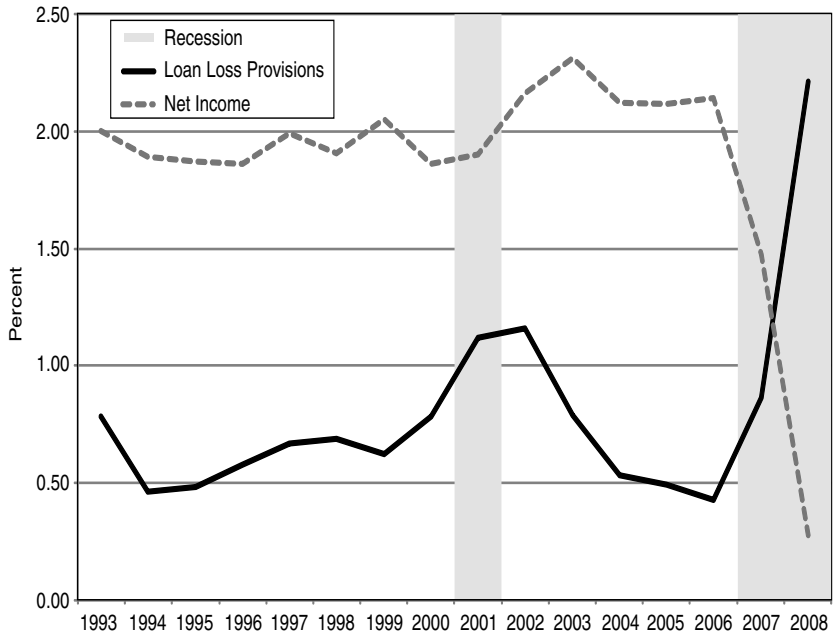
Using aggregate U.S. data from the year-end quarterly FDIC banking reports (key actual data presented in Figure 6),[30] we created an example bank. We have populated the example bank's financial report with aggregates from a hypothetically consolidated U.S. commercial banking industry so as to observe the interaction of a dynamic provisioning approach with historical U.S. banking data. We apply the 2004 Spanish provisioning methodology to the financials of the example bank so as to display and describe the technical nuances of dynamic provisioning. The purpose of this exercise is to demonstrate a well-functioning variant of the policy, that is, how the policy should work, not how it would or would have worked. Accordingly, several simplifying assumptions were required to complete the simulation.

First, we selected representative α and β parameters to compute the statistical provision on a yearly basis. We selected 1.8 percent and 0.65 percent, respectively. While an attempt could have been made to approximate the parameters for the example bank, the available banking data does not permit that level of analysis.[31] Further, these parameters allowed us to show the results under a case where the policy is effective and, by examining the results under

[30] FDIC Quarterly Banking Profile: 1998–2008

[31] The coefficients used in the Spanish standard approach are estimated from a large credit database maintained by the Bank of Spain. No equivalent consolidated credit database exists in the United States as of the publication of this article.

Figure 6 Aggregate U. S. Bank Loan Loss Provisions and Net Income as a Percentage of Total Loans, 1993–2008



Source: FDIC data.

this example, readers should be able to also grasp how the results would vary under the case where the parameters were either too high or low.

Second, we had to select a time period for illustrative purposes. The simulation for the example bank begins in 1993 with five years devoted to building the statistical fund. We targeted a build-up of about 2 percent of total loans by 1999 (close to the largest statistical fund possible of 1.25 α). Subsequent to 1999, the statistical fund was built using only the statistical provision, shown in line 6 of Table 6 (all line references for the simulation are from Table 6). Line 13 is the actual allowance for loan and lease losses (ALLL),[32] while line 14 is the ALLL plus the statistical fund, equal to the total stock of provisions. Line 16, the actual loan loss provision for the year divided by total loans and leases (LLP/TLL), and line 17, the five-year moving

[32] ALLL is the formal accounting term for loan loss reserves.

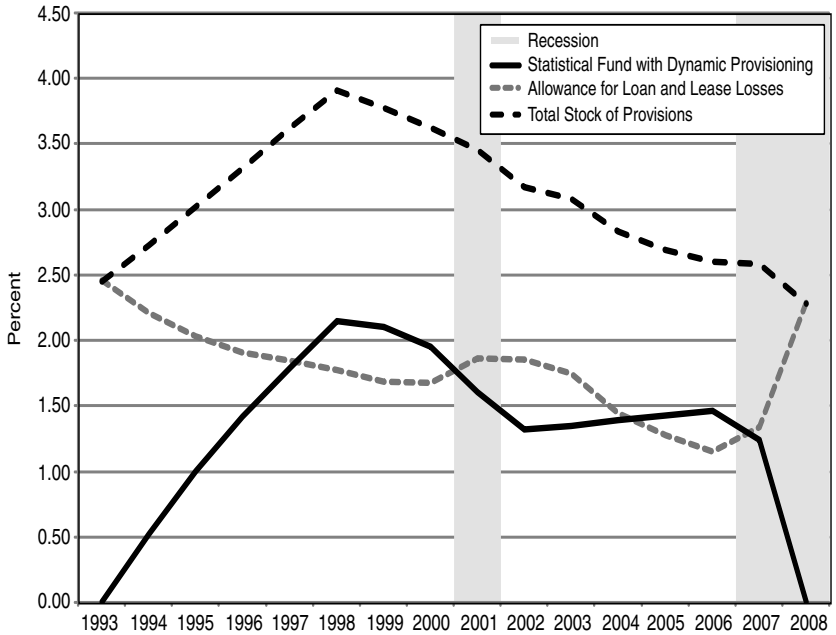
Table 6 Dynamic Provisioning Simulation: Example Bank, 1999–2008 (in millions, except TLL)

Balance Sheet Accounts	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
1 TLL, in billions	3,491	3,820	3,889	4,156	4,429	4,904	5,380	5,981	6,626	6,840
2 Actual Rate of Growth	7.81%	9.40%	1.83%	6.86%	6.55%	10.74%	9.70%	11.17%	10.80%	3.22%
Parameter Scenarios										
3 $\alpha = 1.8\%$										
4 $\beta = 0.65\%$										
5 Statistical Fund, Beginning	69,395	73,238	74,494	62,229	54,873	59,550	67,942	76,329	86,976	82,077
6 Statistical Provision	4,411	5,714	976	4,492	4,677	8,392	8,387	10,646	11,249	2,862
7 Statistical Provision after Shift	3,843	1,256	0	0	4,677	8,392	8,387	10,646	0	0
8 Amount Needed from Statistical Fund	0	0	12,265	7,356	0	0	0	0	4,899	86,404
9 Statistical Fund after Shift	69,395	73,238	62,229	54,873	54,873	59,550	67,942	76,329	82,077	0
10 Statistical Fund Limit	78,554	85,940	87,513	93,519	99,649	110,351	121,051	134,567	149,094	153,900
11 Statistical Fund Limit Reversal	0	0	0	0	0	0	0	0	0	0
12 Statistical Fund, End	73,238	74,494	62,229	54,873	59,550	67,942	76,329	86,976	82,077	0
13 ALLL	58,770	64,137	72,323	76,999	77,152	70,990	68,688	68,984	89,004	156,152
14 Total Stock of Provisions	132,008	138,631	134,552	131,872	136,702	138,932	145,017	155,960	171,081	156,152
Income Statement Accounts										
15 LLP	21,814	30,001	43,433	48,196	34,837	26,098	26,610	25,583	57,310	151,244
16 LLP/TLL	0.62%	0.79%	1.12%	1.16%	0.79%	0.53%	0.49%	0.43%	0.86%	2.21%
17 Five-Year Moving Average	0.61%	0.67%	0.78%	0.87%	0.89%	0.88%	0.82%	0.68%	0.62%	0.91%
18 Amount Needed from Statistical Fund	569	4,458	13,242	11,847	0	0	0	0	16,147	89,266

Table 6 (Continued) Dynamic Provisioning Simulation: Example Bank, 1999–2008 (in millions. except TLL)

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
19 Amount Taken from Statistical Fund	569	4,458	13,242	11,847	0	0	0	0	16,147	84,939
20 New Loan Loss Provision	21,245	25,543	30,191	36,349	34,837	26,098	26,610	25,583	41,163	66,305
21 Profits Without Statistical Provision	71,556	71,009	73,967	89,861	102,440	104,172	114,016	128,217	97,630	18,726
22 Changes in Provisions under Dynamic Provisioning	3,843	1,256	−12,265	−7,356	4,677	8,392	8,387	10,646	−4,899	−82,077
23 Profits With Statistical Provision	67,713	69,753	86,232	97,217	97,763	95,780	105,629	117,571	102,529	100,803

Figure 7 Aggregate U. S. Bank Reserves Compared to Total and Statistical Reserves as a Percentage of Total Loans, 1993–2008

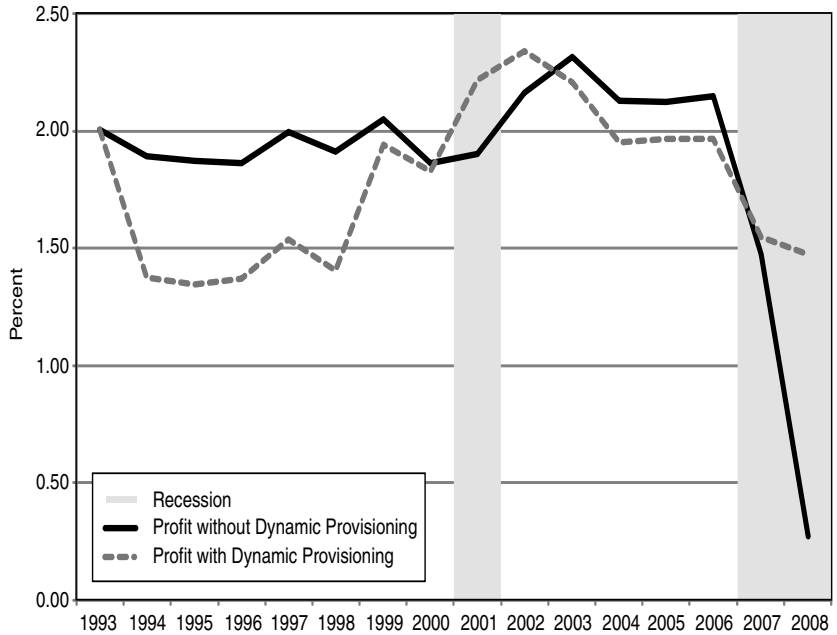


Source: FDIC data.

average of the same ratio, were used to objectively determine if funds were needed from the statistical fund; the difference between a given year's LLP/TLL ratio and the five-year moving average determine exactly how much should be drawn from the statistical fund. The statistical provision in a given year was used to cover as much of the amount needed as possible, and then, if necessary, the remaining amount would be taken from the statistical fund; the amount needed from the statistical fund is listed in line 18, while the actual amount taken is listed in line 19. If funds from the statistical provision and statistical fund were used to cover specific provisions, then the loan loss provision in a given year would change. The new loan loss provision reflecting the changes is shown in line 20, and the changes in provisions under the dynamic provisioning simulation, which has an effect on net income, are shown in line 22. The new net income (profit) under the dynamic provisioning simulation is listed in line 23.

The simulation has several interesting results. First, by allowing the bank to use five years to build the fund (during good times from 1993–1998), the

Figure 8 Aggregate U. S. Bank Net Income Compared to Net Income as a Percentage of Total Loans, 1993–2008



Source: FDIC data.

example bank was able to build a total stock of provisions, both statistical and specific, of 3.9 percent of total loans. When compared to the actual (U.S. aggregate) total allowance for loan and lease losses of 1.7 percent, the bank is clearly better positioned to withstand losses associated with a recession in 1999. The statistical fund was set to be constrained, if necessary, by the aforementioned limit of 1.25 times the latent losses, which was not reached in the simulation.

The second result to note is the impact of a recession, or increase in loan losses, on the statistical fund. Figure 7 shows total provisions and the statistical fund compared to the actual allowance for loan and lease losses. Both the 2001 and current recessions result in declines in the statistical fund, driving down total provisions, compared to large increases in the actual allowance for loan and lease losses. Since the statistical provision is recorded against the income statement, it reduces profits, as demonstrated in the first example

that contrasted dynamic provisioning to the incurred loss model.[33] Figure 8 shows how profits under statistical provisioning compare to actual profits, emphasizing the impact this has on earnings.

The simulation's primary purpose was to depict the results of dynamic provisioning under one possible scenario with representative parameters selected to demonstrate how the policy should work. The selected parameters, combined with the other simplifying assumptions, allow a sizeable fund of statistical provisions to be built in good times and absorb losses in bad times. However, if representative parameters were selected that were too low, higher levels of specific provisioning would have prevented the appropriate fund from being built. Conversely, if higher parameters were selected, then the fund built would have been larger, which could have resulted in inefficiently used capital given the size of realized losses over the time horizon for which the bank provisioned under the policy. This is an important point because the standard approach in the Spanish system is dependent on the parameters estimated by the Bank of Spain.[34] The cost of underestimating the risk weights relates to bank-specific solvency, whereas an overestimation of the risk weights is a tax to the bank and could reduce the overall supply of credit.

The simulation illustrates how dynamic provisioning prepared the example bank to weather the economic downturn while remaining profitable and leaving its capital levels in periods of economic growth accurately stated and allowing managers to further build capital if deemed necessary.

5. CONCLUSIONS

Current accounting guidelines require banks to recognize losses prompted by events that make the losses probable and estimable. But this method may be at odds with the bank regulators' desire for banks to build a "war chest" of reserves in good times to be depleted in bad times. In 1998, the SEC required SunTrust Bank to restate its LLR by $100 million, reflecting the SEC's aversion to what it considered to be an overstated reserve. Remarkably, the action taken by the SEC wasn't long after the banking crises of the 1980s and early 1990s, which, in 1990 alone, resulted in a total provision for loan losses almost twice bank profits (Walter 1991). By the same measure, the financial crisis that began in 2007 was far more severe, resulting in total loan loss provisions greater than eight times bank profits in 2008.[35] Following the current episode,

[33] Note that both examples treat loan growth rate as independent of reserve rules. Dynamic provisioning is likely to smooth out loan growth expansion and contraction. One could argue that the rules could impose a high enough cost as to result in lower loan volume through the credit cycle.

[34] Even banks that use internal models are subject to approval by the Bank of Spain. Presumably, the approval process could involve some reference to the standard risk weights.

[35] FDIC Quarterly Banking Profile, December 31, 2008.

Table 7 Aggregate Condition and Income Data: All Commercial Banks, 2008 (in millions)

Assets		Liabilities and Capital	
Total Loans and Leases	$6,839,998	Deposits	$8,082,104
Less: Reserve for Loss	**156,152**	Other Borrowed Funds	2,165,821
Net Loans and Leases	6,683,846	Subordinated Debt	182,987
Securities	1,746,539	All Other Liabilities	724,353
All Other Assets	3,882,529	Equity Capital	1,157,648
Total Assets	12,312,914	Total Liabilities and Capital	12,312,914

Source: FDIC Quarterly Banking Profiles, 2008.

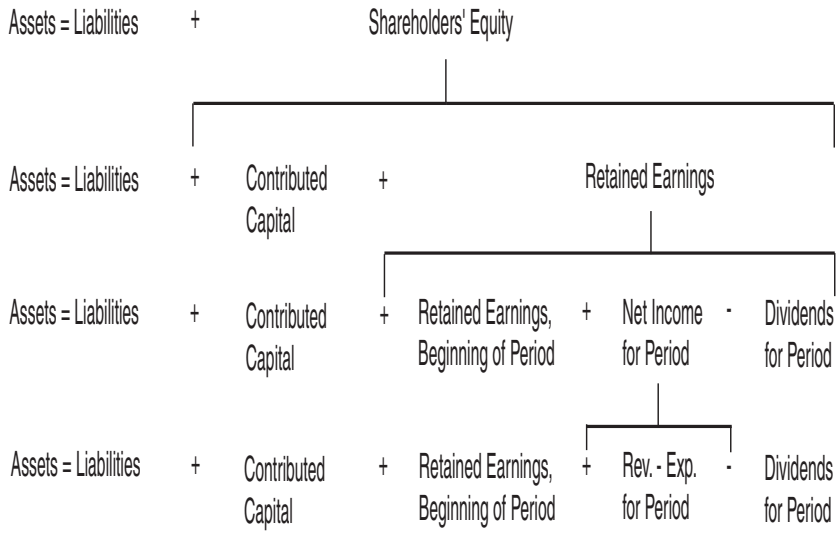
many have called for revisiting the loan loss provisioning method so as to allow for the "war chest" approach and to mitigate the current approach's procyclical effects.

We compared the incurred loss (as implemented in the United States) and dynamic provisioning (as implemented in Spain) approaches. Through a basic example we illustrated that the key difference is not the aggregate level of provisioning but the timing of the provisioning. We conducted an empirical simulation to illustrate that a dynamic provisioning framework, like the one implemented in Spain, could have allowed for a build-up of reserves during the boom years in the United States. The results demonstrate that the alternate framework would have smoothed bank income through the cycle and the need to provision for loan losses would have been significantly lower during the financial crisis of 2007–2009.

Both in the context of a conceptual framework and through an empirical simulation, we highlighted that dynamic provisioning could mitigate some of the problems associated with the current U.S. system. Even so, it would be premature to advocate the adoption of dynamic provisioning in the United States. We have three reasons for this opinion. First, the distortions embedded in the current U.S. system need to be more fully understood and quantified. Second, although 2006 is a relevant point in time for measuring the efficacy of the Spanish policies, Spanish banking needs to be studied through this cycle and the lessons learned need to be included in potential reform. Third, and most important, the U.S. approach to loan loss reserves should not be reformed independently of other bank capital regulation reform.[36]

[36] To allow for a focus on LLR issues, this article abstracted away from any discussion of broader reform of regulatory capital.

Figure 9 The Balance Sheet



Source: Stickney and Weil 2007.

APPENDIX: ACCOUNTING FOR LOAN LOSSES

The primary function of banking is to collect depositors' funds to provide them with transactional support and savings. Banks use depositors' funds to invest in loans and securities that provide the yields necessary to conduct their operations. As shown in Table 7, as of the second quarter of 2009 loans and securities represented 54.3 and 14.2 percent of total assets for all commercial banks, respectively, while deposits represented 65.6 percent of total liabilities and capital.

As discussed in the subsection entitled "Incurred Loss Accounting," accounting for loan losses is handled under FAS 114 and 5. Under FAS 114, bank managers recognize individually identified losses—that is, losses that managers believe are probable and can be reasonably estimated. FAS 5 provides for assessment of losses of homogeneous groups of loans and, similar to FAS 114, should be probable and estimable.[37] In Table 7, listed below

[37] Provided in GAAP guidance in March 1975.

Figure 10 Balance Sheet Accounts

Any Asset Account		Any Liability Account		Any Shareholder's Equity Account	
Beginning Balance			Beginning Balance		Beginning Balance
Increases + Debit	Decreases - Credit	Decreases - Debit	Increases + Credit	Decreases - Debit	Increases + Credit
Ending Balance			Ending Balance		Ending Balance

the total loans and leases, is the reserve for losses, an account that represents the total value of loans and leases that managers expect is probable to not be collected as of the balance sheet data.

To fully understand the reserve for losses, a brief review of accounting basics is helpful. Figure 9 helps provide a simplified understanding of the balance sheet accounts: assets, liabilities, and shareholders' equity. The most important concept is the relationship between the income statement accounts (revenue and expenses) and shareholders' equity. Net income for a firm is computed as revenues for the period minus expenses. Net income flows to shareholders' equity, dividends are paid, and the remaining funds are added to the accumulated retained earnings. And, since retained earnings and contributed capital make up shareholders' equity, negative retained earnings will reduce total capital.

Another important concept is that of credits and debits. As shown in Figure 10, a debit serves to increase an asset account, while it reduces a liability and shareholders' equity account. That is, referencing Table 7, loans and leases, securities, and all other assets are increased with a debit to their respective accounts and therefore, at yearend, should have a remaining debit balance. A credit to those accounts would serve to reduce the value of assets on the balance sheet.

In accounting, there are contra-accounts—assets, liabilities, or shareholders' equity accounts that have a negative balance for the account; the opposite of where Figure 10 has the beginning balances shown. The reserve for loan losses is an example of a contra-asset. As previously mentioned, the purpose of

Table 8 Aggregate Condition and Income Data: All Commercial Banks, 2008 (in millions)

Income Data	
Total Interest Income	$530,513
Total Interest Expense	210,569
Net Interest Expense	319,944
Provision for Loan and Lease Losses	**151,244**
Total Noninterest Income	193,853
Total Noninterest Expense	329,050
Securities Gains (Losses)	−14,066
Applicable Income Taxes	6,163
Extraordinary Gains, Net	5,452
Net Income	18,726
Net Charge-offs	87,990
Cash Dividend	42,724
Retained Earnings	22,355

the reserve for losses is to accurately reflect the value of total loans and leases on the balance sheet. As a contra-asset—meaning it has a negative (credit) asset balance—it serves to reduce the value of assets by the amount that bank managers believe it will not collect; hence, it reflects the true value of loans and leases to the bank. The reserve for losses is increased on a quarterly basis when losses are recognized and charged against income through an account called the provision for loan and lease losses. The income statement results for all FDIC-insured commercial banks (Table 8) show that the total provision for loan and lease losses was $151,244, greater than eight times bank profits (net income).

The provision for loan and lease losses is an expense account. All income statement accounts (i.e., revenues and expenses) affect shareholders' equity. An expense account is a debit to shareholders' equity and reduces its value. Therefore, the provision for loan and lease losses reduces net income and shareholders' equity.

REFERENCES

Ahmed, Anwer S., Carolyn Takeda, and Shawn Thomas. 1999. "Bank Loan Loss Provisions: A Reexamination of Capital Management, Earnings Management, and Signaling Effects." *Journal of Accounting and Economics* 28 (November): 1–25.

Angklomkliew, Sarawan, Jason George, and Frank Packer. 2009. "Issues and Developments in Loan Loss Provisioning: The Case of Asia." *BIS Quarterly Review* December: 69–83.

Banco de España. 2008. "Report on Banking Supervision in Spain." www.bde.es/informes/be/supervi/2008/MBS2008.pdf.

Benston, George J., and Larry D. Wall. 2005. "How Should Banks Account for Loan Losses?" Federal Reserve Bank of Atlanta *Economic Review* Q4: 19–38.

Berger, Allen N., and Gregory F. Udell. 2003. "The Institutional Memory Hypothesis and the Procyclicality of Bank Lending Behavior." *Journal of Financial Intermediation* 13 (October): 458–95.

Bernanke, Ben S., and Cara S. Lown. 1991. "The Credit Crunch." *Brookings Papers on Economic Activity* 22: 205–48.

Bernanke, Ben. 2009. "Financial Reforms to Address Systemic Risk." Remarks at the Council on Foreign Relations. Washington, D.C.: March 10, 2009.

Bouvatier, Vincent, and Laetitia Lepetit. 2008. "Banks' Procyclical Behavior: Does Provisioning Matter?" *Journal of International Financial Markets, Institutions, and Money* 18 (December): 513–26.

Catan, Thomas, and Jonathan House. 2008. "Spain's Bank Capital Cushions Offer a Model to Policy Makers." *The Wall Street Journal*, 10 November, A12.

Cohen, Adam. 2009. "EU Ministers Criticize Banking Rules." *The Wall Street Journal*, 8 July.

Davis, Peter O., and Darrin Williams. 2004. "Credit Risk Measurement: Avoiding Unintended Results: Part 4: Loan Loss Reserves and Expected Losses." *The RMA Journal*, October.

Dugan, John. 2009. "Loan Loss Provisioning and Procyclicality." Remarks before the Institute of International Bankers, March 2, 2009.

Eisenbeis, Robert. 1998. "Comment on Hancock and Wilcox." *Journal of Banking and Finance* 22 (August): 1,015–7.

Fernández de Lis, Santiago, Jorge Martínez Pagés, and Jesús Saurina. 2000. "Credit Growth, Problem Loans and Credit Risk Provisioning in Spain." Working Paper 0018. The Bank of Spain.

Financial Stability Forum. 2009. "Report of the FSF Working Group on Provisioning." www.financialstabilityboard.org/publications/r_0904g.pdf.

Garcia-Herrero, Alicia, and Santiago Fernández de Lis. 2008. "The Housing Boom and Bust in Spain: Impact of the Securitization Model and Dynamic Provisioning." Working Paper 0808, Economic Research Department, BBVA.

Hancock, Diana, and James Wilcox. 1998. "The 'Credit Crunch' and the Availability of Credit to Small Businesses." *Journal of Banking and Finance* 22 (August): 983–1,014.

International Monetary Fund. 2006. "Spain: Financial System Stability Assessment, including Reports on the Observance of Standards and Codes on the following topics: Banking Supervision, Insurance Supervision, Securities Supervision, Payment Systems, Securities Settlement Systems, and Financial Policy Transparency." IMF Country Report No. 06/212.

Laeven, Luc, and Giovanni Majnoni. 2003. "Loan Loss Provisioning and Economic Slowdowns: Too Much, Too Late?" *Journal of Financial Intermediation* 12 (April): 178–97.

Lucas Jr., Robert E. 1976. "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy* 1 (January): 19–46.

Mann, Fiona, and Ian Michael. 2002. "Dynamic Provisioning: Issues and Application." *Financial Stability Review* December: 128–36.

Peek, Joe, and Eric Rosengren. 1995. "The Capital Crunch: Neither a Borrower, Nor a Lender Be." *Journal of Money, Credit and Banking* 27 (August): 625–38.

Pennacchi, George G. 2005. "Risk-based Capital Standards, Deposit Insurance and Procyclicality." *Journal of Financial Intermediation* 14 (October): 432–65.

Poveda, Raimundo. 2000. "Reform of the System of Insolvency Provisions." Speech given at the APD, Madrid: January 18, 2000.

Rajan, Raghuram G. 1994. "Why Bank Credit Policies Fluctuate: A Theory and Some Evidence." *Quarterly Journal of Economics* 109 (May): 399–441.

Saurina, Jesús. 2009a. "Dynamic Provisioning: The Experience of Spain." Crisis Response: Public Policy for the Private Sector, Note Number 7 (July). The World Bank Group.

Saurina, Jesús. 2009b. "Made in Spain – and Working Well." *Financial World*, April, 23–4.

Saurina, Jesús. 2009c. "The Issue of Dynamic Provisioning: A Case Study." Presentation to the Financial Reporting in a Changing World, European Commission Conference, Brussels: May 7–8.

Saurina, Jesús. 2009d. "Loan Loss Provisions in Spain. A Working Macroprudential Tool." www.bde.es/webbde/Secciones/Publicaciones/InformesBoletinesRevistas/RevistaEstabilidadFinanciera/09/Noviembre/ief0117.pdf.

Stickney, Clyde P., and Roman L. Weil. 2007. *Financial Accounting: An Introduction to Concepts, Methods, and Uses*. Mason, Ohio: Thomson Higher Education.

Wall, Larry D., and Timothy W. Koch. 2000. "Bank Loan Loss Accounting: A Review of Theoretical and Empirical Evidence." Federal Reserve Bank of Atlanta *Economic Review* Q2: 1–20.

Walter, John R. 1991. "Loan Loss Reserves." Federal Reserve Bank of Richmond *Economic Review* 77 (July): 20–30.

Economic Quarterly—Volume 95, Number 4—Fall 2009—Pages 419–454

The U.S. Establishment-Size Distribution: Secular Changes and Sectoral Decomposition

Samuel E. Henly and Juan M. Sánchez

Establishment heterogeneity has been modeled in economics at least since the seminal work of Lucas (1978). More recently, this feature has been incorporated into calibrated models to provide quantitative evaluations of different mechanisms. This article aims to contribute to this literature by providing a set of facts about the establishment-size distribution since the 1970s that may be used to calibrate and test the predictions of these models.

First, this article analyzes establishment data from 1974–2006.[1] During this period, the number of workers (size) of a "representative establishment" is relatively constant. Next, the analysis turns to the dispersion of establishment sizes. The size distribution of establishments has become slightly more even. The same analysis is then applied at the sector level. Service establishments became larger and service labor became more concentrated in large establishments while opposite trends were observed in manufactures. Although these *intrasector* shifts played an important role in explaining aggregate movements, *intersector* changes were also found to be important. Finally, this article considers whether trends in the firm-size distribution resemble those

We gratefully acknowledge comments from Anne Stilwell, Devin Reilly, Kartik Athreya, Marianna Kudlyak, and Ned Prescott. All remaining errors are our own. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: sam.henly@rich.frb.org; juan.m.sanchez@rich.frb.org.

[1] Two alternative production units will be considered—firms and establishments. A firm may be a collection of establishments. For instance, Walmart is one firm but it has more than 4,000 establishments in the United States.

found in establishments. They are similar, although labor became slightly more concentrated in large firms.

Davis and Haltiwanger (1989) also analyze secular trends at the establishment level.[2] In particular, they study changes in the establishment-size distribution during the period 1962–1985. First, they study how workers are distributed across establishments; they find that the "representative" worker was working in a larger establishment in 1962 than in 1985. Second, they consider the establishment-size distribution; conversely, they find that the "representative" establishment was smaller in 1962 than in 1985.[3] The opposite behavior of these series reveals a decline in the dispersion of establishment size. Davis and Haltiwanger also decompose these changes by sector. They find that "changes in the industry distribution of employment and movements in the employee size distribution within the average two-digit industry make roughly equal contributions to the secular shift towards mid-size establishments in the aggregate economy." This article extends part of their work through 2006 and complements it with an analysis of firm data and alternative statistics, figures, and decompositions. The earlier change in the first moments contrasts with the finding in this article, while the downward trend in the dispersion of establishment size continued after 1985.

Buera and Kaboski (2008) also study the evolution of the scale of production and sectoral reallocation. They emphasize the difference between the size distribution for manufactures and services establishments. Additionally, they present evidence of the rise in the size of service establishments and the reallocation of resources from manufacturing to services.[4] Our article extends their analysis by studying changes in the size distribution of manufacturing and service establishments over time.

Several studies take an interest in which distribution best fits the firm-size distribution. Gibrat (1931) finds that the log-normal distribution effectively described French industrial firms. This distribution is a consequence of the "law of proportional effect," also known as Gibrat's Law, whereby firm growth is treated as a random process and growth rates are independent of firm size (Sutton 1997). As noticed by Axtell (2001), census data display monotonically increasing numbers of progressively smaller firms, a shape the log-normal distribution cannot reproduce. Using data from the U.S. Census Bureau from 1988–1997, Axtell (2001) shows that firm size is approximately Zipf-distributed. Although we find that the aggregate distribution is relatively

[2] See also Davis and Haltiwanger (1990) and Davis, Haltiwanger, and Schuh (1996).

[3] Our article also considers the distribution of employees by establishment size and the distribution of establishments by size. Notice that while the latter describes which proportion of the establishments is of a given size, the former studies which proportion of employees work in an establishment of a given size.

[4] They also show evidence of sectoral reallocation for 30 countries.

stable, results for manufacturing and services suggest that it would be interesting to extend Axtell's analysis to the sectoral level.

Recent articles use establishments data to study economic development. They argue that the misallocation of resources among heterogeneous establishments may be a key determinant of cross-country income differences. Banerjee and Duflo (2005) conclude that "the microeconomic evidence indeed suggests that there are some sources of misallocation of capital, including credit constraints, institutional failures, and others." Restuccia and Rogerson (2008) illustrate this mechanism using a model with establishment heterogeneity similar to Hopenhayn and Rogerson (1993). In a similar framework, Hsieh and Klenow (2007) find that productivity would increase by 30–50 percent in China and 40–60 percent in India "if capital and labor were reallocated to equalize marginal products across plants to the extent observed in the U.S." Similarly, Greenwood, Sánchez, and Wang (2008) study the role of informational frictions for economic development in a model with establishments heterogeneity.[5] All the theories above analyze mechanisms that may contribute to an explanation of differences in income across countries. The calibrations of these and similar models generally use targets from the size distribution. For instance, Restuccia and Rogerson (2008) use the 2000 establishment size distribution and Greenwood, Sánchez, and Wang (2008) use the Lorenz curve for the distribution of employment by establishment size for 1974. The subsequent sections of this article present evidence for size distributions of establishments and firms and supply a set of stylized facts that new theories in this strand of literature may find useful as calibration targets. Perhaps more importantly, these sections analyze secular changes in the size distribution that could be used to test the predictions of these models. For example, we find that the average size of establishments is fairly constant (or slightly decreasing) over the last 30 years. This finding supports models in which the average size is constant on the balanced-growth path.

The remainder of the article is organized as follows. Section 1 introduces and summarizes our findings. Section 2 describes the secular changes in the establishment-size distribution. The decomposition of secular changes into changes in the sectoral composition (intersector) and distribution changes within each sector (intrasector) is undertaken in Section 3. A description of the data on firms, as an alternative to establishments, is presented in Section 4. Finally, Section 5 concludes. An Appendix presents detailed information about data sources, formulae used to compute the statistics, and some figures and tables.

---

[5] See also Caselli and Gennaioli (2003); Amaral and Quintin (2007); Alfaro, Charlton, and Kanczuk (2008); Bartelsman, Haltiwanger, and Scarpetta (2008); Buera and Shin (2008); Guner, Ventura, and Yi (2008); and Castro, Clementi, and McDonald (2009).

1. PRODUCTION UNIT SIZE TRENDS, 1970–2006

In the sections below, several statistics are defined and used to evaluate the distributions of productive units and their workers from the 1970s to 2006. The aggregate economy, as well as two component sectors (manufacturing and services), are considered in each analysis.

Section 2 develops statistics and functions that are used in the analysis of trends in establishment size and shifts in the dispersion of establishments and workers. We find that the aggregate establishment size changes negligibly. Manufacturing establishments are very large and shrink over time, while service establishments are initially smaller than average but become much larger by 2006. Variation of establishment size does not change significantly apart from a small increase in the service sector. The distribution of employees across establishments becomes slightly more even. This trend is driven by the decline of large manufacturing firms and dampened by increased labor concentration in services.

Section 3 decomposes, by sector, several statistics introduced in Section 2. The results are used to disentangle changes in aggregate statistics caused by intrasector distribution movements from those caused by shifts in the sectoral composition of the aggregate (intersector changes). We find that both intra- and intersector movements are important, but the importance of each varies by statistic.

Section 4 examines the question of whether and when firm distribution patterns should resemble those found in establishments. We argue that movements in establishment distributions should be more similar to those in firms when large firms are composed of relatively large establishments, and present evidence is consistent with this hypothesis. Trends in the aggregate and sectoral distributions of firms and employees across firms generally conform to trends at the establishment level.

2. SECULAR CHANGES IN THE SIZE DISTRIBUTION OF ESTABLISHMENTS

The U.S. Census Bureau (USCB) publishes annual data on establishments in their County Business Patterns series. This section presents a variety of statistics derived from these data. The statistics describe the size distribution of establishments and the dispersion of labor and establishments across establishments. Major trends in these statistics since 1974 are noted and depicted in Figures 1–8.

County Business Patterns Data

County Business Patterns (CBP), released by the USCB annually since 1964, contains tables listing establishment quantity, worker quantity, and payroll by

Table 1 Example Establishment Data

Size Group	Establishment Size	Number of Establishments
1–2 Workers (Small)	1	5
	2	2
3–4 Workers (Large)	3	2
	4	1

establishment size groups. For example, CBP tables in any given year list the number of establishments employing 20–49 workers, the number of people employed by those establishments, and other data (like payroll) not used in this article. Similar data are provided for other establishment size groups (1–4 workers, 5–9 workers, etc.). This information is given for the aggregate and also by SIC (1997 and earlier) or NAICS (1998 onward) industry category. We use data for years 1974 and later due to a significant methodological shift taking place between 1973 and 1974.[6]

A caveat is in order. In the service sector, data for years before and after 1997 are not directly comparable: After 1997, an establishment's sector was determined by the North American Industrial Classification System (NAICS), which is not easily reconciled with the Standard Industrial Classification (SIC) system used for the same purpose in previous years.[7] Consequently, analysis of labor concentration across service sector establishments treats SIC years (1974–1997) and NAICS years (1998–2006) separately. The composition of the manufacturing sector also changes with NAICS, but a single series is available under each system and differences are minimal.

Mean Establishment Size and Coworker Mean Size

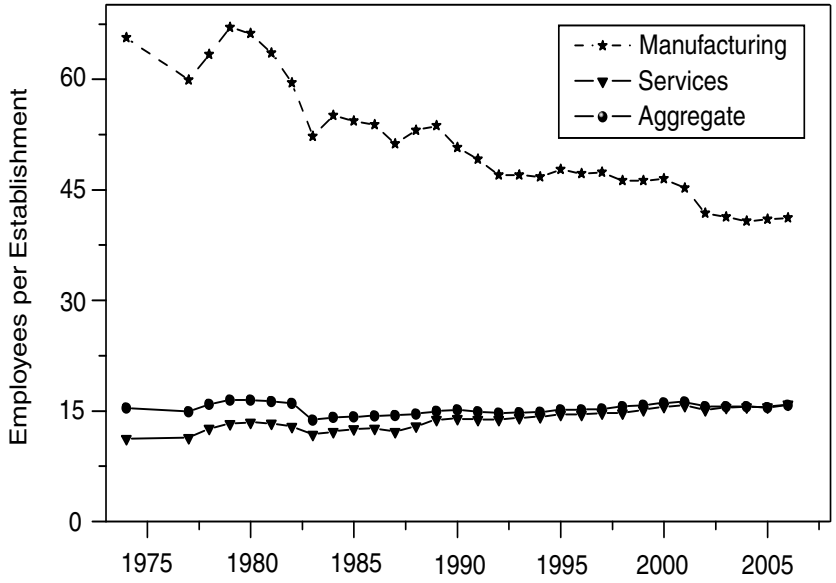
Two different measures of mean size will be considered to describe the size of a "representative" establishment. Given data restrictions, the comparison of these two measures will be used later to study the dispersion of establishments by size.

It may be useful to consider the world described in Table 1, where establishments have between one and four employees (inclusive) and are separated into two size groups: two or fewer workers and three or more workers. In

[6] Some data were retrieved from the National Historical Geographic Information System, an online database operated by the Minnesota Population Center (Ruggles et al. 2009).

[7] Under the SIC system, a single series summing all portions of a "service" sector was available. NAICS split the sector into numerous constituents (educational services; health care and social assistance; professional, technical, and scientific services; and so on). A composite service sector was constructed from these NAICS service subsectors (see Appendix) but it was not possible to precisely recreate the SIC service sector's composition.

Figure 1 Mean Size of Establishments, 1974–2006



this world, a "small" group is comprised of seven establishments employing a total of nine workers; the remaining three establishments form a "large" group employing 10 workers.

Establishment mean size

We begin by asking: What is the average establishment size across establishments? The answer is the mean of the distribution of establishments by establishment size, referred to hereafter as the *mean size of establishments* (or simply as the *establishment mean*) and denoted $E(esize)$. Denote index establishment size groups by i. Then, we obtain the establishment mean by taking a weighted sum of the expected size of establishments within each size group i:

$$E(esize) = \sum_i E(esize \mid egroup = i) * P(egroup = i). \qquad (1)$$

Here, $egroup = i$ is the condition in which an establishment is a member of size group i.[8] Considering our example world, we find that

$$E(esize) = [9/7] * (7/10) + [10/3] * (3/10) = 1.9. \qquad (2)$$

[8] Calculations of expected values and probabilities are detailed in the Appendix.

Figure 1 displays the mean size of establishments between 1974 and 2006. Across the period, this mean changes negligibly: In 1974, the average establishment employed about 15 workers, a figure that ranged between 14 and 16 workers in subsequent years through 2006. This constancy in the aggregate masks significant shifts at the sector level. The average manufacturing establishment size fell from almost 70 employees in the late 1970s to about 41 employees in 2006. The greatest decline occurred between 1979 and 1983, when the average size dropped from 67 employees to 52 employees. In spite of this decline, manufacturing establishments tend to be much larger than other establishments in all years. For instance, in 1974 the average manufacturing establishment employed about 50 more workers than the aggregate economy's average establishment; this gap was halved by 2006. Contemporaneously, the average service sector establishment increased in size, from about 11 workers in 1974 to 14.7 workers in 1997 and from 14.8 workers in 1998 to 16 workers in 2006.

Coworker mean size

What is the average number of coworkers across workers? The answer is the mean of the distribution of workers by establishment size, referred to hereafter as the *coworker mean size of establishments* or simply the *coworker mean*, denoted $E(wsize)$. This statistic is interesting because it may vary even when the mean size of establishments is constant.[9] The following formula can be used to compute this measure:

$$E(wsize) = \sum_i E(wsize \mid wgroup = i) * P(wgroup = i), \qquad (3)$$

where $wgroup = i$ denotes a worker who is employed by an establishment in size group i. In our example, we have data that allow us to compute $E(wsize)$ directly:

$$\begin{aligned} E(wsize) &= [\frac{((1*5)+(2*4))}{9}] * (\frac{9}{19}) + [\frac{((3*6)+(4*4))}{10}] * (\frac{10}{19}) \\ &\approx 2.47. \end{aligned} \qquad (4)$$

Unfortunately, $E(wsize)$ cannot be computed directly from public CBP data because we are unable to obtain $E(wsize \mid wgroup = i)$ without information about the distribution of workers within size groups. We use an alternative method of computation that employs an assumption about the distribution of establishments within size groups.[10]

[9] This was actually the case for the time period studied by Davis and Haltiwanger (1989).
[10] See details in the Appendix.

Figure 2 Coworker Establishments Mean Size, 1974–2006

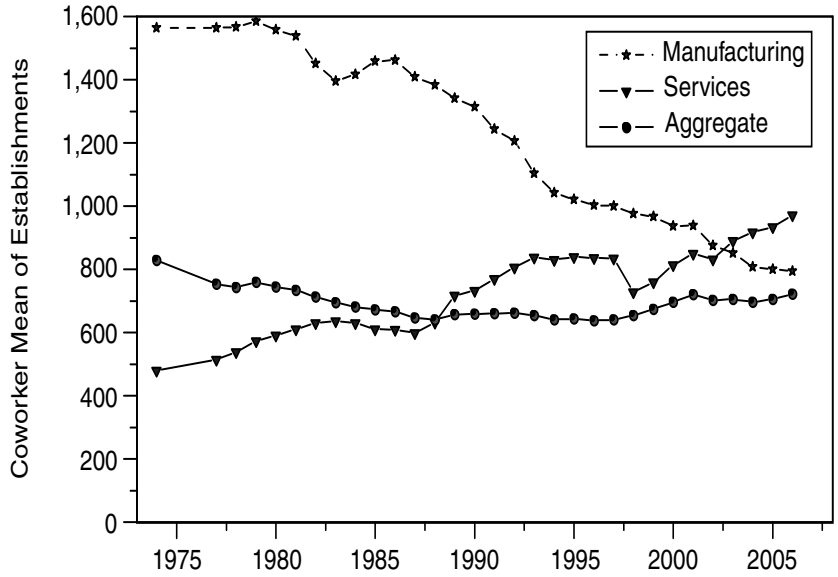


Figure 2 shows the coworker mean size of establishments. As expected, worker mean size is much greater than establishment mean size. In 1974, the worker mean stands around 830 at the aggregate level, 1,560 for manufactures, and 480 for services. Subsequent trends resemble those for the mean size of establishments. The aggregate worker mean remains fairly flat through 2006, dropping 11 percent. Simultaneously, the coworker mean in manufactures is halved (falling from 1,560 to 760) even as the services coworker mean doubles (480 to 970).

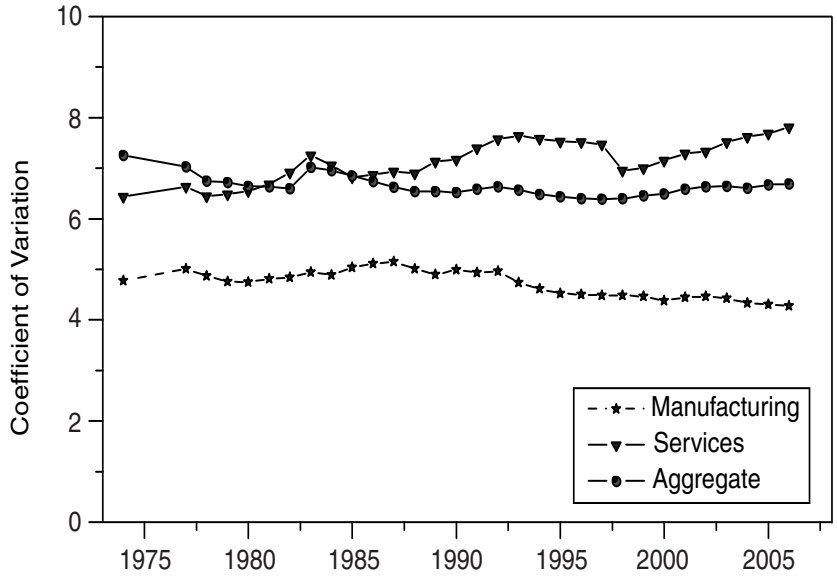
Establishment Size Dispersion and Employment Concentration

Coefficient of variation

The statistic used to analyze the dispersion of establishment size is the coefficient of variation (CV). It measures the dispersion of establishment size relative to the mean size.[11]

[11] This statistic is computed from equation (18) in the Appendix.

Figure 3 Coefficient of Variation of Establishments Size, 1974–2006

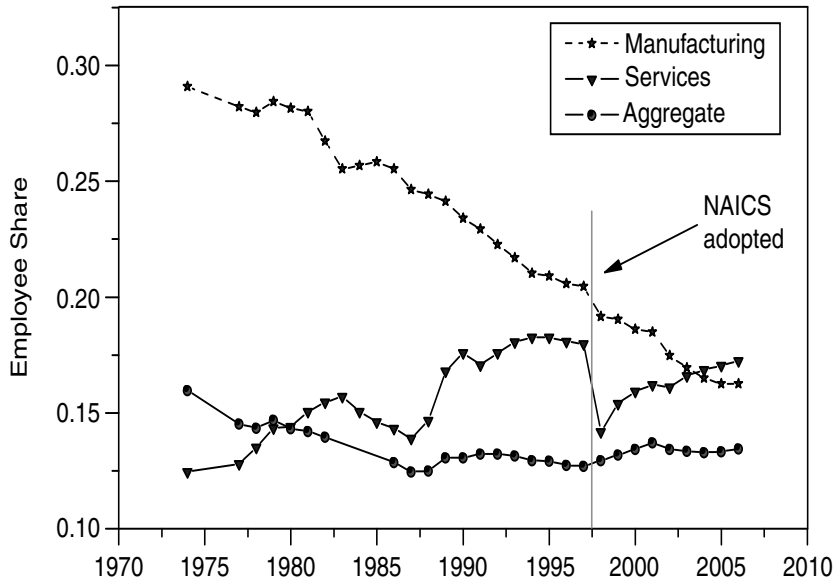


The coefficients of variation for the aggregate and for industries are displayed in Figure 3. In the aggregate, this measure fell about 8 percent from 1974 to 2006 (7.2 to 6.1). The coefficient also fell slightly in the manufacturing sector, from 4.7 to 4.2; note that this figure indicates a much lower variation in establishment size than is present in services or the aggregate. Service establishments actually saw their coefficient increase about 21 percent (6.3 to 7.8).

Large establishment employment share

The fraction of workers employed by very large establishments (those with more than 1,000 workers) serves as a simple measure of labor concentration (Figure 4). In the aggregate this figure decreased slightly. Very large establishments employed about 16 percent of all workers in 1974. By 2006, they were responsible for only 13 percent of employment, although this number had earlier dipped to a 1987 nadir of 12.5 percent. In the manufacturing sector, a decline in large establishment employment share was observed. Large establishments employed 29 percent of manufacturing workers in 1974; in 2006, they employed only 16 percent. Finally, the large establishment share of service labor moved erratically upward. In this sector the employment share

Figure 4 Employment Share of Large Establishments, 1974–2006



increased from 12.5 percent in 1974 to about 18 percent between 1990–1997; from 1998–2006, the share increased from 14 percent to 17 percent.

Lorenz curve

One frequently employed instrument for the analysis of inequality is the Lorenz curve. This measure of the distribution of labor across establishments is independent of the absolute size of establishments. Thus, if all establishments grow or shrink proportionally, there are no changes in the Lorenz curve.

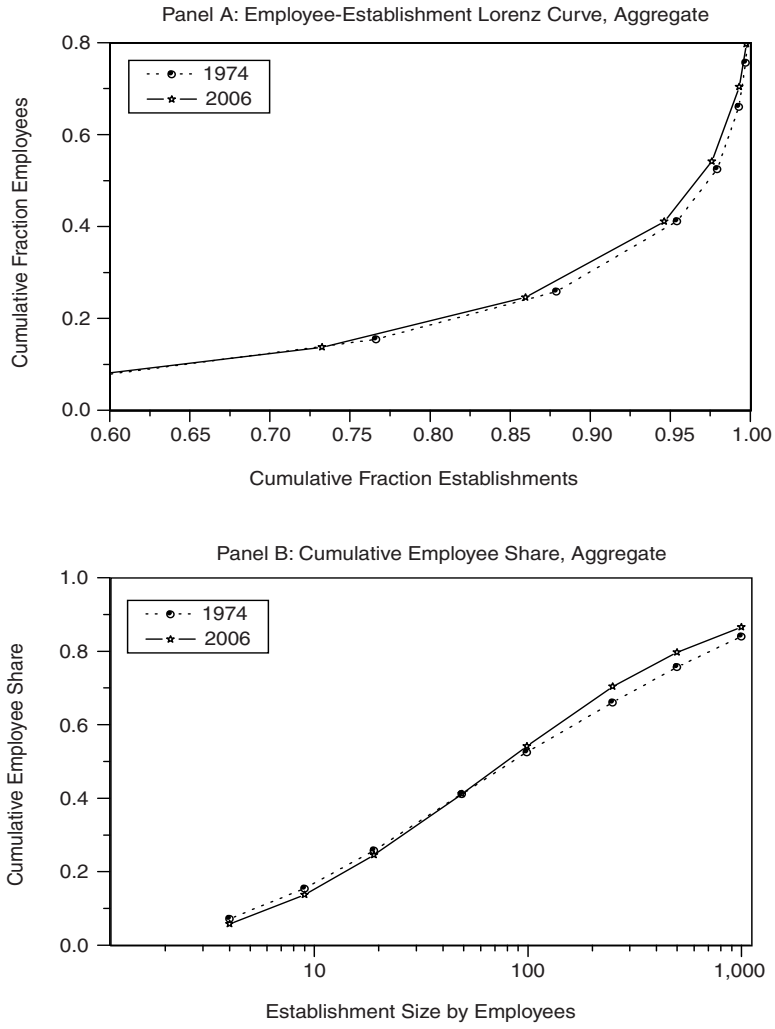
Here, a Lorenz curve represents the fraction y of total workers employed by the fraction x of total establishments employing the smallest number of workers. A 45° line means that all establishments employ the same number of workers; the further a curve is below this line, the greater the unevenness in worker distribution across establishments. Given the data restriction, we have values for the Lorenz function, L, at the upper bound of each size group i:

$$L(P(egroup \leq i)) = P(wgroup \leq i). \tag{5}$$

The function is linearly interpolated elsewhere.

Panel A of Figure 5 shows the Lorenz curve for the distribution of labor across establishments. This curve shifted slightly upward over time, suggesting a decrease in labor concentration. This movement is minor: In 1974, the largest 5 percent establishment employed about 60 percent of the country's

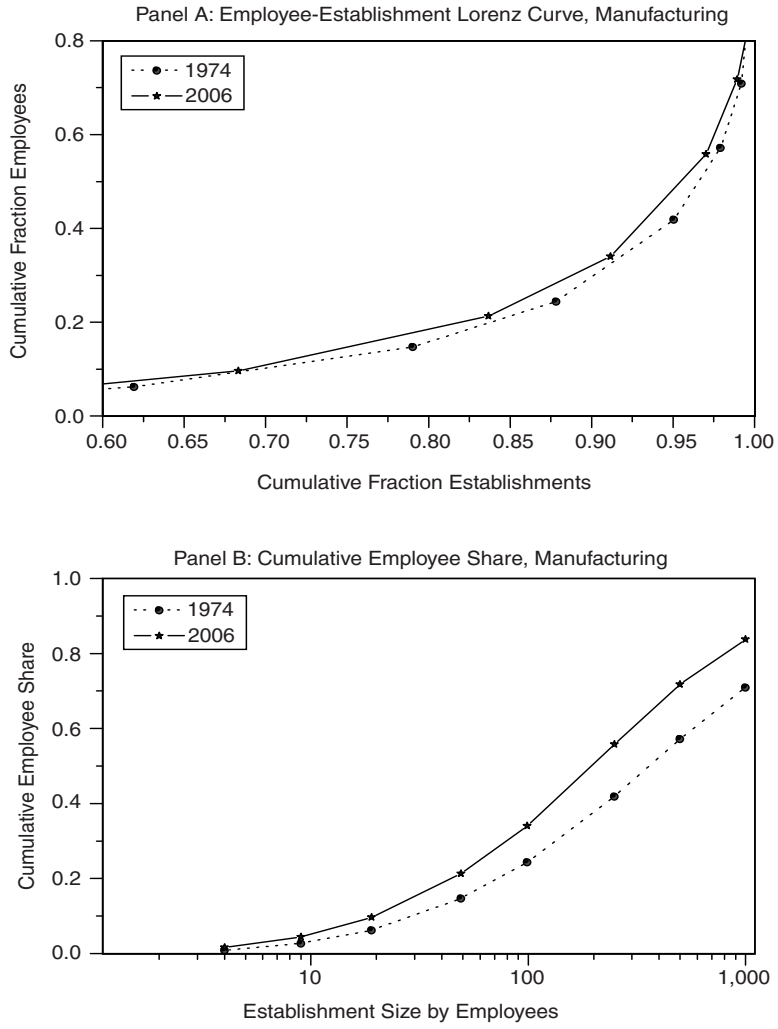
Figure 5 Establishment-Size Distribution; Aggregate Economy, 1974–2006



workers. In 2006, the same icosile employed about 57 percent of the work force.

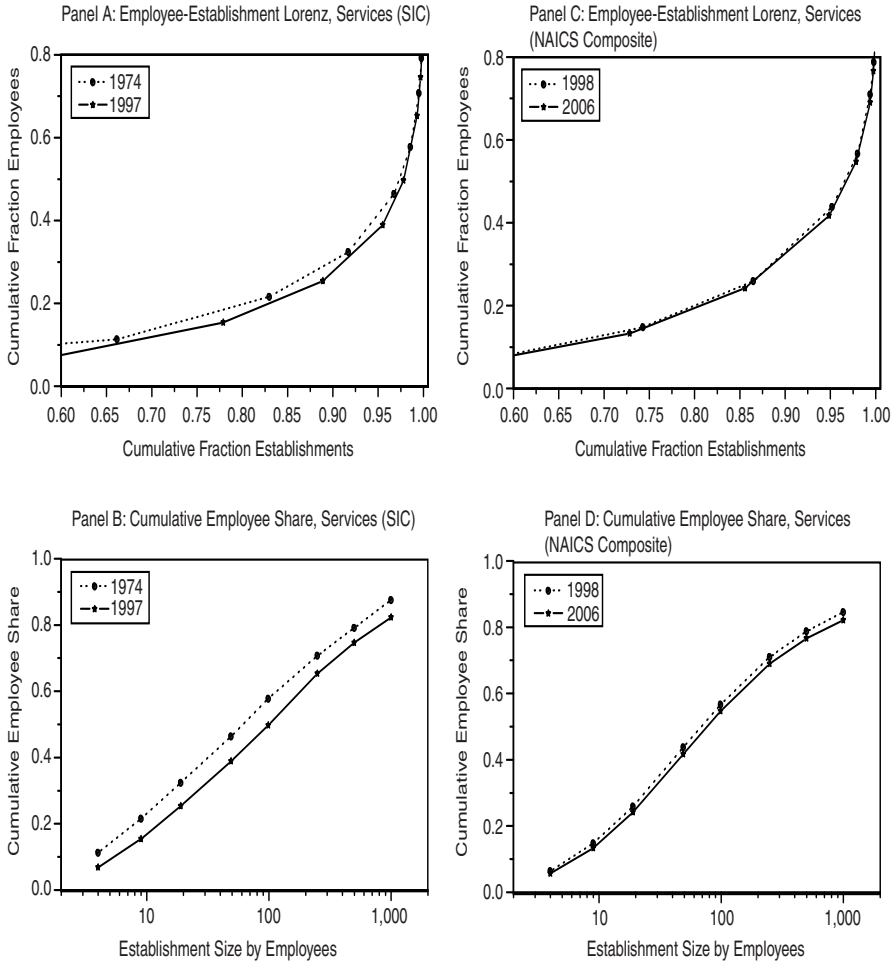
The manufacturing sector's Lorenz curve is found in Panel A of Figure 6. The curve shows a clear shift upward near the top of the scale from 1974 to 2006, as the employee share of the top 5 percent establishments fell from 58.2 percent to 51.7 percent. Workers, then, became more evenly distributed among manufacturing establishments.

Figure 6 Establishment-Size Distribution; Manufacturing Sector, 1974–2006



Service-sector Lorenz curves are located in Panels A and C of Figure 7. Over the SIC years (Panel A) the employee-establishment Lorenz curve shifted downward: The top 5 percent establishments employed about 58 percent of all service workers in 1974 and 62 percent in 1997, reflecting a greater concentration of employment in the largest service establishments. Service labor also became more concentrated in large establishments in the NAICS

Figure 7 Establishment-Size Distribution; Service Sector, 1974–2006



period (Panel C) when the largest 5 percent establishment employment share rose from 1998 (56.6 percent) to 2006 (57.6 percent).

Cumulative employee distributions

To consider the distribution of workers across establishments without explicit disregard for the absolute size of establishments (in contrast to the Lorenz curve), we construct the cumulative distribution function (CDF). This function provides the share of employment held by establishments of or less than a

particular size and is computed at the upper bound of each size group, max_i:

$$CDF(max_i) = P(wgroup \leq i), \tag{6}$$

and then linearly interpolated elsewhere.

Panel B of Figure 5 plots the CDF for the aggregate. This graph shows that the distribution of labor across establishments shifted toward mid-size firms between 1974 and 2006. In 1974, small establishments (10 or fewer employees) and larger establishments (more than 500 employees) are responsible for larger shares of total employment than in 2006. This change is visible as the 2006 curve begins below the 1974 curve but rises more quickly through the mid-size establishments. In both years, employment is nearly evenly divided between establishments with more than and fewer than 100 workers: Establishments with 99 or fewer workers employed 53 percent of the work force in 1974 and 54 percent in 2006.

The cumulative employment curve in Panel B of Figure 6 shows that every size group of manufacturing establishments below 500 workers increased its employee share from 1974 to 2006. Manufacturing establishments employing fewer than 250 workers held 56 percent of the manufacturing employment share in 2006, up from only 42 percent in 1974.

Conversely, in both SIC and NAICS periods, cumulative employment share curves for services (Figure 7, Panels B and D) moved to the right, implying a broad increase in the size of service establishments (recall data in Figures 1 and 2). Establishments employing fewer than 1,000 workers saw their employee share drop from 88 percent to 82 percent between 1974 and 1997 and from 85 percent to 82 percent between 1998 and 2006.

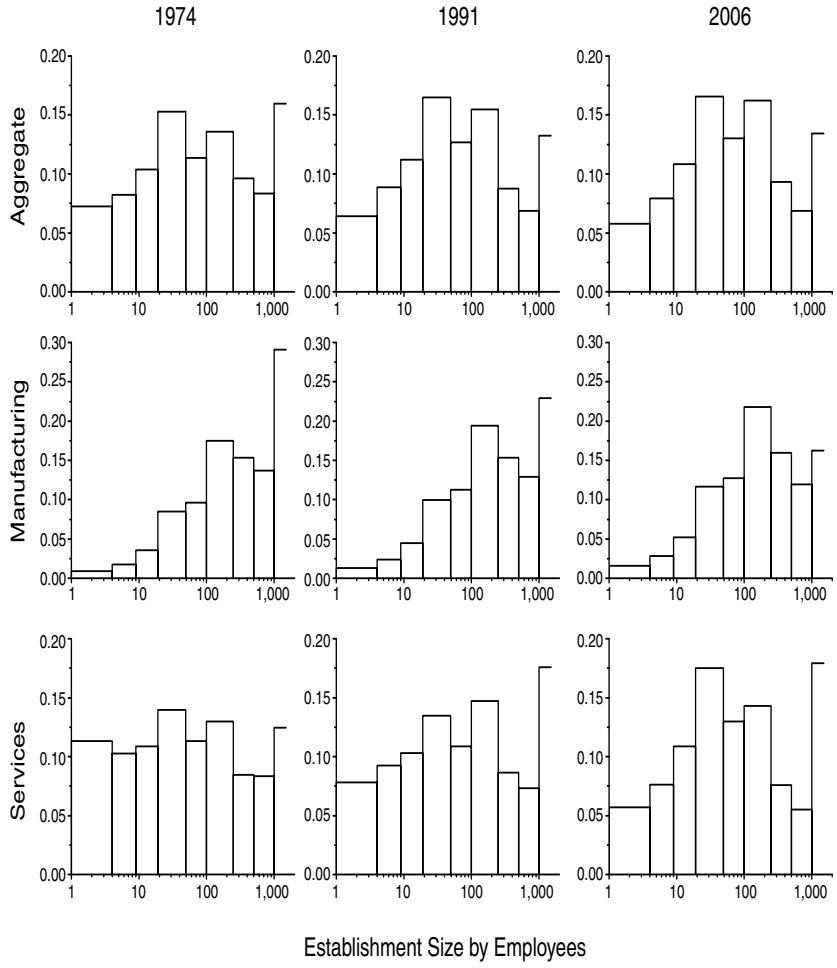
Histograms

While the CDF is useful for revealing shifts in the distribution of labor across establishments, simple histograms of the distribution of labor across establishments are helpful to identify which size groups are actually responsible for those shifts. This function is computed as

$$f((min_i - 1, max_i]) = P(wgroup = i). \tag{7}$$

where min_i and max_i are, respectively, the lower and upper establishment size bounds for size group i. The histogram for distribution of labor among size categories at the aggregate level is depicted in the top row of Figure 8. These histograms show movement of worker share from the smallest and largest establishments into establishments of intermediate size. The employee share of the smallest establishment size group decreases (1–9 workers, 15.5 percent to 13.7 percent) while intermediate size categories see their employee share increase. Establishments with 10–249 workers employed 50.6 percent of the labor force in 1974, and their share increased to 56.7 percent by 2006. Larger establishments (250–999 employees) lose employment share (18 percent to

Figure 8 Histograms for the Distribution of Labor Across Establishments



16.2 percent) as do the largest establishments (1,000 or more employees; 16 percent to 13.4 percent). Large establishments lost the most share before 1991, while small establishments lost the most after 1991.

Figure 8 also contains histograms illustrating the labor distribution across manufacturing establishments. As in previous figures, it is apparent that manufacturing sector employment was less concentrated in large establishments in 2006 than in 1974. Every establishment size group of 499 employees or fewer saw significant increases in its employment share from 1974 to 1991 and again from 1991 to 2006. Establishments employing 100–249 workers saw

the greatest increase over the entire period, employing about 17.5 percent of manufacturing workers in 1974 but 21.8 percent in 2006. By contrast, the size group 500–999 workers saw its employment share decrease from an initial 13.7 percent to 12.0 percent over the same period. This movement is in the same direction as the 13-percentage-point decline in the employment share of manufacturing establishments with more than 1,000 workers.

As noted earlier, the service sector is more difficult to probe due to differences in its composition before and after 1997. The last row of histograms in Figure 8 show that between 1974 and 1991, both years using the SIC service sector, the smallest service establishments (1–19 workers) saw their employee share drop from 32 percent to 27 percent. Intermediate size categories (20–249 workers) increased their employee share slightly, from 38 percent to 39 percent, and the largest size categories depicted (250–999 workers) lost 1 percentage point of total employee share (17 percent to 16 percent). The largest size group (1,000 or more employees) accounted for most of the balance as between 1974 and 1991 its share increased from 12 percent to about 18 percent. A histogram for 2006 shows further erosion in the employment share of the smallest and largest establishments depicted, but these data cannot be directly compared with data from 1974 or 1991.

3. SECTORAL DECOMPOSITION OF SECULAR CHANGES

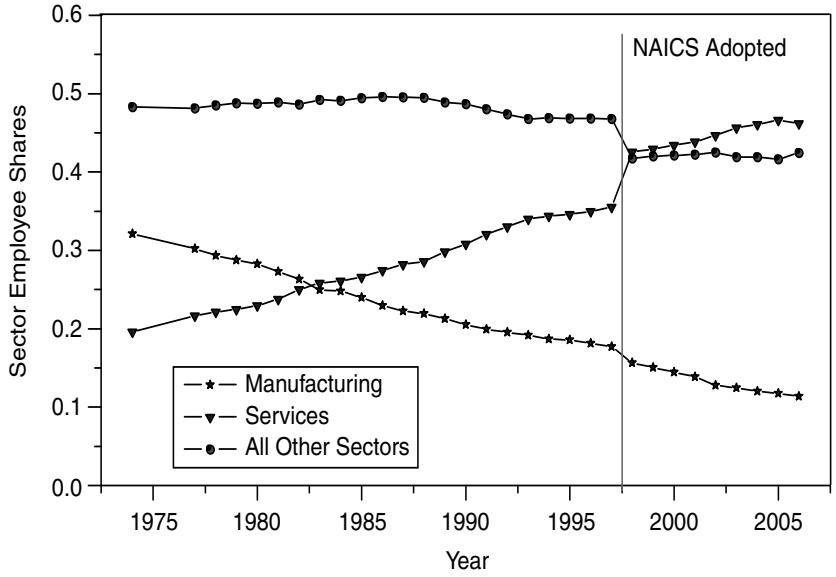
Changes in the Sectoral Composition

Previous sections demonstrated that, broadly speaking, manufacturing establishments have become smaller and service establishments have become larger since the mid-1970s. The distribution of workers became more even across manufacturing establishments and less even across service establishments. These sector level trends offset one another in the aggregate economy. However, to better understand the cause of the slight decline in overall establishment size and labor concentration, it is also necessary to consider changes in the relative share of the service and manufacturing sectors over time.

Two types of effects can be cited as contributors to observed trends in the aggregate distribution of labor across establishments. First are *intrasector* movements of labor; these are described for manufacturing and service sector establishments in the previous section. Intrasector movements of labor include shifts of employment share of different establishment size categories and changes in the dispersion of labor across establishments. The aggregate can also be affected by *intersector* forces as the relative labor and establishment share of different sectors change.

Figure 9 displays the sector shares of total employment from 1974 to 2006, and Figure 10 shows the sector share of establishments for the same period. The pattern is similar in both figures. The participation of other sectors

Figure 9 Employment Share by Sector, 1974–2006



is relatively constant,[12] only decreasing slightly in establishments; service sector participation rose and manufactures participation fell. Changes are more notable in terms of worker shares: manufacturing had 32 percent in 1974 and 11 percent in 2006, while services had 19 percent in 1974 and 46 percent in 2006. During the same period, the establishment share of manufacturing dropped from 8 percent to 4 percent while the services establishment share rose from 27 percent to 47 percent.

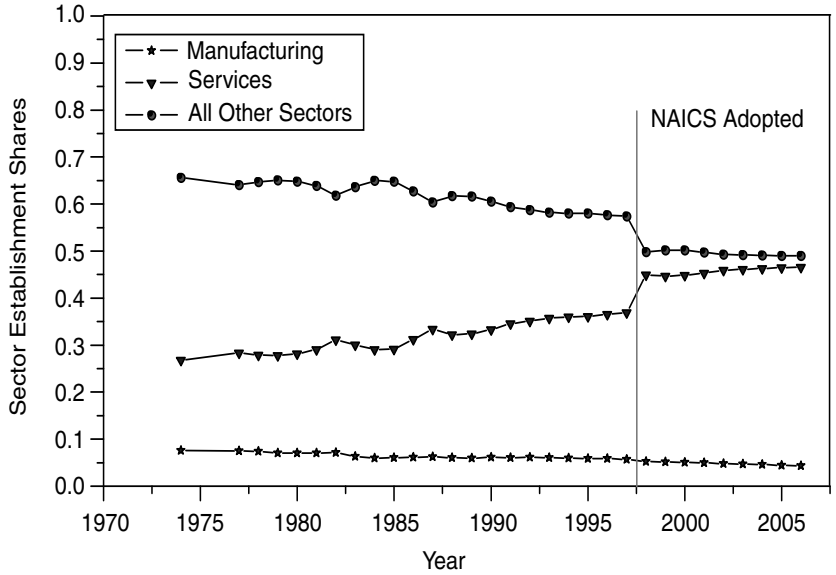
Computation

Any aggregate statistic is a weighted average of the sectoral values of that statistic. Therefore, it can be decomposed into its sectoral constituents. As an example, consider the mean size of establishments, the first statistic that will be decomposed. It can be written as

$$E(esize) = \sum_s E(esize \mid esector = s) * P(esector = s), \qquad (8)$$

[12] The main change seems to be in 1997, when a new sector classification system was adopted (NAICS). Of course, this implies that this change does not have economic meaning. These data were derived from County Business Patterns figures.

Figure 10 Establishment Share by Sector, 1974–2006



where s is a sector index and $esector = s$ denotes that an establishment operates in sector s. By separating services, manufacturing, and the combined other sectors, and simplifying the notation, the mean size of establishments can be written

$$E(esize) = n^{\text{serv}} E^{\text{serv}} + n^{\text{manuf}} E^{\text{manuf}} + n^{\text{other}} E^{\text{other}}, \tag{9}$$

where $E^{\text{s}} = E(esize \mid esector = s)$ and n^{s} is the establishment-share of each sector, $n^{\text{s}} = P(esector = s)$.

This decomposition may be used to answer two questions: (1) What would the value of a statistic (the establishment mean in this example) be if the intersector weights had stayed at their 1974 values? and (2) what value would the statistic have taken if the intrasector value of the statistic had stayed the same as in 1974? The first question is answered by computing a counterfactual statistic,

$$E(\widetilde{esize})_t = n^{\text{serv}}_{1974} E^{\text{serv}}_t + n^{\text{manuf}}_{1974} E^{\text{manuf}}_t + n^{\text{other}}_{1974} E^{\text{other}}_t. \tag{10}$$

Similarly, the second question is answered by computing another counterfactual statistic,

$$E(\widehat{esize})_t = n^{\text{serv}}_t E^{\text{serv}}_{1974} + n^{\text{manuf}}_t E^{\text{manuf}}_{1974} + n^{\text{other}}_t E^{\text{other}}_{1974}. \tag{11}$$

Other statistics can be decomposed in a similar manner. The only difference is that some of them require a different weight, the sector employment share,

Table 2 Sectoral Decomposition of Changes Between 1974–2006

Statistic	Aggregate		Manufactures		Services		Other Sectors	
	Value		Weight	Value	Weight	Value	Weight	Value
Mean Size								
Year 1974	15.447	=	0.076	65.6	0.268	11.3	0.656	11.4
Year 2006	15.776	=	0.044	41.2	0.466	15.6	0.491	13.6
Intrasector	16.263	=	0.076	41.2	0.268	15.6	0.656	13.6
Intersector	13.690	=	0.044	65.6	0.466	11.3	0.491	11.4
Coworker Mean								
Year 1974	845.095	=	0.321	1,563.4	0.196	479.1	0.483	516.2
Year 2006	754.655	=	0.114	793.9	0.462	970.8	0.424	508.8
Intrasector	690.784	=	0.321	793.9	0.196	970.8	0.483	508.8
Intersector	618.081	=	0.114	1,563.4	0.462	479.1	0.424	516.2
Coefficients of Variation*								
Year 1974	7.329	=	0.076	102,598.9	0.268	5,400.1	0.656	5,871.8
Year 2006	6.844	=	0.044	32,687.4	0.466	15,185.9	0.491	6,944.3
Intrasector	6.401	=	0.076	32,687.4	0.268	15,185.9	0.656	6,944.3
Intersector	7.186	=	0.044	102,598.9	0.466	5,400.1	0.491	5,871.8

Notes: *Aggregate coefficients of variation are calculated here as the square root of the sum of the products of sector weights and variances, all over the mean establishment size.

defined as $e^s = P(wsector = s)$, where $wsector = s$ is the condition that a worker is employed at an establishment in sector s. Notice that e^s and n^s are the shares presented in Figures 9 and 10, respectively.

Decomposition Results

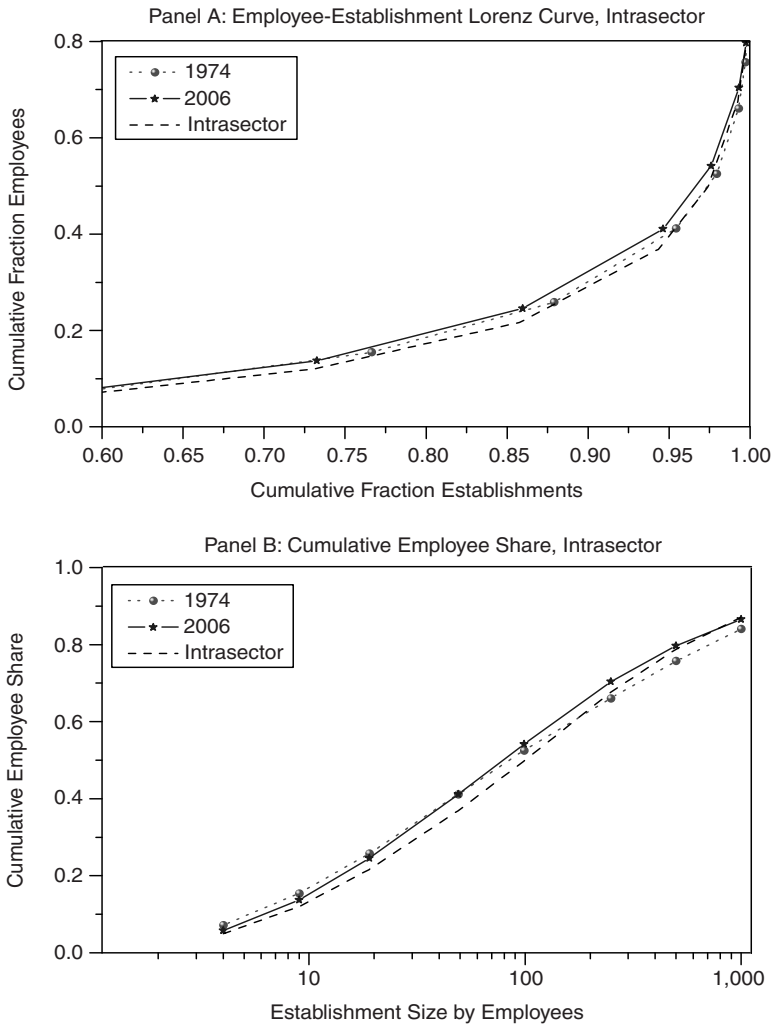
Table 2 presents the decomposition of trends in intra- and intersectoral changes. It shows how each statistic can be constructed as a weighted average of sectoral values. It also illustrates the computation of the counterfactual statistics used for the decomposition following the logic of equations (10) and (11). Considering only intrasector changes, the mean size of establishments would have increased 5 percent. Only the establishment mean of the manufacturing sector fell during this period, and its weight is relatively small. Keeping intrasector changes constant, the mean size would have dropped 12 percent. This is clearly because services, a sector with relatively small establishments in 1974, nearly doubled its share during this period.

Coworker mean results are substantially different. The main reason is that when labor shares are used instead of establishment shares, manufacturing is far more important than services. Consequently, when only intrasector changes are permitted, the drop in the coworker mean of manufacturing dominates the rise in services, and the coworker mean drops by 20 percent. Similarly, considering only intersector changes, the coworker mean size would have dropped 31 percent.[13] Finally, Table 2 presents the decomposition of the coefficient of variation of the establishment size distribution. The drop at the aggregate level is 7 percent. The decomposition shows that this drop is mainly due to intrasector changes. Keeping the share constant at 1974 levels, the drop would have been −14 percent; if one allows only changes in the share a fall of −2 percent is observed.

Figures 11 and 12 further resolve changes in the concentration of labor across establishments. Notice that these figures describe the distribution of workers across establishments, while the coefficient of variation mentioned earlier describes the distribution of establishments across establishment sizes. The results of this decomposition are different than those of the decomposition of the coefficient of variation. Allowing only intrasector changes, there would be a less equal distribution of labor across establishments in 2006 (see Figure 11). In contrast, intersector changes imply a greater shift toward a more even distribution than the one observed during this period.

[13] It is surprising in this case that with inter- or intrasector changes alone the coworker mean would have decreased more than when both changes occurred. This happens because the coworker mean size of services is higher than that of manufacturing in 2006, while the reverse is true in 1974. Thus, when the shares are allowed to change (not just the sectoral means), the aggregate coworker mean size increases.

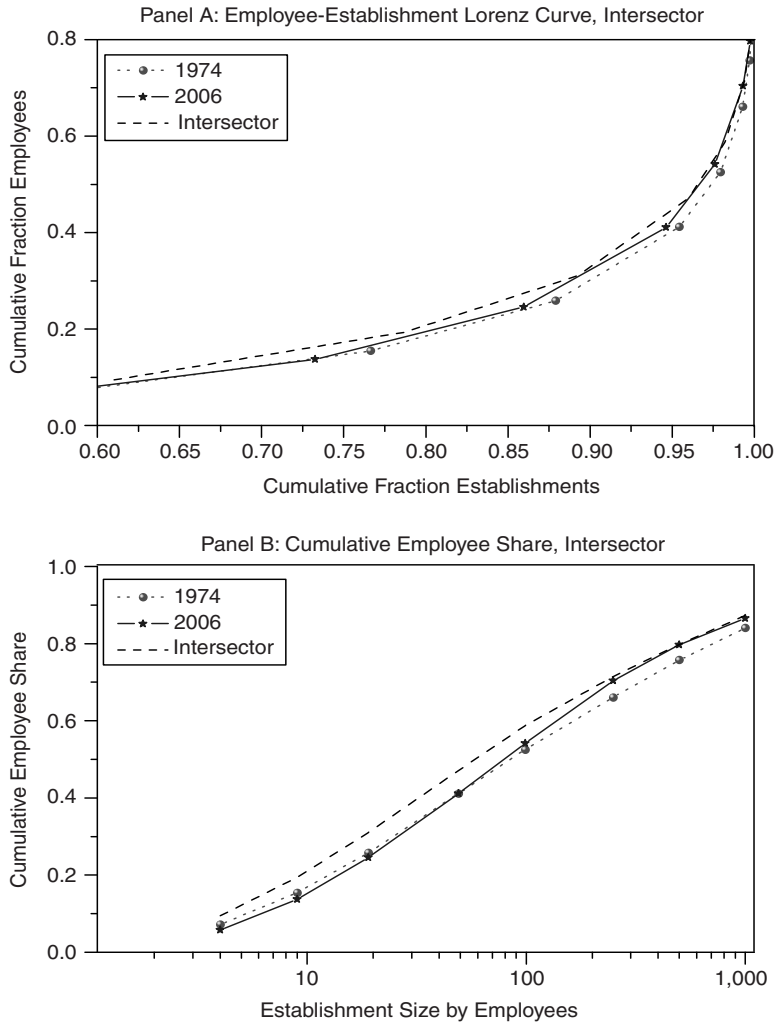
Figure 11 Intrasectorial Changes in the Establishment-Size Distribution, 1974–2006



4. FIRMS VERSUS ESTABLISHMENTS

Although the establishment is usually used as the production unit in models with heterogeneity in productivity, it is conceivable that the firm might also serve in that role. Because production units in these models vary in productivity or in their managers' ability, one could argue that they resemble establishments. However, since financial decisions are also made at the

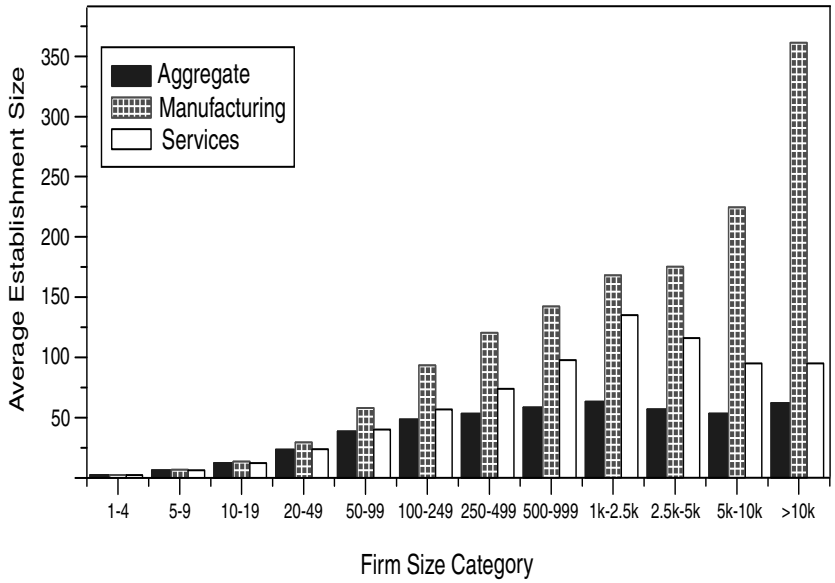
Figure 12 Intersectorial Changes in the Establishment-Size Distribution, 1974–2006



production unit level, it might also be argued that firm data is more appropriate. If it could be shown that the distribution of labor across firms tracks labor patterns across establishments, however, this distinction might be irrelevant.

It might be expected that small firms and small establishments, and large firms and large establishments, will see their labor distributions move together. Trivially, all small firms are composed entirely of small establishments, and all large establishments are constituent parts of large firms. If large firms contain

Figure 13 Establishment Size by Firm Size Group, 1991

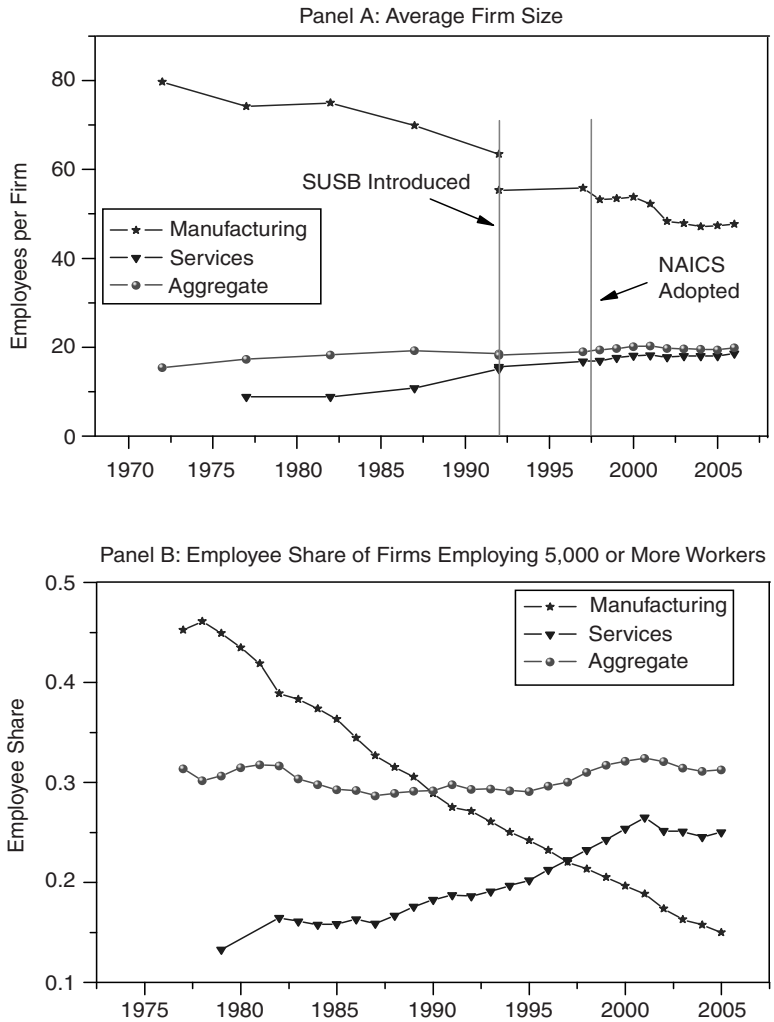


few small establishments, then the employment share of small establishments will correlate strongly with the employment share of small firms; the same will be true of large establishments and large firms. However, one may imagine a world in which large firms are mostly composed of many small establishments, and in this case movements in the distribution of labor across establishments might not be clearly reflected in movements of workers among firms. Consequently, it might be expected that co-movement in labor across establishments and across firms tends to be greater when large firms are composed of larger establishments.

Firm Data Sources

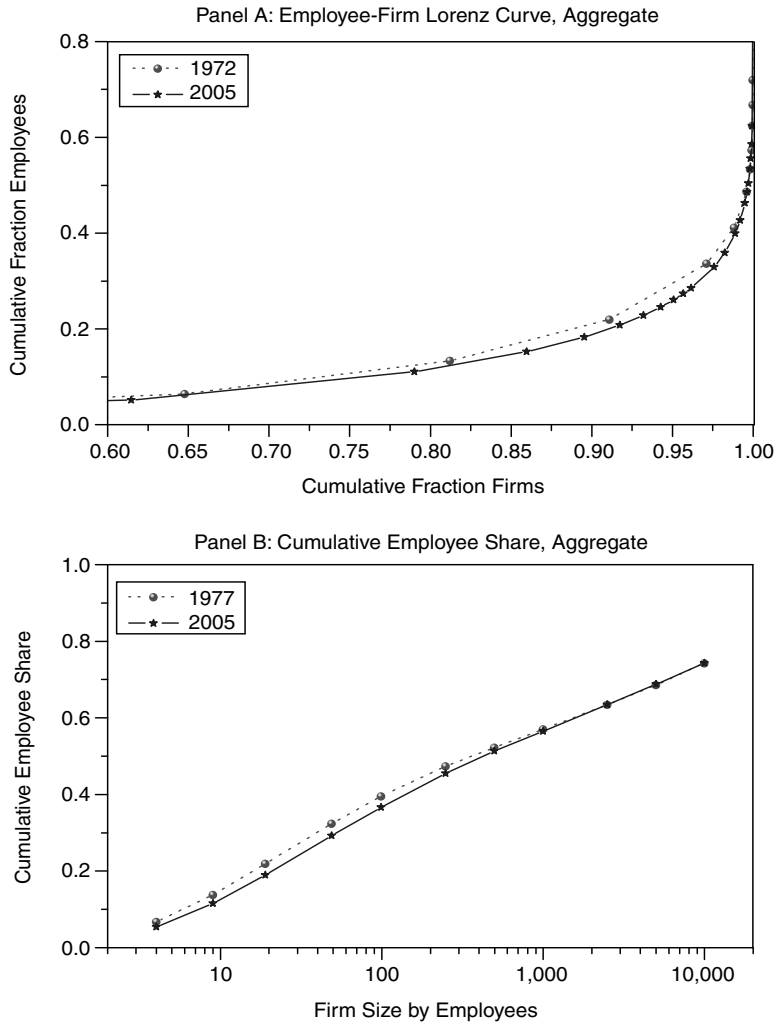
Firm data were obtained from three Census Bureau series: Enterprise Statistics, Statistics of U.S. Businesses (SUSB), and Business Dynamics Statistics (BDS). All series contain tallies of establishments and employees by firm size; Enterprise Statistics and SUSB also contain a count of firms in each firm size group. Enterprise Statistics was published consistently every five years from 1967 to 1992; SUSB was published in 1992 and annually after 1997. BDS was constructed retrospectively from several internal census databases and is available annually from 1977.

Figure 14 Firm Size and the Share of Large Firms in Total Employment



Whenever possible, BDS data are utilized. The publication is consistent in scope and methodology over the entire period of study. SUSB and especially Enterprise Statistics suffer from shifting definitions and sector coverage. These deviations, and the methods used in this article to mitigate their effects, are discussed in the Appendix.

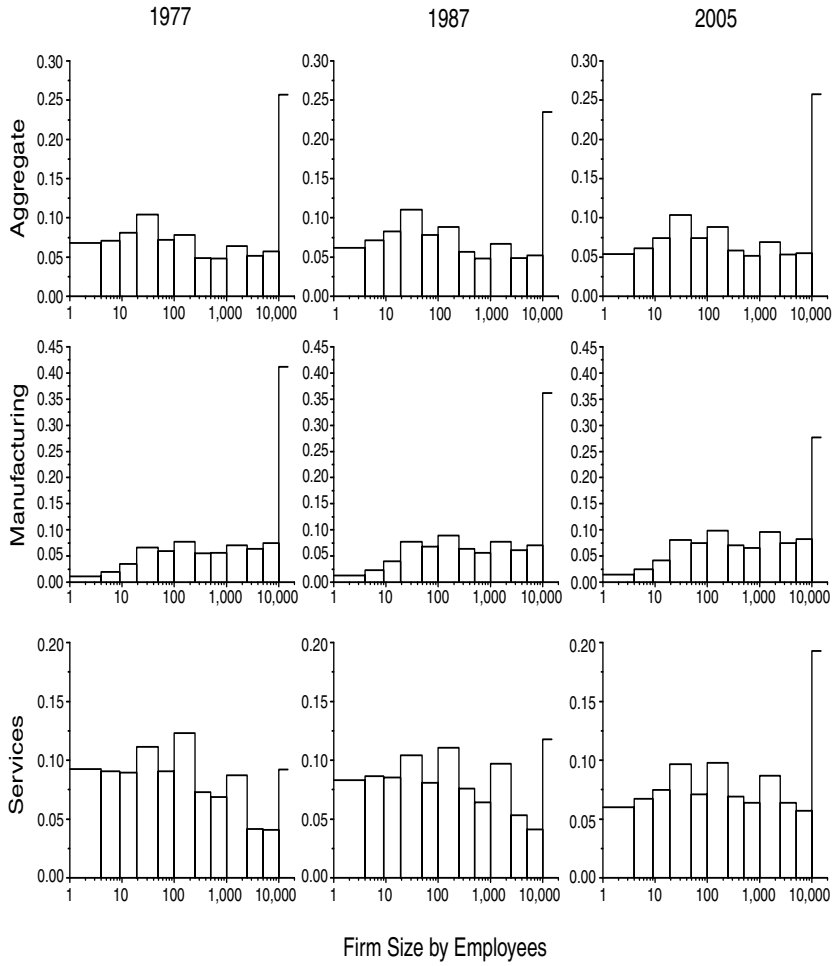
Figure 15 Firm-Size Distribution; Aggregate Economy, 1977–2006



Comparison Results

Figure 13 shows the average size of establishments for firms in 12 size categories in 1991; the data in this figure are typical for the sectors depicted and for the years 1979–2005. These data were obtained from BDS. Large firms, unlike small firms, do seem to be composed of larger establishments, and this is even more true in the manufacturing sector than in the rest of the economy. Movements in labor distribution should be similar across establishments

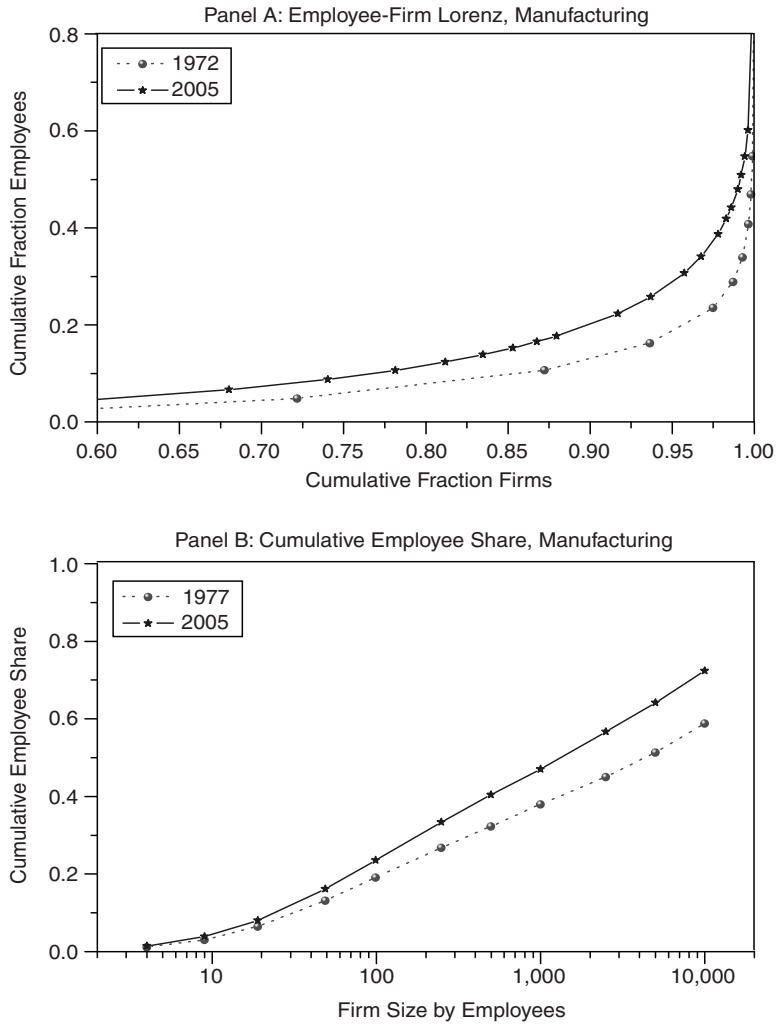
Figure 16 Histograms for the Distribution of Labor Across Firms



and firms, then, especially within the manufacturing sector. Indeed, evidence presented below generally confirms firm-establishment labor co-movement in these sectors, and to a degree in the aggregate economy, at least in the period under examination.

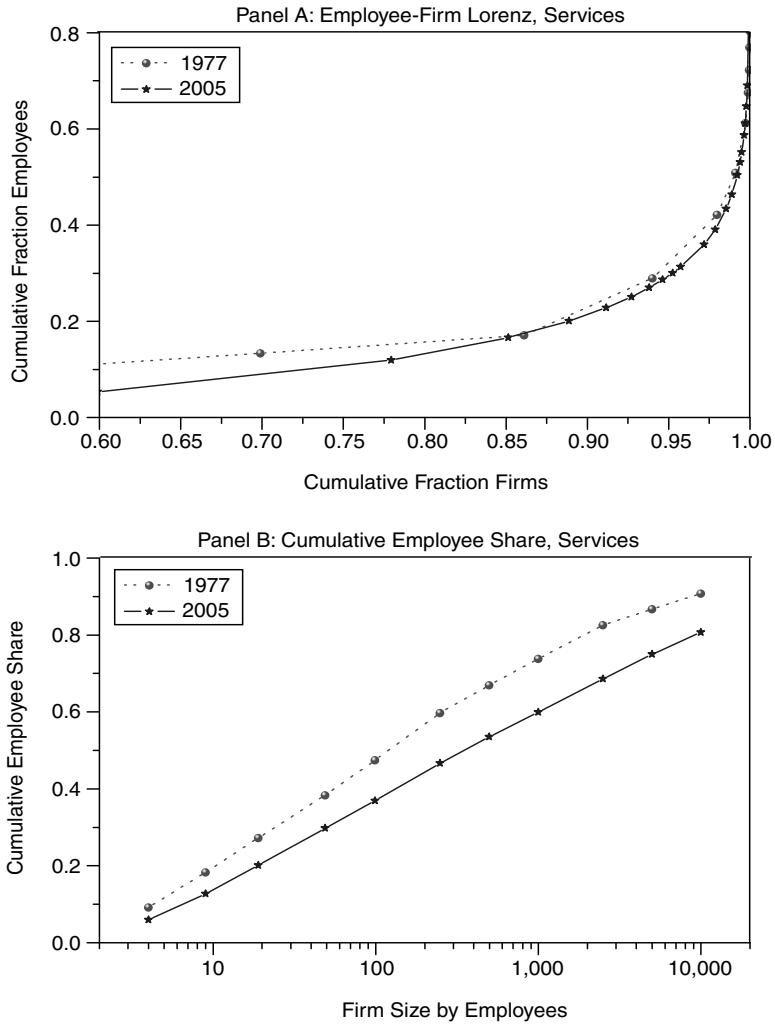
Figures 14 through 18 display firm data analogous to the establishment data. Data used in the creation of Lorenz curves (Panel A in Figures 15, 17, and 18) and mean firm size series (Figure 14, Panel A) were obtained through Enterprise Statistics and SUSB. Other firm figures (Panel B in Figures 14, 15, 17, and 18, as well as all of Figures 13 and 16) were derived from the BDS series.

Figure 17 Firm-Size Distribution; Manufacturing Sector, 1974–2006



It is clear that labor distribution movements across establishments track those in firms. Both the aggregate and the sectoral mean size series display the same patterns between the early 1970s and mid-2000s that are seen at the establishment level. Intrasector changes in the distribution of employment by firm size resemble those in establishment data: labor in the manufacturing sector became less concentrated (more clearly for firms than establishments), while service sector labor grew slightly more concentrated. Perhaps the only qualitative departure from establishment trends is a decrease in the

Figure 18 Firm-Size Distribution; Service Sector, 1977–2006



evenness of the aggregate labor distribution across firms that occurred between 1972 and 2005.

5. CONCLUSIONS

This article collects and analyzes publicly available data from the 1970s onward to obtain a set of statistics that can be used to calibrate and evaluate models with establishment heterogeneity. Recently, these models have

become widely used in economics to explain phenomena as important as economic development.

At the aggregate level, there is a minor shift of labor to mid-size establishments and away from the smallest and largest establishments. This change is partially explained by intrasector changes. The largest manufacturing establishments have consistently lost employee share since 1974, and manufacturing establishments smaller than 500 employees have uniformly seen their employee share increase. Trends in the distribution of labor across service sector establishments are complicated by inconsistencies in the definition of the service sector, but service establishments seem to have become larger since 1974, and the largest service establishments have grown at a disproportionately fast rate. Thus, the distribution of labor across service establishments has become less even, with most change occurring before 1997. Changes in the aggregate distributions of establishments and labor across establishments are also the result of changes in the share of sectors. Between 1974 and 2006 the worker share of manufacturing, a sector with large establishments and concentrated labor, decreased as the employment share of the services sector, characterized by smaller establishments, increased. In combination with movements in intrasector distributions, this trend explains observed changes in the aggregate distributions of establishments and labor across establishments.

Labor movements across firms should, hypothetically, resemble the movement of labor across establishments. This will be true to a greater degree when large firms contain fewer small establishments. This hypothesis is not contradicted by the data presented in this article.

APPENDIX

Data Sources

Enterprise Statistics

The Enterprise Statistics (ES) data set was first published in 1954; later publications came in 1958, 1963, 1967, and every five years after 1967 until the series was discontinued after 1992. The primary virtue of ES for this article is the provision of tables detailing quantities of firms, establishments, and employment; these values are provided for firms in different employment size groups similar to establishment size groups in CBP. These size groups are available for the aggregate economy as well as for sectors that generally replicate SIC definitions.

Unfortunately, ES's coverage and content changes significantly from publication to publication. The number of SIC sectors covered varies wildly;

using sector-level data we were able to homogenize the aggregate data, but the adjusted series lacks coverage of entire sectors (transportation and communication; finance and real estate; and most services). Moreover, the 1972 publication inflates its count of small firms by including certain non-employers; this can be corrected for the aggregate using a table found in that publication's appendix. The manufacturing sector from this year is still usable because there are no manufacturing firms in the small size group affected by the 1972 methodology, but the sector-level data for service firms must be set aside.

Adjustment of ES data to obtain a homogenous aggregate composition requires the subtraction of some sectors from each year's aggregate. This is a simple arithmetic task complicated in some cases by the lack of subsector data: The Census Bureau occasionally withholds employment information for certain firm size groups if its publication might result in the disclosure of private information. These missing values are estimated by multiplying the number of firms in the size group with the missing data by the mean number of employees per firm for the size group at the aggregate level. An example adjustment is displayed in Table 3. There, the original employee count for each aggregate firm size group was reduced by the deduction of employees in public warehousing, travel agencies, and dental laboratories—three small sectors not present in the ES aggregates in all years. Values in bold were missing from the original publication and estimated using the procedure previously described. Similar exercises were also carried for firm and establishment series and in all ES years.

The composition of the services sector also varied from publication to publication. Unfortunately, homogenization was not a feasible solution: very few firms would remain in an intertemporally consistent services sector. Consequently, the service sector is presented for each ES year unaltered with the caveat that it is inconsistent.

Statistics of United States Businesses

Statistics of U.S. Businesses (SUSB) replaced ES in 1992; it was published in 1992, and annually from 1997 onward. Although SUSB provides data similar to those found in ES, there are several important differences. First, SUSB covers many sectors not covered by ES. This leaves aggregate data somewhat incomparable across the two publication series, especially after this article's sectoral homogenization of aggregate ES data. Second, SUSB uses enterprise size groups rather than firm size groups. In ES these terms were interchangeable and each enterprise was assigned a single industry code; in SUSB an enterprise is composed of many firms, each of which represents the enterprise's production in a given industry. With this convention, it is possible to find a 5,000–9,999 employee size group containing three firms employing 2,000 workers between them. This data is not well-suited for the creation of Lorenz curves because it does not permit the sorting of firms by size.

Table 3 Adjustment to ES Sectoral Composition; Example

Firm Size Group	Original Total Total Employees	(Subtracted) Public Warehousing (42A)	(Subtracted) Travel Agencies (47)	(Subtracted) Dental Laboratories (80)	Final Figure Adjusted Total Employees
0	0	0	0	0	0
1–4	2,938,355	5,235	8,898	6,355	2,917,867
5–9	3,209,609	8,471	10,447	5,107	3,185,584
10–19	3,945,190	14,670	7,869	5,284	3,917,367
20–49	5,372,937	27,508	5,997	6,072	5,333,360
50–99	3,446,571	13,739	3,100	1,670	3,428,062
100–249	3,459,628	14,281	1,967	1,526	3,441,854
250–499	2,126,488	**5,833**	2,100	**686**	2,117,869
500–999	1,837,286	**688**	**688**	0	1,835,910
1,000–2,499	2,330,673	**4,618**	**3,079**	**1,539**	2,321,437
2,500–4,999	1,981,793	0	0	0	1,981,793
5,000–9,999	2,376,041	0	0	0	2,376,041
= 10,000	12,786,233	0	0	0	12,786,233
Total	45,810,804	94,464	44,888	27,744	45,643,708
Column Error	0	−579	743	−496	331

Values in bold were missing from the original publication and are estimated using the procedure described in the text of this article.

Table 4 Services Sector Assembled from NAICS

NAICS Number	NAICS Service Sector Component
54	Professional, scientific, and technical services
56	Administrative and support and waste management and remediation services
61	Educational services
62	Health care and social assistance
71	Arts, entertainment, and recreation
72	Accommodation and food services
81	Other services (except public administration)

Moreover, it prevents any adjustment of the SUSB aggregate by the subtraction of sector data, because too many firms would be dropped. For example, if the construction and mining sectors are subtracted from the aggregate, and a single enterprise has constituent firms in each sector, then two firms will be removed from the aggregate despite the fact that the enterprise is represented in the aggregate by a single firm. Consequently, sectoral and aggregate data are only marginally comparable between the two series.

The utility of SUSB is further reduced by the switch to the NAICS classification system from the SIC system after 1997; it is difficult to compare sectors between systems, and, as with CBP, it was necessary to construct a composite service sector from several NAICS subsectors (see Table 4). Because of the SUSB definition of a firm, the number of service firms in large size groups is probably overstated in NAICS.

Business Dynamics Statistics

BDS is consistent in methodology and coverage; derived from a number of internal USCB databases, it has annual data on employment for firm size groups reaching back to 1977. For the purposes of this article, BDS has one major shortcoming: For each firm size group, only data on establishments and employment are provided. When firm quantities are required for a calculation, ES and SUSB are used.

Because the series was assembled from microdata retrospectively, BDS industry classifications are internally comparable for all years. These classifications are based on the SIC system, and so the comparability of BDS sector data with CBP and SUSB sector series from 1998 on is somewhat compromised.

Computing Establishment and Coworker Means and Probabilities

We compute the expected establishment mean for a size group by dividing the total number of workers in a size group ($workers_i$) by the total number of establishments in the size group ($establishments_i$):

$$E(esize \mid egroup = i) = \frac{workers_i}{establishments_i}. \tag{12}$$

Obtaining the expected coworker mean for a size group is more involved and the next subsection is devoted to this effort. Meanwhile, the probabilities $P(egroup = i)$ and $P(wgroup = i)$ are obtained by dividing the establishments or workers (respectively) in i by the total number of establishments or workers over all size groups j:

$$P(egroup = i) = \frac{establishments_i}{\sum_j establishments_j}, \text{ and} \tag{13}$$

$$P(wgroup = i) = \frac{workers_i}{\sum_j workers_j}. \tag{14}$$

Probabilities $P(egroup \leq i)$ and $P(wgroup \leq i)$ are calculated in a similar manner by summing the probabilities for each size group j less than or equal to i:

$$P(egroup = i) = \frac{\sum_1^i establishments_j}{\sum_j establishments_j}, \text{ and} \tag{15}$$

$$P(wgroup = i) = \frac{\sum_1^i workers_j}{\sum_j workers_j}. \tag{16}$$

Computing the Size-Group Coworker Mean

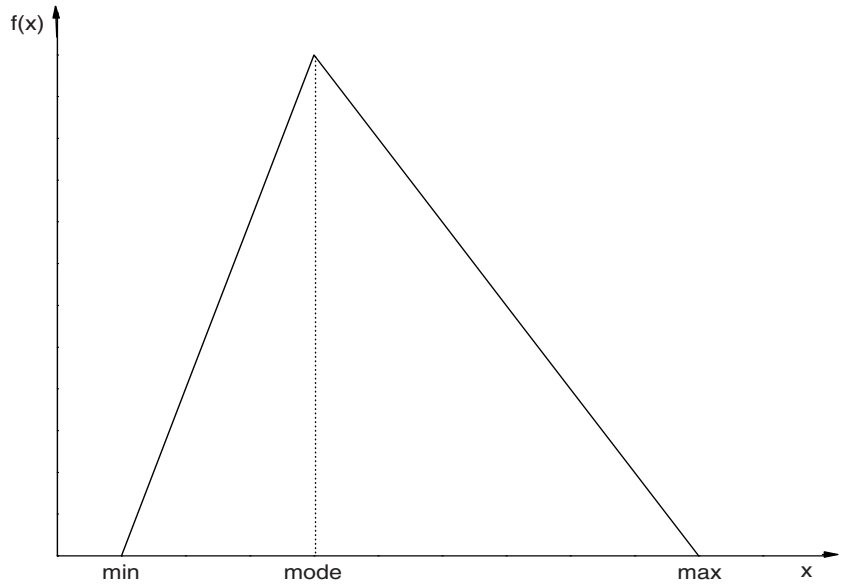
For each size group i, the available information is

- the minimum and maximum size in the group, min_i and max_i, respectively;
- the total number of workers, $workers_i$; and
- the total number of establishments, $establishments_i$.

With this information it is simple to compute the mean size of the group,

$$E(esize \mid egroup = i) = \frac{workers_i}{establishments_i}. \tag{17}$$

Figure 19 Triangular Distribution; Example



Unfortunately, it is not possible to compute the coworker mean of this group. Davis and Haltiwanger (1989) show that the coworker mean can also be written as

$$E(wsize \mid wgroup = i) = E(esize \mid egroup = i) + \frac{V(esize \mid egroup = i)}{E(esize \mid egroup = i)}, \tag{18}$$

where $V(esize \mid egroup = i)$ is the variance of the establishment size for the size group i. Equation (18) indicates that once $E(esize \mid egroup = i)$ is known, only an estimate of $V(esize \mid egroup = i)$ is needed to obtain an estimate of $E(wsize \mid wgroup = i)$. With a distributional assumption for the distribution of establishments inside each size group, this statistic can be recovered. A useful assumption is that this distribution is triangular. This distribution has three parameters: the lower bound, min; the upper bound, max; and the mode, $mode$. The probability density function increases linearly from min to $mode$ and decreases linearly from $mode$ to max (see Figure 19 for an example). With this assumption, the mean size can be written as

$$E(esize \mid egroup = i) = \frac{min_i + max_i + mode_i}{3}. \tag{19}$$

Since $E(esize \mid egroup = i)$, min_i, and max_i are available, one can use the equation above to solve for $mode_i$. Then, it is simple to compute the variance using the formula for the triangular distribution,

$$V(esize \mid egroup = i) =$$

$$\frac{min_i^2 + max_i^2 + mode_i^2 - min_i * max_i - min_i * mode_i - max_i * mode_i}{18}. \tag{20}$$

Finally, equation (18) can be used to compute the coworker mean of size group i.

REFERENCES

Alfaro, Laura, Andrew Charlton, and Fabio Kanczuk. 2008. "Plant-Size Distribution and Cross-Country Income Differences." Working Paper 14060. Cambridge, Mass.: National Bureau of Economic Research.

Amaral, Pedro S., and Erwan Quintin. 2007. "Limited Enforcement, Financial Intermediation, and Economic Development: A Quantitative Assessment." Manuscript, Federal Reserve Bank of Dallas.

Axtell, Robert L. 2001. "Zipf Distribution of U.S. Firm Sizes." *Science* 293 (September): 1,818–20.

Banerjee, Abhijit V., and Esther Duflo. 2005. "Growth Theory through the Lens of Development Economics." In *Handbook of Economic Growth*, edited by Philippe Aghion and Steven Durlauf. Amsterdam: Elsevier, 473–552.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta. 2008. "Cross-Country Differences in Productivity: The Role of Allocative Efficiency." Manuscript, University of Maryland.

Buera, Francisco J., and Joseph P. Kaboski. 2008. "Scale and the Origins of Structural Change." Federal Reserve Bank of Chicago Working Paper 2008-06.

Buera, Francisco J., and Yongseok Shin. 2008. "Financial Frictions and the Persistence of History: A Quantitative Evaluation." Mimeo, Northwestern University.

Caselli, Francesco, and Nicola Gennaioli. 2003. "Dynastic Management." Working Paper 9442. Cambridge, Mass.: National Bureau of Economic Research.

Castro, Rui, Gian Luca Clementi, and Glenn McDonald. 2009. "Legal Institutions, Sectoral Heterogeneity, and Economic Development." *Review of Economic Studies* 76 (April): 529–61.

Davis, Steven J., and John Haltiwanger. 1989. "The Distribution of Employees by Establishment Size: Patterns of Change in the United States, 1962–1985." Manuscript, University of Chicago and University of Maryland.

Davis, Steven J., and John Haltiwanger. 1990. "Size Distribution Statistics from County Business Patterns Data." Manuscript, University of Chicago.

Davis, Steven J., John Haltiwanger, and Scott Schuh. 1996. *Job Creation and Destruction*. Cambridge, Mass.: The MIT Press.

Gibrat, R. 1931. *Les Inégalités Économiques; Applications: Aux Inégalités des Richesses, à la Concentration des Entreprises, Aux Populations des Villes, Aux Statistiques des Familles, etc., d'une Loi Nouvelle, La Loi de l'Effet Proportionnel*. Paris: Librarie du Recueil Sirey.

Greenwood, Jeremy, Juan M. Sánchez, and Cheng Wang. 2008. "Financing Development: The Role of Information Costs." Federal Reserve Bank of Richmond Working Paper 08-08.

Guner, Nezih, Gustavo Ventura, and Xu Yi. 2008. "Macroeconomic Implications of Size-Dependent Policies." *Review of Economic Dynamics* 11 (October): 721–44.

Hopenhayn, Hugo, and Richard Rogerson. 1993. "Job Turnover and Policy Evaluation: A General Equilibrium Analysis." *Journal of Political Economy* 101 (October): 915–38.

Hsieh, Chang-Tai, and Peter J. Klenow. 2007. "Misallocation and Manufacturing TFP in China and India." Working Paper 13290. Cambridge, Mass.: National Bureau of Economic Research (August).

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

Restuccia, Diego, and Richard Rogerson. 2008. "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments." *Review of Economic Dynamics* 11: 707–20.

Ruggles, Steven, Matthew Sobek, Trent Alexander, Catherine A. Fitch, Ronald Goeken, Patricia Kelly Hall, Miriam King, and Chad Ronnander. 2009. *Integrated Public Use Microdata Series: Version 4.0*. Minneapolis: Minnesota Population Center. http://usa.ipums.org/usa/.

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