

Process and Progress in Risk Management

Based on a speech presented by President Santomero at the BAI - Treasury, Investment, ALM, and Risk Management Conference, New York, New York, October 28, 2002

BY ANTHONY M. SANTOMERO

The banking industry and its regulators recognize that they share a vested interest in enhancing the industry's risk-management strategies. In his first *Business Review* message of 2003, President Santomero discusses three points: risk management as its own distinct discipline; the financial industry's work to improve risk-management techniques and regulators' increased commitment to risk-focused examinations; and the need to improve risk-management systems even further.

Risk management is a topic close to my heart and my career as an academic, consultant, and now policy-maker. In this message, I'd like to reflect on the advances that the financial services industry has made over the past decade in risk-management practices, as well as the challenges we still face.

But before I recount that story, I'd like to point out a subtle but substantive change that has occurred in this area since I began my study of risk management in the financial services sector. Now, both industry and regulators recognize that we share a vision and a vested interest in enhancing the industry's risk-management strategies. As a result, a joint effort is taking place to raise risk-management standards for the entire industry. As I will note later, this is both a major advance and a substantial improvement for the industry and its regulation.

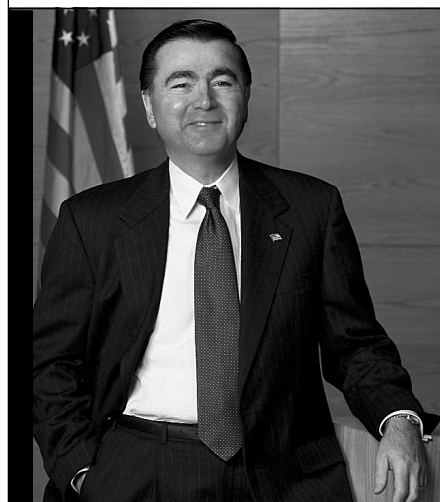
I will emphasize three main points here: One, over our recent past, the practice of risk management has evolved as its own separate and distinct discipline. Two, as this evolution has taken place, the financial industry has worked to improve risk-management techniques, and regulators have indicated an increased commitment to risk-focused examinations. Three, as the industry evolves, risk-management systems will need to improve even further and become a greater part of firms' decision-making process. In fact, an emphasis on risk-management capability will be an increasing part of the supervisory process.

I will also look at what is needed to be successful in the effort to raise risk-management standards. But before we embark on the future of risk management, let's look at where we have been.

PROGRESS SINCE 1996

In 1996, I helped write a best-practices study of risk management in the financial services industry, which was financed by the Sloan Foundation and conducted under the auspices of the Wharton Financial Institutions Center. That study included discussion of what institutions were actually doing and what had yet to be done. At the time, the industry had finally reconciled itself to the fact that banking is an inherently risky business. Institutions had gone from trying to avoid risk to developing techniques to manage different types of risk, with credit risk and market risk taking center stage. In each case, the risk was being identified and quantified.

For the most part, credit risk concerns centered on newly discovered concentrations. And the growing threat of volatility here and abroad led to new developments to measure and limit trading risk. In fact, it could be argued



Anthony M. Santomero, President,
Federal Reserve Bank of Philadelphia

that industry interest in risk management as a profession developed almost by accident. While efforts were already under way to formalize the activity, the practice really received industry buy-in only when everyone realized the mistakes of the past: over-concentrations in credit and excessive trading risk levels.

Back then, lip service was given to operational risk too, but little analytical work had been done on the subject. Less quantifiable risks, such as reputation, regulatory, or strategic risk, were managed less formally or simply ignored. CEOs had these issues on their radar screens, but virtually no substantive analysis existed.

Risk aggregation was seen as a major issue for the industry; there was a clearly perceived need to install an organizational structure to oversee risk management at the firm level. But this was a new and controversial concept. The title “risk czar” was being floated, but not everyone was sure where this function fit into the organizational structure. In fact, different organizations had different solutions to the management of firm-level risk.

Since then, much has been accomplished. Risk management has become an increasingly prevalent — and accepted — industry discipline. Firms have rushed to develop systems and install processes to manage the risks that are an inevitable part of the financial landscape. Firms now understand that risks of various types, embedded throughout their portfolios, must be managed both carefully and rigorously. Collectively, they impose an aggregate level of risk that can threaten the very solvency of the firm.

Now, risk assessment is a standard part of every deal, every strategic discussion, and every financial review. Firms recognize all risks are ultimately related and strive to focus their efforts on total enterprise risk.

Moreover, a clear role for the firm-level risk manager has emerged. We have come a long way. Risk-management systems have been developed and implemented as firms have forged a new risk-management culture. And the result, at least in part, has been a decade of high earnings and overall stability in the banking industry.

In fact, the last decade can be distinguished by what did not happen, rather than what did. Volatile markets did not lead to the spectacular losses of past cycles. The Asian crisis left trading firms relatively unscathed. The technology industry bubble brought down no major financial institutions and resulted in manageable credit losses. And profits, capitalization, and solvency ratios improved throughout the industry — despite the recent recession and a series of extraordinary domestic and international events.

Interestingly, as the financial system became more complex, regulators encouraged more private-sector innovation, in the belief that markets are quite efficient at sorting out their own best practices.

In short, the past decade has proven the increased ability of the industry to manage risk and has demonstrated the benefits of substantially improved risk-management capacity.

This has not been lost on regulators, who themselves have embraced the new discipline of risk management. The results of regulators’ efforts are also evident. By the mid-1990s, regulators had made the very practical move to risk-based examinations rather than just looking at point-in-time balance sheets and financial ratios.

Interestingly, as the financial system became more complex, regulators encouraged more private-sector innovation, in the belief that markets are quite efficient at sorting out their own best practices. The Fed’s own philosophy is that flexible yet watchful supervision, complemented by market discipline, is the best approach to ensure a safe and stable financial system.

Yet, more needed to be done on the regulatory front and still does. As the industry’s approach to risk management became more sophisticated, so did its systems and business practices. Early in this process regulation had to play catchup. For example, capital arbitrage became a common practice, and this led regulators to reassess the very foundation of capital regulation framework instituted in the late 1980s.

The once-innovative regulatory regime established with the Basel

Accord concentrated exclusively on credit risk but had only a handful of risk categories and totally ignored both trading and interest-rate risk, as well as correlations across risk categories. Changes in the intervening decade attempted to retrofit the regulations by adding trading risk and interest-risk considerations to the standards. But the results were never fully satisfactory. The outcome was both regulatory arbitrage and avoidance. In time, it became clear that Basel I had become obsolete. Regulation had fallen behind, and it was time for something new.

BASEL II

The Basel Accord's shortcomings prompted the Bank for International Settlements to revamp and update international capital regulation in the living document known as Basel II. In general terms, the goal of this effort is to update the earlier model of risk-related capital regulation in light of current market instruments and modern financial techniques.

Nonetheless, Basel II should be seen as quite distinct from predecessor regulations in at least one important respect: It is an effort to engage both the industry and the regulators in using advanced risk-management techniques. This shows through in two important ways. First, whereas Basel I focused only on regulatory capital adequacy, Basel II gives equal consideration to minimum capital ratios, supervisory review, and market discipline. Second, substantial effort has been made to incorporate the risk management practices that firms actually use into the process and to increase the risk sensitivity of the minimum capital requirements.

While a considerable advance, Basel II has its critics. One common complaint is that the current proposal is too complex. Is it? Yes. However, its complexity reflects the underlying complexity of risk and risk management in modern banking institutions. Is it doable? I believe so. In fact, by proposing the use of a bank's own risk-management system in the advanced internal risk-based (or IRB) approach, Basel II engages the banking industry's risk-management community in determining appropriate bank risk levels and regulatory capital ratios.

In its present form, the advanced IRB approach is designed to employ the advanced risk-management systems that banks have in place for day-to-day operations in the determination of capital adequacy. Regulators will need to certify that a bank's systems are

up to the task, and in many cases, this may require substantial improvement of the systems. However, the approach is one where banking firms and regulators will need to work together to improve the existing best practices in the industry. Just as Basel I became obsolete, Basel II will not be the final word on risk-management regulation. But it is a step forward in that its structure works

Basel II sets the right incentives for the industry to continue to seek advances in risk management and for regulators to continue to improve their skills in assessing the adequacy of risk-management systems in use.

to encourage the industry and regulators toward better risk-management practices. In this way, the industry itself can lead in the evolution of risk management — as ultimately it should.

Some critics of Basel II feel that it crosses the line and makes bank regulators into bank managers. This is not our intention, and we recognize that this is a potential danger that needs to be avoided. In addition, some bankers question whether regulators have the expertise to properly assess banks' systems. This is a legitimate concern, and it highlights the necessity of regulators everywhere to intensify their efforts in the areas of appropriate staff development and training.

While these are potential problems, we cannot gain the benefits of incorporating banks' internal risk-management practices into regulatory capital unless regulators conduct appropriate analysis to ensure the adequacy of industry practice. Without sufficient supervisory review, it would simply be imprudent to defer to internal ratings for appropriate oversight of industry risk levels.

The banking industry is littered with firms that confidently talked the talk of safety and soundness but fell flat when it came to walking the walk. As regulators, we need the assurance that risk-management systems are, in fact, advanced in both theory and practice. This assurance can come only from supervisors' gaining first-hand knowledge of bank operations.

To illustrate this point, let me tell you about an experience I had during my days as a risk-management consultant. I visited a major financial institution in New York to assess its approach to trading risk. The CEO assured me that the bank had a highly sophisticated VAR risk-management system already in place. The CFO said they had just implemented it. The head of trading said they were about to implement it. And the traders — well, they'd never heard of it.

So, in this case, senior management thought that it had an advanced trading risk-management system in place and everything was under control. But the facts were that the organization had its traders taking million-dollar positions with few controls in place. You can see why regulators might get nervous.

THE FUTURE

As to the future of risk management, the most important thing to keep in mind is that Basel II sets the stage for a joint effort and further advances in the science and art of risk

management. Basel II sets the right incentives for the industry to continue to seek advances in risk management and for regulators to continue to improve their skills in assessing the adequacy of risk-management systems in use. Together, industry leaders and regulators can work to raise the standards of risk management for the industry.

The next stage in this process is at hand: the third round of quantitative impact studies, the so-called QIS 3, was launched last October. This study allows banks to assess how Basel II will affect their particular institution.

On the international level, a regulatory Accord Implementation Group has been formed to assure the industry that common approaches and a level playing field will emerge from the implementation process scheduled for the end of 2006 for internationally active banks.

The special and unique element of Basel II is that it allows for an internal risk-based approach. The best banks will be able to step forward and define best practices for the industry. While earlier methods dictated across-the-board regulation, now regulators are looking to the industry for valid approaches and new insights. Perhaps under Basel II, bankers and regulators will build the true relationship of trust and understanding that did not emerge under Basel I.

As a central bank responsible for the financial integrity of the financial system, the Federal Reserve sees the development of adequate risk-management systems as an important part of its balanced approach to bank supervision. As such, evaluating a bank's ability to establish an appropriate risk-management regime in the bank's culture has become a more important part of the bank regulation and supervision process.

The challenges still facing

bankers, regulators, and risk managers are well known. Everyone involved in risk management probably has his or her own list of projects that warrant industry attention. This is just part of the evolution of risk management.

Let me offer my list, for what it is worth. On credit risk, while techniques have improved, much work still needs to be done on consistency, transparency of process, and the timeliness of review.

We need to address the pro-cyclicality of risk in any risk-based capital allocation system.

On the commercial loan side, data on actual outcomes are still too scarce. On the retail side, many of the risk models are largely proprietary and of unknown reliability. The recent controversy surrounding the regulators' approach to retail risk quantification speaks more to the lack of a consensus on a standard approach to retail risk management than anything else.

On the market risk side, many questions still require ongoing investigation and continual monitoring. The robustness of the models and systems continues to be questioned. Market valuations of complex instruments are subject to debate — perhaps now more than ever. And the estimated correlations across markets seem to change too frequently to provide a useful guide for risk-management purposes.

On operational risk, we have even less knowledge and capability, even though contingency issues seem to loom larger now than ever before. Basel II has included operational risk in pillar one but permits an advanced management approach to the setting of an

appropriate capital level. Yet, little work has been done on measuring operational risk systematically, and insufficient public data exist to test the validity of different approaches.

However, perhaps the greatest challenge is the issue of appropriate risk aggregation. Whether it is the correlation of risks within product lines or across them, this is one area of significant disagreement. This was an open issue some years ago, and it is still the subject of much discussion and debate.

As we know, risk aggregation presents a fundamental problem. What is the correlation across different credit exposures? How can we aggregate different types of risk to measure the firm's total exposure? What is the correlation across different types of risk? How can we add up the risks associated with September 11, WorldCom, Argentina, and retail loan losses? Quantifying divergent risks and reaching some logical conclusion have proven to be a daunting task. We don't have all the answers yet, which is why we had better keep working.

But risk aggregation isn't the only open issue. We also need to figure out how to allocate capital within the firm to create incentive schemes that foster appropriate risk attitudes. And we need to address the pro-cyclicality of risk in any risk-based capital allocation system. This last issue remains a challenge. Exactly how stable should capital allocation algorithms be over a business cycle? And does the answer to this question differ at the firm level and at the regulatory level? Another open issue is how organizations should be structured to reflect risk-management priorities.

These are complex issues, and they raise questions we are still trying to answer. All of this suggests the status quo will not be good enough for tomorrow — indeed, it is probably not good enough for today.

Basel II, while a significant improvement, is just a step in an industry-wide movement toward better and more effective risk management. This is where the joint effort of industry leaders and regulators is invaluable. We must work together to make our best practices even better. It will take a lot of innovation and leadership from the industry. It will also take a lot of flexibility and direction from regulators.


Basel II sets a deadline of 2006 for implementation of adequate risk-management systems. Meeting that deadline will require the same effort and speed of scientific advance that we have seen thus far. And it is imperative that risk-management systems improve and become an even greater part of bank management's decision process. Indeed, the industry has already become more mindful of the new regulatory guidelines — guidelines that hold the industry and its systems to a

higher standard. But improvement in risk-management practices is not just imperative because of regulatory mandate; it is a necessary component of good banking in a world of increasing complexity and evolution of the financial services industry.

CONCLUSION

The financial sector has come a long way in its risk-management efforts. From the early days of simple ratios or simply risk avoidance, risk management has evolved into a complex, dynamic discipline of its own.

Basel II offers an unusual opportunity for banking issues to be resolved by those who will live with the result on a daily basis — the bankers. It makes sense, and I believe it will be very effective. Bankers know what is best for their banks. They will have the principal responsibility for setting their own course.

But there is more at stake here than the profitability and health of any single institution. The integrity and stability of the financial system is critical to the health of our economy. So banks must be prepared to defend their own assessments and procedures to their regulators and the market. If a bank is using the internal ratings-based approach, it should be prepared to provide concrete evidence and support for its systems. Regulators will expect it. I believe banks have the capability to successfully innovate and restructure to meet the requirements of this new, more rigorous environment. Banks and regulators should and will continue to work together to ensure risk-management processes are sufficiently robust and ultimately effective. We can take pride in the fact that we have already done so much. Yet, we have much work ahead of us. It will not be easy, and it will not be completed overnight. 

A Confluence of Events?

Explaining Fluctuations in Local Employment

BY GERALD A. CARLINO

L

ocal economies are subject to different types of shocks. Usually, the fortunes of local economies depend on a confluence of national, sectoral, and local shocks.

That raises a question: Does one type of shock systematically buffet local economies more than another? The answer has important implications for both academic researchers and policymakers. Jerry Carlino examines the evidence to see which type of shock most likely explains fluctuations in local employment.

In the late 1980s residents of Los Angeles saw an ominous cloud forming over the city. The local housing market, which for the previous several years had been red hot, began to falter. Soon, Angelinos would have to deal with the unthinkable — declining housing prices and a generally worsening regional economy.

The loss of the area's economic vigor had several sources. For one, Congress had been rolling back some of the increased defense spending it had legislated earlier in the decade, in response to the public outcry over large and growing federal budget deficits.



Jerry Carlino is a senior economic advisor and economist in the Research Department of the Philadelphia Fed.

Because the defense sector has a large presence in southern California, the cutbacks meant widespread job losses and decreased disposable income.¹

Problems in the banking sector added insult to injury. Loan defaults by Mexico and other developing countries combined with significant problems from the savings and loan crisis led to a noticeable contraction of employment and output in the financial services industry. Finally, the national economy went into a recession in 1991 causing the Los Angeles economy to stumble further.

¹ A sector consists of a group of industries whose firms produce a similar good or service. Nonagricultural employment is grouped into one of eight broad employment categories: government; mining; construction; transportation and public utilities; manufacturing; wholesale and retail trade; finance, insurance, and real estate; and nonfinancial services.

Throw in an earthquake or two, and the southern California economy was in serious trouble.

While the particulars of the preceding example are unique, virtually all local economies face the general experience much of the time. That is, local economic fortunes depend on a confluence of national events (for example, changes in interest rates), sectoral events (for example, changes in the defense and financial services industries), and local events (for example, earthquakes).

Some events, of course, are more important than others, depending on the time and place. But a question does arise: Does one type of shock, or disturbance, systematically buffet local economies more than another type? For example, do national events affect local economies more than sectoral ones do?

The answer has important and interesting implications, both for academic researchers who attempt to better understand the nature and sources of business cycles and for policymakers who wish to diminish the resulting swings in employment and output. For example, if national shocks were mainly the culprit, perhaps a monetary or fiscal policy action would be most helpful. If, however, disturbances to specific sectors were the main driver of fluctuations in activity, perhaps a policy that helped workers move from an economically troubled sector to a healthier one would be the better choice.

TYPES OF ECONOMIC SHOCKS

The phrase “economic shock” represents economists’ shorthand for a

factor or force that causes unexpected changes in economic variables. When studying changes in local employment, economists usually discuss three types of shocks. The first is a national, or aggregate, disturbance, such as a monetary or fiscal policy action, that typically affects all industries. Although the shock is national in origin, its impact will be felt locally and will cause fluctuations in local employment, albeit to different degrees in each locality.

In a 1998 study, Robert DeFina and I found that shocks to interest rates induced by changes in monetary policy affected different regions in quantitatively distinct ways, primarily because of differences in industry mix across regions. Personal income in the Great Lakes region, for example, showed the largest response to an unexpected increase of 1 percentage point in the federal funds rate: It dropped almost 50 percent more than did income at the national level. Personal income in the Rocky Mountain and Southwest regions, by contrast, responded only half as much as national income.

A second type of shock is one that affects a specific industry, such as a change in defense spending, the imposition of a tariff on particular imported goods, or a strike by workers in a particular sector, such as the automobile industry. Shocks that affect a specific sector of the economy will also have differing local effects because of variations in the concentration of industries. Mark Hooker and Michael Knetter found that changes in national military procurement spending have a modest impact on employment in most states but a sizable impact on those states that depend heavily on the military. Hooker and Knetter also found that changes in military spending have an asymmetric impact on income and employment: Large cutbacks in military spending have proportionately greater effects than do large awards.

As another example, the recently imposed tariffs on imported steel will have a larger effect in steel-producing states such as Pennsylvania, Ohio, and West Virginia, than in other locales. The 1996 strike at a General Motors brake plant in Dayton, Ohio, which crippled General Motors' North American operations, adversely affected

national versus sectoral shocks.³ The majority of this work has used quarterly or annual data and data that cover the nation and broad regional areas, such as the regions defined by the Bureau of Economic Analysis (BEA).⁴ These studies have generally found that national shocks account for slightly more than one-half of the variation in

The phrase “economic shock” represents economists’ shorthand for a factor or force that causes unexpected changes in economic variables.

the economies in both southeastern Michigan and northeastern Ohio.

The third type of economic shock is one that directly affects the locality itself: the San Francisco earthquake and fire of 1906 or the flooding that has recently afflicted Prague. The imposition or elimination of taxes by a municipal government also directly affects the local economy, as can passage of “living wage” legislation, which mandates that firms pay wages higher than the national minimum wage.²

SORTING OUT SOURCES OF LOCAL EMPLOYMENT FLUCTUATIONS

Economists have conducted a substantial amount of research trying to sort out how much of the fluctuations in employment growth at the national, regional, and local levels is due to

economic activity. However, the particular data that underpin the analyses turn out to have an important influence on the studies’ conclusions. Some very recent studies that look at monthly data, instead of quarterly or annual, for smaller geographical areas (cities and metropolitan areas) have found a considerably smaller role for national shocks and a correspondingly higher one for sectoral shocks.

Studies Using Quarterly and Annual Data for Regions. In an influential article, David Lilien suggested that frictions associated with the reallocation of workers across industries in the economy accounted for a substantial portion of fluctuations in aggregate unemployment. Lilien’s paper inspired a considerable amount of research examining the extent to which sectoral disturbances contribute to fluctuations in national economic

² In December 1994, Baltimore Mayor Kurt Schmoke signed into law one of the nation’s first living wage ordinances. Nearly 40 cities have passed some form of living wage law since then. According to David Neumark, these ordinances entail much higher wage requirements than traditional minimum wage legislation.

³ Much less research has focused on identifying how much of the fluctuations in employment growth in metropolitan areas is due to the effects of local shocks, such as natural disasters.

⁴ The BEA regions are New England, Mideast, Great Lakes, Plains, Southeast, Southwest, Rocky Mountain, and Far West.

activity. A majority of the studies indicate that sector-specific shocks play an important role in variations in employment and output for the national economy. In a 1996 review of the literature, Michael Horvath and Randal Verbrugge concluded that, on average, 40 percent to 45 percent of the variation in total economic activity in the nation could be attributed to sectoral shocks.

The majority of studies done at the sub-national level have looked at broad regions. A number of these studies have looked at the extent to which sector-specific disturbances contribute to fluctuations in employment or output in each of the major regions of the United States.⁵ These studies indicate that sector-specific disturbances account for between 35 percent and 67 percent of total variation in regional economic activity. Taking an average across these studies indicates that about 45 percent of regional fluctuations in output or employment can be attributed to sectoral shocks, roughly the same average that other studies have found for the nation. Several studies that looked at the issue for Canadian regions produced estimates that also yielded an average of 45 percent.⁶

Some Shortcomings of Regional Studies. The studies just reviewed indicate that sectoral shocks explain approximately one-half of fluctuations in regional economic activity. There are reasons, though, to suspect that studies based on quarterly or annual data for broad regions systematically understate the role of sectoral shocks. Michael Horvath and

Randal Verbrugge pointed out that studies based on quarterly or annual data might erroneously characterize shocks that actually have origins in a specific sector as having a common, non-sector-specific source. Over time, shocks initially specific to an industry tend to be transmitted to other industries through trade. For example, a

As an alternative to the approaches used in the previously discussed studies, we can examine monthly data on employment in cities or in metropolitan statistical areas (MSAs).

strike in the automobile industry will eventually affect employment in the steel, rubber, glass, and plastics industries and in all the other industries that supply inputs to auto producers. In turn, firms in industries that supply goods and services to the steel, rubber, glass, and plastics industries will be affected, and so on. If these disturbances propagate rapidly — say, within one quarter — disturbances initially specific to a sector (autos, in our example) will appear as a national shock (because it directly or indirectly affects other industries) in studies that use quarterly data.

In addition, by using data for broad regions, these studies run the risk of further tilting the findings to favor the influence of national shocks. That's because it's possible that positive disturbances to an industry in one area of a broad region (such as a BEA region) are likely to be offset by negative disturbances to that industry in other

areas of that region. Suppose a steel producer closes its Pittsburgh operation, resulting in a loss of 5000 jobs in the steel industry in the Pittsburgh metropolitan area. At the same time, an expansion of an existing steel plant in Philadelphia results in 5000 additional jobs in the Philadelphia metropolitan area. This change would leave steel employment in Pennsylvania unaffected. But steel employment (and total employment) in the Pittsburgh area would fall, while steel employment (and total employment) in the Philadelphia area would rise. Thus, the measured effects of shocks that have their origins in a specific industry are likely to be much smaller for broad regions than for local areas within the region.

Studies Using Monthly City and MSA Data. As an alternative to the approaches used in the previously discussed studies, we can examine monthly data on employment in cities or in metropolitan statistical areas (MSAs). Using monthly data limits the problem of shocks that are rapidly transmitted to other sectors; using city or MSA data limits the possibility that shocks within a broader region cancel each other out.

A 1999 study by Ed Coulson and a study that I did with Robert DeFina and Keith Sill overcame some of the data issues by looking at monthly data for sectors in cities (Coulson's study) and sectors in MSAs (our study). Both studies used a statistical technique known as a vector autoregression (VAR).⁷ Coulson's study looked at employment growth in eight broad sectors in four cities (Baltimore, Denver, Houston, and New York).⁸ The model

⁵ See the articles by Stefan Norrbin and Don Schlagenhauf; Todd Clark; Clark and Kwanho Shin; and Tamin Bayoumi and Eswar Prasad.

⁶ See the article by Joseph Altonji and John Ham and the one by Eswar Prasad and Alun Thomas. See Clark and Shin's article for an excellent review of the literature.

⁷ A VAR is a widely used modeling technique for gathering evidence on business-cycle dynamics. VARs typically rely on a small number of variables expressed as past values of the dependent variable and past values of the other variables in the model. See Theodore Crone's article for a discussion of VARs as applied to regional analysis.

for each city included 16 equations, one equation for local employment in each of the eight sectors to capture local disturbances plus one equation for national employment in each of the eight sectors to capture national disturbances to each sector.⁹

Coulson's findings showed that shocks to local sectors are much more important than shocks to national sectors in accounting for volatility in city employment growth. His study suggested that disturbances to local sectors explain between 67 percent (Baltimore) and 97 percent (Denver) of the variation in total employment growth.¹⁰

Our study looked at employment growth in seven of the eight broad sectors used in Coulson's study but in five MSAs (Chicago, Los Angeles, Oklahoma City, San Francisco, and Tucson). Our model included nine

⁸ In Coulson's study, the beginning date is January 1949 for Baltimore, January 1950 for New York City, and January 1970 for Denver and Houston. The ending date is April 1996 for all four cities.

⁹ Coulson's model does not include variables to separately control for the effects of monetary and fiscal policies or for city-specific shocks, such as natural disasters. The effects on city employment of monetary and fiscal policies are indirectly "picked-up" in the model's variables for national employment by sector. Similarly, the effects on city employment of shocks to the city's economy are indirectly "picked-up" by the model's variables for city employment by sector.

¹⁰ In another study, Kenneth Kuttner and Argia Sbordone provided some details on the factors that affect the performance of New York City, one of the cities also studied by Coulson. They found that while the economy of New York City usually tracks expansions and contractions of the national economy, the relationship is far from a lockstep one. For example, Kuttner and Sbordone found that much of the slow growth in New York City's employment during the late 1980s and early 1990s can be traced to weakness in the financial services industry, although the study did not determine how much of the weakness in the financial services industry was due to shocks to the industry in New York City and how much was due to national shocks to the industry.

equations, one for each of the seven local sectors plus two equations to capture common national economic disturbances (Table 1).¹¹ To account for common or national disturbances, the model included the level of the three-month Treasury bill rate and the

Our study found that among individual sectors, shocks in manufacturing explain more of the variation in total employment growth in the Chicago and Los Angeles MSAs than in the Oklahoma City, San Francisco, and Tucson MSAs.

monthly growth rate of national productivity. The assumption is that the Treasury bill rate and national productivity are influenced by national developments and not by developments in any given sector or any particular metropolitan area. The study used monthly data for 1951 to 1999. Our model did not separately identify shocks specific to a metropolitan area or national shocks to individual industries. These shocks are "picked up" through their effects on specific industries that make up the local economy.

Similar to Coulson's findings for cities, our findings demonstrated that sectoral disturbances are much more important than national disturbances that are common across industries in accounting for fluctuations in metropolitan employment growth. In fact, our study found that sectoral disturbances account for between 87 percent of volatility in employment growth in the

¹¹ Our study omitted the mining sector, since this sector typically accounts for a tiny share of employment in most metropolitan areas. The five metropolitan areas used in our study were chosen because they are the only metropolitan areas for which monthly data are available over a long period, namely, 1951 to 1999, for each of the remaining seven sectors.

Los Angeles MSA to almost 94 percent in the Tucson MSA.

Among sectors, the bulk of the evidence in both studies indicated that shocks to government, manufacturing, and nonfinancial services accounted for a substantial portion of volatility in local

employment growth. Shocks in these three sectors accounted for one-half or more of the variation in total employment in three of the four cities studied by Coulson (Table 2) and in all five metropolitan areas in our study (Table 3); they accounted for 43 percent of the variation in employment in the city of Houston.

Our study found that among individual sectors, shocks in manufacturing explain more of the variation in total employment growth in the Chicago and Los Angeles MSAs than in the Oklahoma City, San Francisco, and Tucson MSAs. Similarly, Coulson reported that manufacturing explained more of the variance in total employment growth in Baltimore and New York than in Denver and Houston. Both studies found that disturbances to the nonfinancial services and government sectors are generally important in accounting for total variance in employment growth.¹²

¹² Nonfinancial services consist of employment in personal services, health care, business support services, legal services, and social services. Employment in services such as finance, insurance, and real estate is excluded from the nonfinancial services category, as are jobs in wholesale and retail trade.

TABLE 1

Share of Total Employment Accounted for by Major Industry (Average share for the period 1951-99)

Metropolitan Areas						United States
	CHICAGO	LOS ANGELES	OKLAHOMA CITY	SAN FRANCISCO	TUCSON	
Government	11.6	13.2	24.4	18.7	23.2	15.3
Mining	0.2	0.4	3.4	0.2	3.3	0.9
Construction	4.0	3.9	5.1	5.0	7.9	4.6
Manufacturing	27.2	26.7	12.3	15.1	11.2	22.3
Trans. and Public Utilities	6.9	5.9	6.2	9.3	6.1	5.7
Wholesale and Retail Trade	22.6	22.4	24.4	22.5	22.4	28.1
Finance, Insurance, and Real Estate	6.8	5.8	6.0	8.0	4.3	5.0
Nonfinancial Services	20.7	21.7	18.1	21.3	21.7	18.0

Source: Carlino, DeFina, and Sill (2001).

Changes in the Treasury bill rate and national productivity growth typically played a considerably smaller role than disturbances to industries within the metropolitan area in accounting for the volatility in a metropolitan area's employment growth. Our results indicated that the largest *combined* effect of these common national disturbances is 13 percent in Los Angeles; the smallest is 6.3 percent in Tucson (Table 4). Thus, national disturbances are relatively unimportant for understanding fluctuations in individual MSAs' employment growth. Still, our study showed that the five

metropolitan areas responded differently to changes in the national economic variables used in the study.

The Importance of a Region's Size. Earlier, I pointed out that looking at the major regions of the country rather than at MSAs might increase the measured impact of national disturbances and decrease the impact of sectoral disturbances. Is that the case?

To look at this issue, DeFina, Sill, and I estimated two additional VARs. The first is called a five-metropolitan-area aggregate model. For this model, we constructed an aggregate

region by summing the data over the five metropolitan areas in our study for each industry. The second model, called the nation model, simply used national data for each industry.

An important finding from these additional estimates is that the measured impact of national disturbances does, in fact, increase as the level of the data increases, first from metropolitan area to region and again as we move from the region to the nation (Table 4). Changes in national economic variables account for about 16.7 percent of the fluctuations in the five-metropolitan-area model and 41.1

percent of fluctuations in the nation model. Similarly, changes in national economic variables explain a much smaller share of the variation in employment growth in each of the five metropolitan areas than was found for either the nation or the five metropolitan areas as a whole.

CONCLUSION

An important issue facing economists and policymakers is the degree to which fluctuations in economic activity can be attributed to sector-specific disturbances and the degree to which the fluctuations are due to forces common across sectors. Research on this issue for the national economy suggests that sectoral disturbances account for approximately one-half of the fluctuations in total economic activity.

Studies at the local level find a more significant role for sectors in accounting for fluctuations in economic activity than national studies, suggesting

TABLE 2

How Local Sectors Contribute to City Employment Growth*

(Average response, in percent)

	Baltimore	Denver	Houston	New York City
Source of Variance				
Government	14.0	28.7	14.8	41.9
Manufacturing	25.6	14.1	11.4	27.4
Nonfinancial Services	18.5	14.5	16.8	5.8
Construction	2.3	14.6	16.3	1.7
Trans., Comm., and Utilities	3.2	5.4	6.2	3.4
Trade	3.3	16.6	10.0	3.1
Finance, Insurance, and Real Estate	0.3	1.9	1.6	0.9
Mining	0.0	0.9	3.4	0.0

*Percent of variation in total employment growth accounted for by disturbances to specific sectors. Columns do not sum to 100 because national sectoral contribution to city growth is not included in the table.

The beginning dates are 1949 for Baltimore, 1950 for New York City, and 1970 for Denver and Houston.

The ending date is 1996 for all four cities.

Source: Coulson.

TABLE 3

How Sectors Contribute to Local Employment Growth*

(Average response for the period 1951-99, in percent)

	Chicago	Los Angeles	Oklahoma City	San Francisco	Tucson
Source of Variance					
Government	16.3	5.6	14.9	10.4	32.3
Manufacturing	32.6	34.5	21.0	18.4	17.3
Nonfinancial Services	13.0	18.9	24.3	20.9	7.9
Construction	7.3	7.9	6.5	14.6	16.8
Trans., Comm., and Utilities	7.5	5.9	7.6	11.6	6.3
Trade	7.9	8.3	14.0	10.3	10.8
Finance, Insurance, and Real Estate	6.4	5.9	4.8	2.6	2.4
Treasury Bill Rate	7.9	6.6	2.6	4.3	1.9
Productivity Growth	1.1	6.4	4.3	6.8	4.4

* Percent of variation in metropolitan area total employment growth accounted for by disturbances to specific sectors, Treasury bill rate, and productivity growth.

Source: Carlino, DeFina, and Sill.

that at least two-thirds of the fluctuations are due to sectoral disturbances.

These findings raise an important issue facing national and local policymakers. Large differences in fluctuations in economic activity across metropolitan areas can make it difficult for national policymakers to maintain low unemployment and low inflation in all parts of the country. Attempts at stimulating the economy during a national downturn in business conditions, for example, may lead to tight labor markets and falling unemployment rates in some parts of the country while others lag behind. If most disturbances to local economies have their origins in

REFERENCES

Altonji, Joseph G., and John C. Ham. "Variation in Employment Growth in Canada: The Role of External, National, Regional, and Industrial Factors," *Journal of Labor Economics*, 8 (1990), pp. 198-236.

Bayoumi, Tamin, and Eswar Prasad. "Currency Unions, Economic Fluctuations, and Adjustments: Some New Empirical Evidence," *IMF Staff Papers*, 44 (1997), pp. 36-58.

Carlino, Gerald A., Robert H. DeFina, and Keith Sill. "Sectoral Shocks and Metropolitan Employment Growth," *Journal of Urban Economics* 50 (2001), pp. 396-417.

Carlino, Gerald A., and Robert H. DeFina. "The Differential Regional Effects of Monetary Policy," *Review of Economics and Statistics* 80 (1998), pp. 572- 87.

Clark, Todd E. "Employment Fluctuations in the U.S. Regions and Industries: The Roles of National, Region-Specific, and Industry-Specific Shocks," *Journal of Labor Economics* 16 (1998), pp. 202-29.

TABLE 4

Percent of Variation in Total Employment Growth Accounted for by National Disturbances

Chicago	Los Angeles	Oklahoma City	San Francisco	Tucson	Five-Metro-Area Aggregate	Nation
9.0	13.0	6.9	11.1	6.3	16.7	41.1

Source: Carlino, DeFina, and Sill.

specific sectors, as these studies suggest, national and local policies that promote

labor mobility across sectors might serve as a useful adjustment mechanism for local economies. 

Clark, Todd, and Kwanho Shin. "The Sources of Fluctuations Within and Across Countries," in G.D. Hess and E. von Wincoop, eds., *International Macroeconomics*. Cambridge, UK: Cambridge University Press, 2000.

Coulson, N. Edward. "Sectoral Sources of Metropolitan Growth," *Regional Science and Urban Economics*, 29, (1999), pp. 723-43.

Crone, Theodore M. "A Slow Recovery in the Third District: Evidence From New Time-Series Models," Federal Reserve Bank of Philadelphia *Business Review* (July/August 1992).

Hooker, Mark, and Michael M. Knetter. "The Effects of Military Spending on Economic Activity: Evidence from State Procurement Spending," *Journal of Money, Credit, and Banking* 29 (1997), pp. 400-21.

Horvath, Michael T.K., and Randal Verbrugge. "Shocks and Sectoral Interactions: An Empirical Investigation," mimeo (June 1996).

Kuttner, Kenneth N., and Argia M. Shordone. "Sources of New York Employment Fluctuations," Federal Reserve Bank of New York *Economic Policy Review* 3 (1997), pp. 21-35.

Lilien, David. "Sectoral Shifts and Cyclical Unemployment," *Journal of Political Economy* 90 (1982), pp. 777-93.

Neumark, David. "Do Living Wage Laws Help Low-Wage Workers and Low-Income Families?" Public Policy Institute of California, *Research Brief* 55 (March 2002).

Norrbin, Stefan C., and Don E. Schlagenhauf. "The Role of International Factors in the Business Cycle," *Journal of International Economics*, 40 (1996), pp. 84-104.

Prasad, Eswar, and Alun Thomas. "A Disaggregated Analysis of Employment Growth Fluctuations in Canada," *Atlantic Economic Journal* 26 (1998), pp. 274-87.

How Inflation Hawks Escape Expectations Traps

BY SYLVAIN LEDUC

Why did inflation increase so dramatically from the 1960s to the 1970s? That's a question economists are still debating. One possible theory, however, is that once people started *believing* inflation would rise, the Fed was forced to validate those expectations by increasing the money supply. Sylvain Leduc discusses this "expectations-trap" hypothesis and uses a direct measure of expectations to see if the theory is consistent with the data.

In the early 1960s, inflation in the U.S. was below 2 percent, but by the late 1970s, it was in double digits. Why the inflation rate increased so much over such a relatively short period is still highly debated. Among the different views, one is particularly controversial. The expectations-trap hypothesis suggests that inflation rose dramatically over that period because the Fed, by projecting a dovish image, painted itself into a corner: For whatever reasons, once the public started *believing* inflation would rise, the Fed was forced to validate those expectations by increasing the money supply in the economy. According to this view, doing otherwise

would have been too costly. This article discusses the expectations-trap hypothesis, then uses survey data on inflation expectations to see if a sudden rise in that variable could have led to a burst of inflation.

The expectations-trap hypothesis is controversial because it implies that the same set of economic fundamentals, such as industrial production and the unemployment rate, can lead to a drastically different inflation rate, depending on how the public interprets the data and their effects on future inflation. One practical implication of the expectations-trap hypothesis is that it becomes very difficult for theorists and forecasters to predict inflation rates because any inflation rate can be rationalized from a given set of economic fundamentals. The theory could be right or wrong, but in general, it's hard to tell from the data, since we don't know how people will interpret any given piece of economic news.

In this article, I will present an analysis that tries to get around this problem using a data set, maintained by the Federal Reserve Bank of Philadelphia, specifically designed to gather information on expected inflation. By using this direct measure of expectations, we can verify whether the theory is consistent with the data. The empirical analysis will show that the predictions of the expectations-trap hypothesis match the U.S. experience surprisingly well.

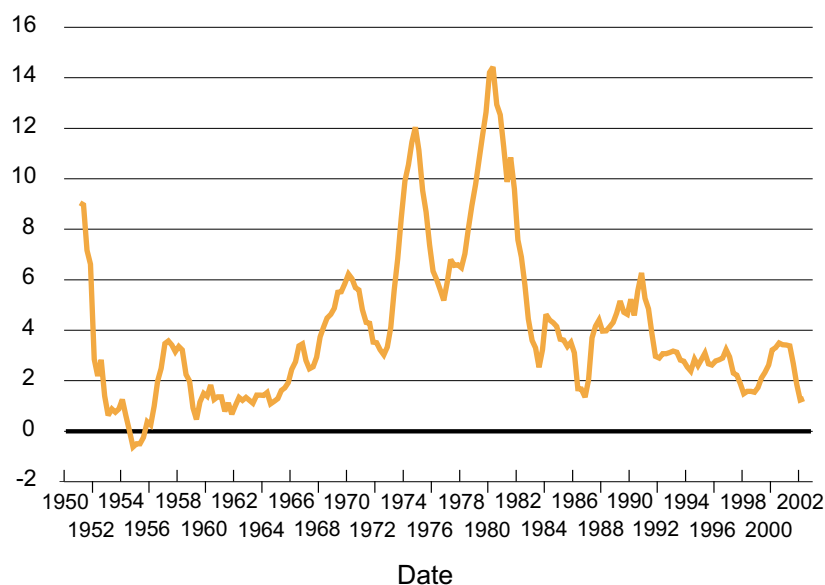
Obviously, the economy has changed substantially since the 1970s. The inflation rate has come down dramatically since the end of that decade; it averaged only 2.5 percent a year in the 1990s (Figure 1). Therefore, it may seem that understanding the causes of the inflation run-up of the 1970s would be mainly of academic interest. Yet, this is hardly the case. What triggered inflation to take off has important consequences for policymakers and the conduct of monetary policy today. Has a change in Fed policymaking kept inflation under control since the 1970s? Or has the structure of the economy changed to one favoring low inflation? The 1970s, after all, were much more turbulent than the 1990s. In the 1970s, there were two oil embargoes, and the Vietnam war was still going on. In the 1990s, there was an amazing increase in labor productivity growth. It is certainly easier to control inflation when you are in an environment of fast productivity growth that keeps production costs under control. But if policymaking hasn't changed, and we've just been lucky since the end of the 1970s, the corollary



Sylvain Leduc is a senior economist in the Research Department of the Philadelphia Fed.

FIGURE 1**CPI Inflation**

Percent



Note: CPI inflation rate is the percent change year-to-year using quarterly data.

is that inflation can take off again when our luck runs out and economic conditions change. Thus, knowing the causes of the inflation run-up of the 1970s is relevant today.

The empirical work in this article demonstrates that changes in the economic environment are not the only reason for inflation's performance over the past two decades: The conduct of monetary policy must share some of the praise. Indeed, our results show that monetary policymakers have become much more "hawkish" since the early 1980s: They have been more ready to forcefully raise interest rates to fight off sudden increases in expected inflation, a policy they weren't ready to follow in the 1970s.¹

¹ Just as the term hawkish describes a central banker who places more weight on achieving low inflation rates, a "dovish" central banker is one who is less prone to fight inflation, especially if fighting inflation entails lower output growth in the short run.

THE WORLD ACCORDING TO A.W. PHILLIPS

To understand what may have gone wrong in the 1970s, we first have to take a small detour to the world of British economist A. W. Phillips, who, in the late 1950s, published an article that would come to heavily influence policymaking and theoretical economics. His research documented a simple inverse relationship between the rate of growth in nominal wages and the unemployment rate in the U.K. Subsequently, a similar relationship was found between the rate of growth of the prices of goods and the unemployment rate in many different countries. This empirical relationship became known as the Phillips curve, and it led many academics and policymakers to believe that a lower rate of unemployment could be achieved by tolerating a higher inflation rate. That is, by exploiting the Phillips curve, academics and

policymakers thought they could reduce unemployment in the face of adverse events by increasing the money supply and, in so doing, generate inflation.

In a nutshell, the belief was that the unemployment rate could be as low as policymakers desired as long as they were ready to live with a higher rate of inflation. More important, a policymaker, basing his analysis on the Phillips curve, might believe that he could *permanently* lower the unemployment rate by simply creating more inflation.² That is, there would be a permanent tradeoff between inflation and unemployment. Since, in general, the costs of higher inflation are less apparent than those of higher unemployment, policymakers thought they had found an easy cure for the regular slumps associated with the business cycle. And if we look at U.S. economic performance in the 1960s, there were reasons to be optimistic: In 1969, the U.S. economy was in its eighth year of expansion, the longest such episode up to that time. However, the following decade would discredit this view, as the tradeoff between inflation and unemployment suddenly disappeared.

THE NATURAL RATE

As one example of the great power of good theorizing, Milton Friedman in the late 1960s argued that a long-run tradeoff between the inflation rate and the unemployment rate was pure fiction. He predicted that, in the long run, people would come to anticipate changes in monetary policy, adjust their expectations of future inflation rates, and thus neutralize monetary policy's effect on the real economy. In his view, only unanticipated changes in the money supply could affect output.

² For a broader discussion of these issues, see the *Business Review* article by Satyajit Chatterjee.

Suppose the central bank wants to lower interest rates to boost the economy. To achieve that goal, the Federal Reserve would reduce the federal funds rate, which is the rate banks charge one another for overnight loans. Although most people are not directly affected by the federal funds rate, the goal is to change very short-term interest rates, such as the fed funds rate, which then affect long-term real interest rates, which, in turn, do influence people's decisions to buy a car or a house or to save.³ The real interest rate affects people's decisions to spend or save because it dictates the tradeoff between consuming goods today or consuming them in the future. An increase in the real interest rate motivates people to increase their savings, which translates into a lower level of consumption today but a higher one in the future.

To lower the federal funds rate, the central bank would typically need to increase the money supply, which tends to generate inflation. Since the nominal interest rate is the sum of the real interest rate and the rate of expected inflation, a fall in the nominal interest rate would bring about a corresponding fall in the real interest rate only if the public does not expect a change in inflation in the future. But, with time, the public would come to realize that, to keep interest rates low, the central bank needs to increase the money supply, an action that tends to be inflationary. The obvious consequence is that the public would then adjust upward its expectations about the rate of inflation. People would then demand to earn a higher nominal rate of interest on their savings to compensate them for the

³ The real interest rate is the difference between the nominal interest rate — that is, the posted interest rate at which consumers borrow or save — and the rate of expected inflation.

higher expected inflation, which erodes the value of their savings in the future. Similarly, because of higher expected inflation, borrowers would be willing to pay a higher nominal interest rate. This process ultimately leaves the real interest rate unchanged, since rising nominal interest rates offset the increase in

In the long run, a strategy of pursuing an expansionary monetary policy that creates inflation to lower the unemployment rate will not work.

expected inflation. As a result, monetary policy will lose its ability to affect components of the real economy, such as output, once the public comes to anticipate the change in monetary policy.

Obviously, this is more likely to happen as time passes. In the short run, there may be a tradeoff between inflation and unemployment, but given time, people can gather more evidence that the Fed has instigated a change in policy and can adapt their expectations accordingly.⁴ Therefore, in the long run, a strategy of pursuing an expansionary monetary policy that creates inflation to lower the unemployment rate will not work. An expansionary policy will indeed increase the rate of inflation, but because it fails to lower real interest rates, it will leave the unemployment rate unchanged at its so-called natural rate.⁵

⁴ More specifically, all *nominal* variables, such as the price level and inflation, would be affected by a change in monetary policy in the long run, while all *real* variables, like unemployment, would be unchanged.

⁵ The natural rate of unemployment is determined by fundamental economic factors that tend to change slowly over time, such as demographics, technology, laws and regulations, and social mores.

Friedman's prediction that the tradeoff between unemployment and inflation would vanish as soon as policymakers tried to exploit it received a stunning confirmation just a few years after it was originally stated in his 1967 presidential address to the American Economic Association. By 1975, a new

term, stagflation, had indeed appeared in the economic jargon to characterize the state of the U.S. economy. Stagflation describes an economy with high and rising inflation and high unemployment.

CREDIBILITY AND THE EXPECTATIONS TRAP

What Friedman really pointed out is the importance of inflation expectations for the way changes in monetary policy are transmitted through the economy. His argument implies that monetary policy will lose its ability to stir the economy if the public comes to anticipate changes in policy and alters its inflation forecasts and that policymakers need to keep surprising the public for monetary policy to have some bite.

How changes in monetary policy affect expected inflation is particularly important when central banks have no way of committing to a particular policy, as pointed out in the work of Robert Barro and David Gordon. These authors argued that the rate of inflation would be higher than desired because a central bank, such as the Federal Reserve in the U.S., could not credibly commit to achieving a specific low inflation rate. Central banks often have multi-purpose mandates,

such as maintaining full employment and price stability, which may conflict in the short run.

To see this, consider an economy characterized by a short-run tradeoff between high inflation and low unemployment. That is, to lower the unemployment rate, the central bank needs to engineer a higher inflation rate. The central bank, having a mandate to maintain full employment and price

Proponents of the expectations-trap hypothesis argue that credibility is exactly what the Federal Reserve was missing in the 1970s.⁶ But, more important, because it was perceived as dovish, the Federal Reserve, according to this theory, could be caught in an expectations trap.

The story of the expectations trap usually goes as follows. Suppose there is a sudden rise in expected

the U.S. experience over that decade consistent with the predictions of this theory?

AN EMPIRICAL STUDY OF EXPECTED INFLATION IN THE 1970S

Economics is a science that likes discipline, and there is no better disciplinarian than data. Typically, models' predictions are compared with the (broad) features of the data to investigate whether a particular theory is consistent with the way the real world works. This is what Keith Sill, Tom Stark, and I did to study the expectations-trap hypothesis. However, applying the data to this particular theory is potentially controversial, since it implies knowing how people's inflation expectations change in response to news about the economy. We got around this problem by using the Livingston Survey, which started compiling data on expected inflation in 1946. Joseph A. Livingston, a journalist at the *Philadelphia Record* (and later at the *Philadelphia Inquirer*), started the survey; he polled business economists on their forecasts of some important economic variables, including the inflation rate. Since Livingston's death in 1989, the Philadelphia Fed has been conducting the survey, which polls forecasters from different sectors of the economy (nonfinancial corporations, academic institutions, and Wall Street investment banks) every six months in June and December.⁷

We introduced this measure of expected inflation into an empirical model — a simple vector autoregression (VAR) — to study the implications of a sudden rise in expected inflation for the economy. A VAR is a system of linear equations that link different variables

Proponents of the expectations-trap hypothesis argue that credibility is exactly what the Federal Reserve was missing in the 1970s.

stability, would like to achieve low inflation and low unemployment rates. But if the central bank announces a policy of price stability (zero inflation) in the future, no one will believe it. If the public does believe the central bank and expects prices to stay constant in the future, the central bank would have an incentive to generate a little bit of inflation to lower the rate of unemployment. Obviously, no one would be fooled by such a policy for very long, and the public would start taking into account this possibility when forming their expectations.

The main problem facing this hypothetical central bank is that its policy of price stability lacks credibility. The public can see through the central bank's rhetoric and understands the incentives the central bank is facing. Since a central bank lacking credibility would have a tendency to deliver too much inflation, Barro and Gordon went on to argue that credibility is thus a necessary ingredient for achieving low inflation rates. And to gain credibility a central bank must have a clear anti-inflation mandate and be shielded from political influences that will often be too willing to raise inflation in the hope of lowering the unemployment rate.

inflation. The central bank could adopt a more restrictive monetary policy and raise the federal funds rate to fight the increase in expected inflation, but this action has a cost. If there is indeed a short-run tradeoff between inflation and unemployment (that is, a Phillips curve), a rise in the federal funds rate will also lead not only to a lower inflation rate but also to a higher rate of unemployment. A dovish central bank, which assigns too much weight to output growth and not enough to inflation, may not be willing to pay that price. Instead, it would simply accommodate (and validate) the rise in expected inflation by leaving nominal interest rates unchanged. The expectations-trap hypothesis dictates that a sudden increase in expected inflation can therefore lead to a long-run rise in the inflation rate because the dovish central bank ends up validating the initial rise in expected inflation. Proponents of this view argue that the Fed was probably caught in such a trap in the 1970s. But is

⁶ See the article by V.V. Chari, Lawrence Christiano, and Martin Eichenbaum and the one by Lawrence Christiano and Christopher Gust for details on the expectations-trap hypothesis.

⁷ For a more detailed description of the Livingston Survey, see the *Business Review* article by Dean Croushore.

together. For instance, a VAR with two variables, let's say the inflation rate and the expected inflation rate, would also have two equations. One equation would try to explain the movements in inflation. The other would try to explain the movements in expected inflation using previous values of the rates of actual and expected inflation. Our VAR included the rates of inflation, expected inflation, and unemployment, as well as data on oil prices and the federal funds rate. The federal funds rate was included as an indicator of monetary policy. A rise in the real federal funds rate was associated with a tightening of policy, while a fall was interpreted as an expansionary policy. To investigate whether the inflation takeoff of the 1970s is consistent with the predictions of the expectations-trap hypothesis, we first looked at data from 1952 to 1979.

Using our model, we estimated what effect a sudden increase in the expected rate of inflation would have on the rest of the economy. We did that by determining the impact that the change in expected inflation would have on the other variables in our statistical model. We were particularly interested in the way inflation and nominal and real interest rates reacted to this change, since the behavior of these variables is at the core of the expectations-trap hypothesis. This theory states that the sudden increase in expected inflation would be followed by an expansionary monetary policy, since the dovish Fed, fearing the impact on economic activity, would not want to fight the rise in expected inflation with higher real interest rates. As a result, the temporary rise in expected inflation would lead to a fall in the real interest rate and a long-lasting increase in the actual inflation rate.

The first column of Figure 2 shows the responses of some of the variables in our model in the 1952-79 period to a one-time, unanticipated

increase in expected inflation. The solid line in the charts represents the estimated response of the variable to the sudden change in expected inflation; the dotted lines tell us how much confidence we can place on this estimate. In particular, when the dotted lines are both above zero or both below zero, we can say with a 90 percent level of confidence that the estimated response of, say, inflation to the unanticipated jump in expected inflation is significantly different from zero — that

Inflation rose dramatically in the 1970s because the Fed was perceived as too dovish and was susceptible to an expectations trap.

is, the unanticipated jump has an impact on the variable. For instance, following the jump in expected inflation, actual inflation increases about 1 percent and climbs to 1.5 percent one year after. Then the rate of actual inflation starts falling and stabilizes at approximately 1 percent higher than it would have been without the sudden increase in expected inflation. Also, if you look at the dotted lines in the figure for actual inflation, you can see that both of these lines remain above zero until 10 years after the initial jump in expected inflation. Our model, therefore, predicts that an increase in expected inflation would have a positive impact on the actual inflation rate for 10 years.

Moreover, the figure shows that this effect is the result of more expansionary monetary policy. Although the figure shows that the nominal interest rate rises following the shock, it does not rise as much as the rate of expected inflation, which translates initially into a *lower* real interest rate, as the expectations-trap hypothesis predicts. For instance, immediately

following the increase in expected inflation, the real interest rate falls a half of a percent. And except for the second year (seen in the bottom chart), the real interest rate is about 0.25 percent lower than it would have been without the sudden rise in expected inflation.⁸

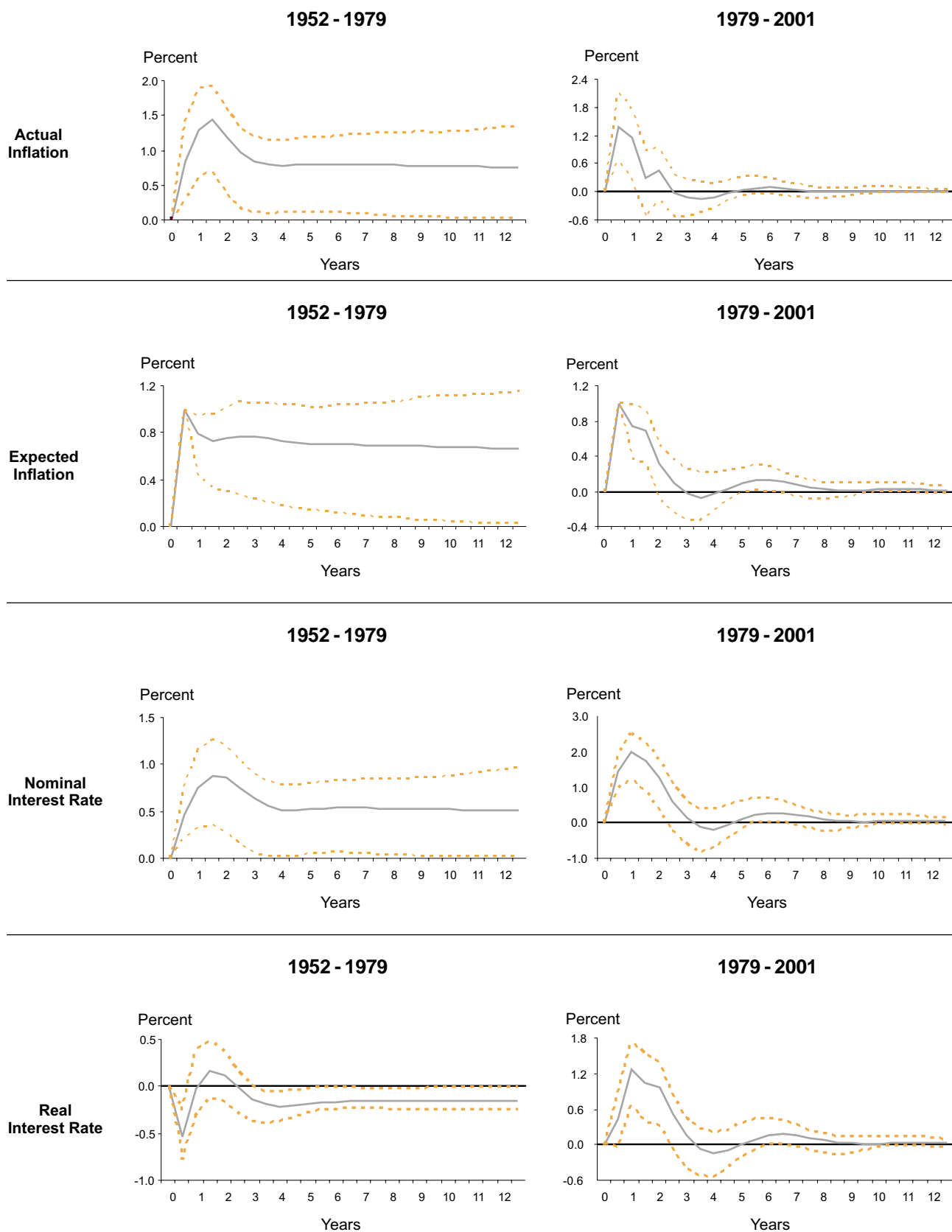
Our VAR model, therefore, offers some evidence to support the argument that inflation rose dramatically in the 1970s because the Fed was perceived as too dovish and was susceptible to an expectations trap.

Nevertheless, over the following two decades, inflation in the U.S. has declined steadily. Have we been enjoying a streak of lucky breaks, or has something more fundamental changed?

LUCK OR POLICY? AN EMPIRICAL STUDY OF THE POST-1979 ERA

As mentioned in the introduction, the inflation rate has come down dramatically since the end of the 1970s; it averaged only 2.5 percent a year in the 1990s (Figure 1). Although it is certainly true that the economy has experienced some changes that may have contributed to our luck, there is still a good reason to think that it was a

⁸ In the case of the real interest rate, both dotted lines are below zero most of the time, implying that the rise in expected inflation has a negative impact on the real interest rate. However, only in the second year does the real interest rate rise above zero, but in this case, the rise is not significantly different from zero, since the dotted lines are on both sides of the zero line.

FIGURE 2**Responses to a Shock to Expected Inflation**

change in policymaking that mainly contributed to bringing inflation to its knees.

In 1979, soon after OPEC agreed to boost oil prices for the second time in the decade, President Carter appointed Paul Volcker as Chairman of the Federal Reserve. In many ways, this appointment is now regarded as one of the most important policy changes enacted since WW II, a change that is often viewed as the Waterloo for rampant inflation. For one thing, Volcker never believed that a little inflation could cure the vagaries of the business cycle. He further believed that tighter monetary policy was, by then, a necessity and that the Fed needed to be immune from political imperatives. His chairmanship (and the following one of Alan Greenspan) would transform the dovish Fed of the 1960s and 1970s into a more hawkish one. For the economy to prosper, the Fed had to provide the business community with an environment in which prices were stable, thus facilitating business decisions.

The Volcker and Greenspan anti-inflation policy should have resulted in making movements in expected inflation less likely to become self-fulfilling. Since they believed that the best way to achieve maximum sustainable economic growth was to keep inflation under control, any indication that the public anticipated a surge in inflation should have been actively fought.

To verify this conjecture, we also conducted our previous experiments using data for the period after Volcker's appointment as Chairman of the Fed. We again looked at the effects on the economy of an unanticipated rise in expected inflation (second column of Figure 2). The figure shows that, since


1979, the Fed has not tolerated sudden increases in expected inflation and has been ready to forcefully raise the nominal interest rate to fight it off — so much so that the real interest rate rises. The figure shows that the Fed's policy response, since 1979, implies that a sudden increase in expected inflation does not generate a permanent rise in actual inflation: The inflation response quickly comes back down toward zero. In this sense, the Fed has been ready to implement a more restrictive monetary policy, by raising real interest rates, whenever it sees the public doubting the Fed's resolve to keep inflation low. With Volcker's appointment as Chairman of the Fed, the central bank stopped validating sudden increases in expected inflation through a more expansionary monetary policy. Therefore, contrary to the pre-1979 era, the post-1979 data show that surges in expected inflation have not had a long-lasting impact on actual inflation.

For many, the change in policy instigated by Volcker largely contributed to the Fed's success in taming inflation, and the results from our analysis agree with that view.⁹ Chance and particular circumstances may have helped, but they alone cannot explain the behavior of inflation since the mid-1960s. Volcker used tight monetary policy to squeeze inflationary expectations out of the U.S. economy, even if that policy turned out to have major consequences for economic activity in the short run. As the Fed kept tight control over the money supply, nominal interest rates ballooned, and real GDP, in 1981-82, suffered its most dramatic drop since the

⁹ For instance, see the article by Richard Clarida, Jordi Gali, and Mark Gertler.

Great Depression of the 1930s. According to the books by William Neikirk and Bernard S. Katz, the newly elected Reagan administration, which emphasized tax cuts to spur economic growth, was concerned that tight monetary policy could hinder the success of its policies. Yet, even though the Federal Reserve often clashed with the administration over the proper type of policies the central bank should adopt, Volcker forcefully defended the independence of the Fed from political influences.¹⁰ In this process, he helped build the credibility that the Federal Reserve enjoys in financial markets today.

CONCLUSION

Your word is often all you have. In some respects, this is also true for central bankers. Without credibility, the central bank has a much more difficult task in keeping inflation under control, in part because it is prone to falling into an expectations trap. And to get an economy out of a trap is not a trivial task: Drastic measures often need to be taken. The appointment of a hawkish Chairman to head the Federal Reserve in 1979 was a necessary decision in the fight against double-digit inflation. The recession of 1981-82 was certainly a high price to pay for bringing inflation under control, but the shift in policy in the early 1980s helped pave the way for 20 years of great economic performance. 

¹⁰ The arguments mostly involved Volcker and Donald Regan, who was then Treasury Secretary and who favored a more expansionary monetary policy. President Reagan and White House officials mostly supported the Fed in its fight against inflation. In fact, President Reagan reappointed Volcker as Chairman of the Federal Reserve for a second term in 1983 (see the books by William Neikirk and Bernard S. Katz).

REFERENCES

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

Chari, V. V., Lawrence J. Christiano, and Martin Eichenbaum. "Expectation Traps and Discretion," *Journal of Economic Theory*, 2, (1998), pp. 462-92.

Chatterjee, Satyajit. "The Taylor Curve and the Unemployment-Inflation Tradeoff," Federal Reserve Bank of Philadelphia *Business Review* (Third Quarter 2002), pp. 26-33.

Christiano, Lawrence J., and Christopher Gust. "The Expectations Trap Hypothesis," Federal Reserve Bank of Chicago *Economic Perspectives*, 24, (2000), pp. 21-39.

Clarida, Richard, Jordi Gali, and Mark Gertler. "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *Quarterly Journal of Economics*, 115, (2000), pp. 147-81.

Croushore, Dean. "The Livingston Survey: Still Useful After All These Years," Federal Reserve Bank of Philadelphia *Business Review* (March/April 1997), pp. 15-27.

Friedman, Milton. "The Role of Monetary Policy," *American Economic Review*, 58 (1968) pp. 1-17.

Leduc, Sylvain, Keith Sill, and Tom Stark. "Self-Fulfilling Expectations and the Inflation of the 1970s: Some Evidence from the Livingston Survey," Federal Reserve Bank of Philadelphia Working Paper 02-13 (August 2002).

Katz, Bernard S. *Biographical Dictionary of the Board of Governors of the Federal Reserve*. Westport: Greenwood Press, 1992.

Neikirk, William R. *Volcker: Portrait of the Money Man*. Chicago: Congdon & Weed, 1987.

Phillips, A.W. "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861-1957," *Economica*, 25 (1958), pp. 283-99.

Let A Hundred Flowers Bloom!

Decentralization and Innovation

BY LEONARD NAKAMURA

W

hich is more likely to encourage creativity and innovation: a centralized or a decentralized system of support? Should large organizations and recognized experts determine which parties get funding for their ideas? Or should small businesses, patrons, and foundations provide the primary support for innovation? Leonard Nakamura looks at the case for both sides using economic analysis, empirical studies, and anecdotal evidence. He also describes the role rivalry plays in innovation.

Michael Tomasello, co-director of the Max Planck Institute for Evolutionary Anthropology, has argued that what separates humans from chimpanzees and other nonhuman primates is their ability to maintain and build upon innovations — teaching children and peers the best ways to act and think. It can be argued that the cumulation of knowledge is not just the most important source of economic growth but also the most important factor in the flowering of human civilization and the dominance of our species on the planet.



Len Nakamura is an economic advisor and economist in the Research Department of the Philadelphia Fed.

Given the importance of knowledge, how should we organize its advance? Should innovation be centralized, with recognized experts determining which parties get funding to develop their ideas?¹ Or should innovation be decentralized, with small groups and individuals — small businesses, patrons, incubators, and foundations — supporting a lot of innovation? And to what extent can we rely on the market system to facilitate developing and disseminating new ideas and cultural products?

Of course, scientific, intellectual, and cultural genius does not appear

¹ Centralization refers to the existence of a single decision maker — a government agency, a monopoly firm, or a cartel — in a given industry, specialty, or product line that determines which innovative efforts to support.

simply because institutions are favorable. Innovation can occur when existing institutions are neglectful of it and even when they actively oppose it. But creativity is more likely to flourish and have its fruits more widely disseminated when it is recognized and supported. After all, artists, scientists, and scholars need offices, laboratories, and studios; they need time for their creative activities; and if their products are to matter, they need to find audiences — art dealers, students, talent scouts, journal editors, and the buying public.

The market system is often viewed as nearly synonymous with decentralization. But modern capitalism rewards innovation with monopoly rights. Copyrights and patents that protect intellectual and cultural property give innovators exclusive right to reproduce cultural, scientific, design, and engineering innovations. Thus, innovators gain property rights that may enable them to monopolize their markets and thereby possibly to control future access to innovation and distribution. Capitalism, by distributing resources to those successful at innovation, may encourage or discourage decentralization. This is currently an important policy issue, one aspect of which has been raised by the antitrust suit against Microsoft. Our question, in this context, becomes: Does market power, such as Microsoft's market power in software, encourage or discourage innovation? Parallel issues may arise, for example, in media mergers or in government research policy.

Similarly, government support for research need not imply centralization. Rather, research may also find

support from large and small profit-making firms and nonprofit organizations such as foundations and universities. So government research agencies may well be important players within an efficient and decentralized innovation network.

THE CASE FOR CENTRALIZATION

In recent decades, economic analysis has made important strides in understanding the advance of knowledge. An earlier strand of economic studies focused on the potential advantages of centralized innovation.

Barriers to Entry Support Innovation. Harvard professor Joseph Schumpeter was the seminal economic thinker on innovation and its role in the economy. He argued that developing and marketing new products was the key to economic development and that innovative firms needed to be repaid for this expensive process.

But if competitors are able to enter the markets for these new products and undercut the innovator, the price of the product will be bid down to its cost of production, and there will be no compensating profit for the innovating firm. To pay for development of new products, innovating corporations need to exclude imitative competitors for some period, to reap temporary supranormal profits. Corporations in some cases may be able to obtain monopoly power over their innovative products with intellectual property rights, such as patents and copyrights, trademarks, and brand names. Often, these will not be enough to adequately protect the innovation. The innovator may have to resort to alternative methods to protect its profits. For example, a firm may field a large sales force that specializes in selling the new product; building such a sales organization would be time consuming and costly for a potential entrant.

Going a step further, Schumpeter also argued that a large incumbent monopolist may have a strong incentive to innovate because a monopolist typically will have a large existing customer base to which it can quickly and easily market new products. Thus, a monopolist with a strong position in the marketplace can turn a profit on a new product far more quickly than a newcomer to the market would —

A good example of the potential for a public solution to the problem of innovation is vaccines.

raising the expected return to innovative activity. To be sure, Schumpeter was no apologist for perpetual monopoly. He believed that as long as entry was not impeded by regulation, all such monopolies were temporary, as entrepreneurs struggled amid a “gale of creative destruction.”²

However, there are drawbacks to innovation through a succession of temporary private monopolies. First, the monopolist uses its market power to sell its product at a high price. Therefore, some customers who would like to use the product — and who could afford to pay its marginal cost of production, but not the monopoly price — may not be able to buy it. Second, a monopolist will be reluctant to introduce innovations that compete directly with their existing products. Therefore, the monopolist's incentive to innovate in a given industry is generally lower than an outsider's. Finally, the monopolist incumbent may use its powerful position within the industry to reduce potential entrants' ability to introduce new products profitably.

² For a further discussion of Schumpeter's theory, see my *Business Review* article.

Public Involvement May Facilitate Innovation. For these and other reasons, economists and others have often argued that governments and public entities are better supporters of innovation and creativity. Indeed, in the United States, the National Science Foundation, the National Institutes of Health, the National Endowments for the Arts and the Humanities, and the military research and development

(R&D) paid for by the Department of Defense all bear testimony to the belief that the federal government is a natural source of such funding. In his 1960s exercise in social forecasting, Daniel Bell predicted the increasing socialization of knowledge production. Bell argued that knowledge (including innovations and creative products) is what economists now call a nonrival good because its transfer to and use by others does not reduce its benefit to original holders, unlike with material objects.³

The government can overcome the monopoly problem of prices being too high because it can pay for the fixed cost of innovation with taxes, then

³ Economists distinguish between public goods and nonrival goods. When an individual or group produces a public good, they can't exclude others from enjoying its benefits. For example, national defense is a public good. A nonrival good may be kept from others, but there is no direct additional cost to providing it to others. For example, an idea for a new innovation that is kept secret is not a public good, but it is a nonrival good. If the idea cannot be kept secret, it is a public good. Since over time knowledge tends to become public, whether it is considered public or nonrival depends on the time period considered. Thus, some economists regard knowledge as a public good.

distribute the innovation at marginal cost. A good example of the potential for a public solution to the problem of innovation is vaccines. When a vaccine is distributed widely enough, it may be possible to eliminate all hosts for a disease and thereby eradicate the disease itself, as appears to have occurred with smallpox. In two articles published in 2001, Michael Kremer argued that governments ought to pay inventors of vaccines the social value of the vaccine, then make the vaccine available at the lowest possible price.⁴

Moreover, government support of research may be valuable because although the research may have no immediate practical applications, it may provide the basis for more profitable research in the future. For example, many important mathematical insights have flowed from theorems that establish that two different and seemingly unrelated branches of mathematics share a common structure: Andrew Weil's recent proof of Fermat's Last Theorem is an example. Proofs of this kind often have no direct potential for profit, and indeed mathematical propositions are generally not patentable. (However, see Robert Hunt's article for changes in tests of patentability.)

⁴In general, Kremer, in a 1998 article, proposed using an auction to disclose the private expected value of the patent. Private parties would bid for the right to patent. Some of the time, the private parties would be allowed to win the patent, so the private parties would have a strong incentive to bid accurately. The winning auction bid should be a reasonable estimate of the value of the patent to a private monopolist. Most of the time, the government would step in and pay the inventor a *premium* over the auction price, and the premium plus the auction price reflects the average social value of the patent, a value that Kremer conservatively estimates is twice the private value. The social value includes both what the monopolist would earn and the consumer surplus (benefits to consumers above and beyond the price they pay).

Government support of basic research has long been accepted in the United States and has been an important source of the country's competitive advantage due to spillovers. Although basic research may not have direct applications, the expertise of those involved in it may be a valuable resource for more directly profitable enterprises.

Another implication of Bell's argument is that societies with strong business-government collaboration (Japan, Singapore) may also be good at creating and using knowledge. Indeed, the success of the East Asian model of economic development from 1960 to 1990 can be viewed as an illustration of this concept.

Centralization can prevent wasteful duplication of research: Private parties racing to produce the same innovation are likely to duplicate one another's efforts unnecessarily. Moreover, since research is highly risky, it is valuable for researchers to hedge the risk that their project will fail. That is, if several different researchers are working on separate lines of research or different attacks on the same problem, it is possible that only one will succeed. Since it is often impossible to predict which line of research is most promising, the successful outcome of one approach need not imply that only one of the researchers was of value and working hard. By centralizing and pooling support and funding, all may receive at least some reward. Indeed, the modern corporation with its research laboratories can be viewed as an institution for pooling risk in this way. Central and secure funding for innovators also may encourage scientists and artists to be more cooperative about sharing discoveries and techniques, further reducing risk and duplication of effort.

Arguments in favor of government support for research are reinforced when the research projects in question are very expensive. Examples

include space travel, particle colliders, the mapping of the human genome project, and nuclear fusion electrical generation. Moreover, coordination among private parties who might profit from the research may be difficult because sharing intellectual property rights can result in excessive competition. And when research is very expensive, even private-sector monopolists may find the project too risky to undertake.

Anecdotal Evidence for Centralized Innovation. Told as anecdotes about the accomplishments of big government and monopoly firms as innovators, much of the evidence from World War II and the two decades following appeared to favor Schumpeter's and Bell's arguments.

Researchers at Bell Labs, the research arm of AT&T and the regional Bell companies before their breakup, produced many inventions crucial to the modern age. Most famous among them was the transistor, the key breakthrough that brought Nobel prizes to three Bell Labs scientists and ushered in the electronic age. During that same period, Bell Labs developed much of the information science that underpins ever-increasing bandwidth, including information theory and coding theory.

IBM, the giant corporation that dominated the computer industry from the mid-1950s to the mid-1980s, developed many inventions crucial to the computer and electronics, including the development of the first major programming language, Fortran.⁵ One of IBM's great breakthroughs was the 360 computer series. Prior to the invention of the 360 series, computer operating systems were usually tailor-made for the particular computer model they ran on. Consequently, when companies wanted to upgrade their

⁵ See the book by Emerson Pugh.

computer systems, they would have to rewrite or adapt all their existing computer programs. The 360 operating system, by contrast, blended the computers in the 360 family so that computer programs written on smaller ones could run almost seamlessly on larger ones.

Government R&D has also had spectacular successes, including the Manhattan Project, which developed the first atomic bomb; the development of ENIAC, the first general-purpose programmable computer; and NASA, whose Apollo program successfully put astronauts on the moon within a decade.

Centralization Also Has Drawbacks. But central authorities — government and business monopolies — can fail to recognize the right path to innovation. If centralization requires consensus, it may be harder to make progress when the consensus is flawed. Encouraging iconoclastic innovation may require subjective judgments from science bureaucrats. But these well-intentioned government bureaucrats may be reluctant to break the mold for fear that they will be accused of arbitrary or self-serving behavior that would conflict with government accountability regulations.

As a consequence, the activity of nongovernmental supporters of research — whether they be for-profit corporations, nonprofit foundations, or universities — can be crucial to the speed of innovation. During the past three decades, a period of exceptionally rapid innovation, the government's share of R&D has declined. When we consider basic and applied research and product development, the federally funded share has fallen from 64 percent in the 1960s to 26 percent in 2000, while industry-funded research has risen to 68 percent (Table 1).

Private, rivalrous industries, such as pharmaceuticals, finance, and

semiconductors, just to name a few obvious ones, have been at the heart of much of modern innovation. This has increased interest in understanding how companies actively competing with one another might be good at conducting R&D. Moreover, more systematic views of the evidence have long suggested that the anecdotes about Bell Labs and IBM research oversold the case for big research centers.

THE CASE FOR DECENTRALIZATION

It may be that the top experts in a given field are not the best judges of innovation. One way of ensuring that many different talents and ideas have the opportunity to find an audience is to have many venues through which the people with talent and ideas can obtain funding and publicity. Decentralization thus may be a superior way to develop new products when it is hard to discern the best talents and ideas.

Free Entry Is Best When “Nobody Knows.” In his path-breaking book on the organization of creative industries, Richard Caves argues that in innovative and creative markets, no one can know in advance who will succeed, a condition he calls “nobody knows.” Decentralized gatekeepers — teachers, book and journal editors, movie producers, department chairs, art dealers, and curators — compete to develop new products and talents. Every success invites entry into the next round, and the right gatekeepers of today may not be the “hot hands” of tomorrow as markets, meanings, and tastes evolve. Under decentralization, the audience — whether scientific peers or customers — rather than the individual gatekeeper becomes far more important to the evolution of the industry in question.

For example, before 1948, the major Hollywood film studios had achieved substantial market power for

their products by vertical integration: The studios owned a large proportion of U.S. movie theaters. This enabled the studios to jointly control production and distribution and made new entry into film-making by independent producers a daunting task. Not only did the studios control their own theaters, but independent movie theaters often either had no access to the most popular films or were required to book multiple titles in advance without the power to review the titles, a practice known as blind booking. After these practices were declared illegal in 1948, the quality of films, as measured by their audience popularity, critical reviews, and awards, became much more important in determining studios' profitability and the success of their management.

In his 1982 article, Boyan Jovanovic developed a theory to model industrial performance under a “nobody knows” condition, in which firms discover whether they are “talented” by

TABLE 1

Sources of Support for All Types of R&D by Source of Funds

All R&D: Basic and Applied Research, and Development

Year	Federal	Industrial	Other
53-59	59.6%	38.1%	2.3%
60-69	63.9%	33.6%	2.5%
70-79	52.9%	43.6%	3.6%
80-89	45.7%	50.5%	3.8%
90-99	34.2%	60.6%	5.2%
2000	26.3%	68.4%	5.3%

Source: National Science Foundation, *Science and Engineering Indicators*, 2002.

“Other” includes universities and colleges, state and local finance of university and college research, and other nonprofit organizations.

facing the market test. We can think of talented firms as including literally talented entrepreneurs and also firms with intellectual property that provides them with a sustained advantage in innovative activity. In this model, talented firms grow bigger and make more profits. Industry productivity and profits increase over time as some firms learn that they are untalented and exit, and new entrants, some of them very talented ones that will survive, take their place.

In his book, Michael Porter argues that having free entry and rivalry keeps companies on their toes, encourages innovation and efficiency, and discourages political favoritism. The United States, by being home to Intel, Texas Instruments, IBM, Hewlett Packard, Motorola, and AMD — all producers of microprocessors — has obtained a sustained advantage in the computer industry because it has this kind of rivalrous industry.

Why does rivalry work so well? In part, because rivals give each firm a yardstick for performance. Excuses — whether made to a superior, to the government, to shareholders, or to oneself — just don't play as well when the competitor across the street or across town is doing better. Moreover, the visibility of the competitors' practices stimulates both emulation and one-upping — spillovers of information. And since new ideas by new entrants may be offered to a variety of bidders, outsiders are encouraged to add fresh talent to the mix. Overall, using a variety of industries and countries, Porter convincingly illustrates that nations that have such rivalrous industries obtain lasting national advantage over other countries.

On the negative side, rivalry often seems to incite deep personal antagonism.

How Rivalry Drives Innovation. In a series of papers, Philippe Aghion and co-authors developed a

formal theory that supports the value of rivalry in innovation. The authors describe industries that have step-by-step innovations and differentiated products, with one company sometimes breaking out of the pack with a new innovation. We can think of an innovation as being a new generation of a product line, such as a new generation of video game players, a new type of car like the minivan, or a new class of drugs. Because products are differentiated and some customers have strong preferences, one company's innovation does not drive its competitors from the market immediately, but the innovator's profits rise dramatically. The possibility of this dramatic rise in profits spurs innovation.

Consider two rivals, which we will call Inventor Bell and Tinker Bell, who are in the business of supplying custom cell-phone chimes. They share the market for 16-bit chimes but are racing to develop 32-bit chimes. Each knows that the first firm to come up with 32-bit chimes will win 80 percent of the market, which is sure to expand because 32-bit chimes will enable phones to play the "Star Wars" theme song. There are thus two effects: the market expands and the innovator gets a larger share.

Now suppose Inventor Bell is the first to invent and market the 32-bit chimes. Since Tinker Bell is able to examine Inventor Bell's product, Tinker Bell's cost of *imitating* are lower than Inventor Bell's cost of finding the *next* innovation.⁶ If Tinker Bell can succeed in imitating the innovation before

⁶ When the original innovation is patented, the follower has to find a way to imitate the product without violating the original patent. Edwin Mansfield, Mark Schwartz, and Samuel Wagner found that a majority of the sample of patented inventions they studied were successfully imitated within four years and that, on average, the cost of imitation was a third less than the cost of the original invention.

Inventor Bell moves on to, say, 64-bit chimes, its profits will rise substantially and Inventor Bell's profits will drop sharply. Therefore, Tinker Bell has a strong incentive to get back in the race. But if Inventor Bell moves on before Tinker Bell can imitate, it knows that Tinker Bell's incentive to innovate may drop sharply, as it will need two rounds of success to catch up. This might leave Inventor Bell with a clear field and a long period of very high profits. Thus, Inventor Bell has a very strong incentive to continue to innovate.

As long as there remain in the industry some competitors who haven't fallen very far behind, the incentives to innovate for all of the competing firms, both leaders and followers, will be high, as long as imitation isn't too easy. If imitation is too easy — if it's much cheaper to imitate than to innovate — the incentive to innovate will be muted because the leader retains its profits for too short a period to justify the expense of innovation.

Put another way, if Tinker Bell can imitate Inventor Bell's invention for one-tenth of the cost Inventor Bell paid to discover the invention, and vice versa, each will prefer the other to be the first innovator. Then it is possible that neither will invest in innovation. The result will be an industry that is competitive precisely because there has been no innovation or progress. In this model, innovation and competition will have an inverted-U relationship: The most innovative industries will be those with some competition, not those with lots of competition or monopoly.

DECENTRALIZATION MAY BE SUPERIOR WHEN ADVANCES ARE UNFORESEEABLE

One factor underlying the relative performance of centralized and rivalrous methods of innovation is how knowledge evolves. A key question is: To what extent does knowledge evolve

along foreseeable paths rather than result from old views being replaced by new ones?

If Scientific Advance Is Foreseeable. One way of looking at the evolution of knowledge is that knowledge is mainly cumulative. If this is true, the world should become more certain. The more we know, the more sure we will be in our knowledge. If our views are built on bedrock, that is, if fundamental theories are correct, new evidence will only confirm them. In this case, new truths do not displace old ones. Past knowledge is a reliable guide to future knowledge. Similarly, our concepts of what is beautiful — what constitutes a good painting or good music — are not subject to radical reconstruction.

In this type of world, senior experts are friends of progress. The theories that the senior experts have learned and taught are citadels. The most valuable new research in this situation extends the reach of existing theory into new areas and applications. This type of scientific advance enhances the value of existing knowledge, rather than conflicting with it. In a world like this, peer review committees function well, since senior scientists don't disagree too much.

Since knowledge is cumulative, older scientists tend to know more than younger scientists. Seniority is a good reason to pay someone more or to allow someone more decision-making authority.

But If Science Advances by Revolutions. Another possibility is that as knowledge advances, what we know often becomes obsolete. When new ideas threaten to make old ones obsolete, the incumbent experts may attempt to block the development of new ideas. As the pace of gathering knowledge accelerates, anomalies that contradict existing theories are likely to accumulate faster; so the advance of

knowledge more often requires making old theories obsolete.

A more subtle effect of this kind can occur when new inventions make old methods obsolete and thereby render old knowledge less useful. For example, the hand-held calculator rendered the slide rule and the ability to manually calculate square roots less

The development of the telescope led Galileo to discoveries that deepened questions about the Aristotelian-Ptolemaic theory of the universe.

useful. A senior scientist's or engineer's knowledge base can become outmoded in this fashion, even though it is not contradicted. Similarly, the advent of photography supplanted the purely documentary function of painting in favor of innovative and imaginative aspects.

The development of the telescope led Galileo to discoveries that deepened questions about the Aristotelian-Ptolemaic theory of the universe. The ability to measure the speed of light overturned the Newtonian universe. The ability to decipher genetic code will change our understanding of biology and evolution and, perhaps, may change what it is to be human.

Who knew that jazz would be the seminal form of American music in the 20th century? Not the musicologists of the time. But the invention of the phonograph, which captured the excitement of improvised music and made it available to the multitudes, made jazz a worldwide musical influence almost overnight.

Universities and Academic Stars. One of America's strengths in innovation is a diverse collection of private and public universities that have substantial freedom to hire academic

innovators. A recurrent question is: How should universities use appointments and tenure decisions to attract and support the best scholarship?

From the perspective of the individual university, the question is: How does a great university stay at the top? One way is to offer professorships to academics who have published path-

breaking research and have already achieved universal acclaim.

The alternative is to attract innovators when they are doing their best work — hiring them when they are doing the work that will win them their Nobel prizes, rather than after they win. This option is clearly riskier, since the innovator's work may not prove to be the best. The university may get stuck with the losers, particularly if other universities poach its stars. But if innovation is proceeding fast enough, hiring professors who are past their prime may gain the university a reputation for standing in the way of progress, rather than representing the best.

On the other hand, it may well be that great scholars or artists will be founts of creativity for a long time. For example, when Joseph Schumpeter came to the Economics Department at Harvard in 1932 at age 49, he still hadn't written three of the four works for which he is best known.⁷

⁷ Those works are *Business Cycles* (1939), *Capitalism, Socialism, and Democracy* (1942), and *History of Economic Analysis* (published posthumously in 1954). *The Theory of Economic Development* was first published in German in 1911.

A diversity of universities, each possessing its own methods for rewarding teachers for producing knowledge and for teaching, is an immense asset for any economy. Within this system of higher learning, competition — and decisions on how to compete — will play a powerful role in determining the overall rate of innovation in the economy.⁸

ANECDOTAL EVIDENCE FAVORING DECENTRALIZATION

The “low level of productivity of Eastern Europe relative to that in Western Europe,” as Stephen Nickell writes, is “an impressive example of what can be achieved by repressing the forces of market competition.” It was not just that the centralized economies of the Soviet bloc were inefficient; they also progressively fell behind at innovating and in adopting innovation.

In basic research, where findings may be too far from direct application to be profitable, the federal government remains predominant with a 49 percent share, although its share has declined (Table 2). But the grant-making of federal science support may not be able to adequately diversify its research base: How do we decide whether a research idea that goes against the mainstream might be successful anyway? Typically, peer review panels will not support such long shots. When this happens, private foundations and for-profit corporations can valuably supplement government support.

Craig Venter’s “shotgun” approach to genomic research was derided by members of the federally

funded Human Genome Project.⁹ Indeed, earlier he had proposed to use the technique to decode the genome of a bacterium, only to have his grant request rejected. He appealed the rejection; meanwhile he used private funding to proceed with his research on the bacterium. The NIH review committee rejected his appeal on the grounds that it was unfeasible. A few months after the rejection, he published his transcription of the bacterial genome in the prestigious journal *Science*. He then went on to use this method in the draft decoding of the human genome, again with private funding, substantially accelerating that landmark event.

As described earlier, IBM was one of the most successful innovators of the 1950s and 1960s. In the late 1960s, alarmed by the rapid advances in technology made possible by miniaturizing integrated circuits, IBM embarked on a project aimed at greatly increasing the usefulness of computers called FS (for Future System). As Emerson Pugh recounts in his history of IBM, although a number of technological breakthroughs occurred, IBM never came close to a marketable product and settled for a modest extension of the 360 series, which it called the 370 series.

After this failure, IBM became much more risk averse, and it became more difficult for innovative projects to advance cost effectively through the IBM project management process. Consequently, as described in Paul Carroll’s book, IBM was slow to enter the personal computing market. After a few in-house failures, IBM was forced to turn to outside sources — Intel for the microprocessor and Microsoft for the operating system. Although the IBM PC thus produced was an instant

success, IBM lost control of the PC market and both Microsoft and Intel profited more in the long run.

The French Academy of Beaux Arts supported students and artists. The identification of Paris with painting continued well into the 20th century. During the 19th century, art students from around the world, including Americans like Thomas Eakins, came to Paris to learn how to paint in the grand style. Yet painting was in the midst of an upheaval, beginning with Impressionism, which was foreign to the tastes of the reigning French painters of the academy and the official salons. As Annie Cohen-Solal’s book shows, the new painting came to prominence despite the opposition of the state-supported institutions of painting. Private art dealers, aristocratic patronage outside the academy, a network of independent teaching artists, and artists’ colonies both in Paris and in the provinces were all important sources of support for the new painting.

These examples point out not that government or large businesses

TABLE 2

Sources of Support for Basic Research by Source of Funds

Basic Research			
Year	Federal	Industrial	Other
53-59	57.1%	32.2%	10.8%
60-69	68.7%	19.2%	12.1%
70-79	70.0%	14.5%	15.5%
80-89	64.6%	19.4%	16.0%
90-99	55.3%	26.3%	18.4%
2000	48.7%	33.9%	17.5%

Source: National Science Foundation, *Science and Engineering Indicators*, 2002.

“Other” includes universities and colleges, state and local finance of university and college research, and other nonprofit organizations.

⁸ This issue made headlines at Harvard University when new President Larry Summers vetoed two appointments to the Harvard faculty, as described in the *Wall Street Journal* article by Daniel Golden.

⁹ See the magazine article by Richard Preston.

cannot successfully support research or creativity but that a proliferation of sources of support can be crucial to rapid progress in the arts, sciences, and commerce.

SYSTEMATIC STUDIES DON'T SUPPORT CENTRALIZATION

Systematic Empirical

Studies. A vast empirical literature investigating the relationship between competitiveness and innovation, most of it produced between 1965 and 1995, argued that there was little systematic relationship between competitiveness and innovation. Work by F. Michael Scherer showed that “there was little evidence of disproportionately great R&D input or output associated with the largest corporations, and market concentration showed no significant positive impact on progressiveness.” Studies summarized in Wesley Cohen and Richard Levin’s article echo this theme.

More recent studies, which were based on detailed and systematic data on innovations by industry in the United Kingdom, have argued that industry innovation declines as industry concentration rises. Richard Blundell, Rachel Griffith, and John Van Reenen found that while dominant firms tend to innovate more than other firms, this dominance dampens innovative activity for other firms in the same industry. On net, they found empirically that the dampening effect on the smaller firms outweighed the innovative activity of the dominant firm.

Stephen Nickell showed that industries that are more competitive have faster rates of innovation as measured by the rate of increase of *total factor productivity* (TFP).¹⁰ TFP measures the growth of industry output

that can't be accounted for by the growth of labor, capital, or materials alone. This is a very good measure of overall growth of innovation, since innovations are the main explanatory factor omitted from the measured inputs.

In 2002, a study by Philippe Aghion and co-authors garnered a result that is perhaps closer in spirit to Schumpeter's original argument. They found an inverted U-shape relationship between patenting activity and competitiveness: Very competitive industries have low patenting activity as do industries that are very profitable.¹¹ This is Schumpeterian in that too much competitiveness appears to be detrimental to innovation, since the rewards to innovation vanish too quickly to repay it. At the same time, the study found that monopolization of an industry results in too little innovation, perhaps for the reasons suggested by Michael Porter. Overall, these results conform to their model of step-by-step innovation.

¹⁰ Nickell's survey asked managers whether the company had more than five competitors in the market for its products. He used this as one measure of competitiveness. Another measure of competitiveness is profit margin, which is the amount a firm can charge for its products above costs, where costs include labor and capital. In general, competition should drive profit margins close to zero, so that large profit margins imply lack of competition. Nickell defined profit margins as profits less capital costs, divided by value added.

¹¹ Like Nickell, Aghion and co-authors use profit margins as an inverse measure of competitiveness. They define profit margins as operating profits divided by sales. Their analysis uses a series of changes in industrial regulation in Britain to identify the role of competitive conditions in influencing gains in industrial productivity.

CONCLUSION

Is our era one of incremental knowledge or of innovation? To the extent that it is an age of innovation, decentralized and competitive structures — whether capitalist or government supported, profit or nonprofit — will favor economic growth.

Can a centralized system decentralize? Ultimately, this is an empirical question. In principle, a monopoly can operate like a decentralized system. That is, a monopolist may be able to use internal competition — between managers or divisions — to obtain results similar to those obtained through market competition.

But the monopoly has its own incentives that may not align with progress. New products may reduce profits on existing products when the new products are successful and waste them when they are not. Moreover, new products may require changes in corporate focus that make production of existing profitable products less efficient.

Clearly the existence of rivalry — competing institutions that encourage innovation — is very valuable in generating innovation. Yet we should be mindful that even where competition has free rein, progress may be unnecessarily slow. Excessive competition may arise where imitation is too easy. Moreover, the return to innovative activity may be too distant from the innovation to provide adequate private incentives to create.

Thus, innovation may best be served when there are a wide variety of sources of support: large and small firms, small foundations and big government agencies, new-firm incubators, and venture capitalists. 

REFERENCES

- Aghion, Philippe, Nicholas Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. "Competition and Innovation: An Inverted U Relationship," Institute for Fiscal Studies Working Paper 02/04, February 2002.
- Aghion, Philippe, Christopher Harris, Peter Howitt and John Vickers. "Competition, Imitation, and Growth with Step-by-Step Innovation," *Review of Economic Studies*, 68, 2001, pp. 467-92.
- Bell, Daniel. *The Coming of Post-Industrial Society: A Venture in Social Forecasting*. New York: Basic Books, reissue edition, 1999.
- Blundell, Richard, Rachel Griffith, and John Van Reenen. "Dynamic Count Data Models of Technological Innovation," *Economic Journal*, 105 (429), March 1995, pp. 333-44.
- Caves, Richard E. *Creative Industries: Contracts between Arts and Commerce*. Cambridge, MA: Harvard University Press, 2000.
- Cohen, Wesley M., and Richard C. Levin. "Empirical Studies of Innovation and Market Structure," in Richard Schmalensee, and Robert D. Willig, eds., *Handbook of Industrial Organization*, Volume II. New York: North Holland, 1989.
- Carroll, Paul, *Big Blues: The Unmaking of IBM*. New York: Crown, 1994.
- Cohen-Solal, Annie. *Painting American: The Rise of American Artists, Paris 1867-New York, 1948*. New York: Knopf, 2001.
- Golden, Daniel. "Course Correction: Roiling His Faculty, New Harvard President Reroutes Tenure Track — Summers Boosts Hopes of Younger Professors, In a Scholarly Gamble," *Wall Street Journal*, January 11, 2002.
- Hunt, Robert M. "You Can Patent That? Are Patents on Computer Programs and Business Methods Good for the New Economy?" Federal Reserve Bank of Philadelphia *Business Review*, First Quarter 2001.
- Jovanovic, Boyan. "Selection and the Evolution of Industry," *Econometrica* 50 (3), May 1982, pp. 649-70.
- Kremer, Michael. "Patent Buyouts: A Mechanism for Encouraging Innovation," *Quarterly Journal of Economics*, 113, November 1998, pp. 1137-67.
- Kremer, Michael. "Creating Markets for New Vaccines: Part I: Rationale," *Innovation Policy and the Economy*. MIT Press, Vol. 1, 2001a.
- Kremer, Michael. "Creating Markets for New Vaccines: Part II: Design Issues," *Innovation Policy and the Economy*. MIT Press, Vol. 1, 2001b.
- Kuhn, Thomas S. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1996.
- Mansfield, Edwin, Mark Schwartz, and Samuel Wagner. "Imitation Costs and Patents: An Empirical Study," *Economic Journal*, 91, December 1981, pp. 907-18.
- Nakamura, Leonard, "Economics and the New Economy: The Invisible Hand Meets Creative Destruction," Federal Reserve Bank of Philadelphia *Business Review*, July/August 2000.
- Nickell, Stephen J. "Competition and Corporate Performance," *Journal of Political Economy* 104 (4) August 1996, pp. 724-46.
- Porter, Michael E. *The Competitive Advantage of Nations*. New York: Free Press, 1998.
- Preston, Richard. "The Genome Warrior," *The New Yorker*, June 12, 2000.
- Pugh, Emerson W. *Building IBM*. Cambridge, MA: MIT Press, 1995.
- Scherer, F.M. *Innovation and Growth: Schumpeterian Perspectives*. Cambridge, MA: MIT Press, 1986.
- Schumpeter, Joseph A. *Capitalism, Socialism, and Democracy*. New York: Harper and Row, 1942.
- Tomasello, Michael. *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press, 1999.