THE ANNUAL REPORT
OF THE
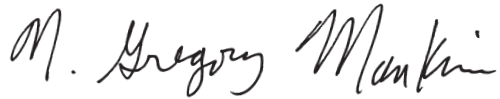## COUNCIL OF ECONOMIC ADVISERS

LETTER OF TRANSMITTAL

COUNCIL OF ECONOMIC ADVISERS,
Washington, D.C., January 30, 2004

MR. PRESIDENT:

The Council of Economic Advisers herewith submits its 2004 Annual Report in accordance with the provisions of the Employment Act of 1946 as amended by the Full Employment and Balanced Growth Act of 1978.

Sincerely,

N. Gregory Mankiw
Chairman

Kristin J. Forbes
Member

Harvey S. Rosen
Member

CONTENTS

<div align="center">LIST OF BOXES</div>

Overview

The U.S. economy made notable progress in 2003, propelled forward by pro-growth policies that led to a marked strengthening of activity in the second half of the year and put the United States on a path for higher sustained output growth in the years to come.

The recovery was still tenuous coming into 2003, as continued fallout from powerful contractionary forces—the capital overhang, corporate scandals, and uncertainty about future economic and geopolitical conditions—was offset by stimulus from expansionary monetary policy and the Administration's 2001 tax cut and 2002 fiscal package. The contractionary forces dissipated over the course of 2003, and the expansionary forces were augmented by the Jobs and Growth Tax Relief Reconciliation Act (JGTRRA) that was signed into law at the end of May.

The economy appears to have moved into a full-fledged recovery, with real gross domestic product (GDP), the most comprehensive measure of the output of the U.S. economy, expanding at an annual rate of more than 8 percent in the third quarter of the year. Based on data available through the middle of January, a further solid gain appears likely in the fourth quarter (the GDP estimate for the fourth quarter was released after this *Report* went to press). Job growth, however, began to pick up only late in 2003.

This *Report* discusses this turning of the macroeconomic tide, along with a number of other economic policy issues of continuing importance. The 14 chapters of this *Report* cover five broad topics: macroeconomic policy, fiscal policy, regulation, reforms of the health care and tort systems, and issues in international trade and finance. In all of these areas, the *Report* highlights how economics can inform the design of public policy and discusses Administration policies.

The Administration's pro-growth tax policy, in concert with the dynamism of the U.S. free-market economy, has laid the groundwork for sustainable rapid growth in the years ahead. Well-timed fiscal stimulus combined with expansionary monetary policy to offset and eventually reverse the contractionary forces impacting the economy. But there is still much to be done. The tax cuts must be made permanent to have their full beneficial impact on the economy. A stronger economy will also result from progress on the other aspects of the Administration's economic agenda, including making health care more affordable; reducing the burden of lawsuits on the economy; ensuring an affordable and reliable energy supply; streamlining regulations; and opening markets to international trade. These initiatives are discussed in this *Economic Report of the President*.

Macroeconomic Policy

Chapter 1, *Lessons from the Recent Business Cycle*, discusses the distinctive features of the recent recession and subsequent recovery, and draws five key lessons for the future. The recent business cycle was unusual in that it was characterized by especially weak business investment but robust consumption and housing investment. This makes clear the first lesson, that structural imbalances such as the "capital overhang" that developed in the late 1990s can take some time to resolve. A number of events contributed to a climate of uncertainty in 2003, including the terrorist attacks of September 11, 2001, corporate governance and accounting scandals, and geopolitical tensions surrounding the war with Iraq. The second lesson from the recent business cycle is that the effects of the uncertainty from these events on household and business confidence can have important effects on asset prices, household spending, and investment. Resolution of some of the uncertainties appears to have contributed to the resurgence of growth.

Monetary and fiscal policies played a critical role in moving the economy back toward potential. The third lesson is that aggressive monetary policy can help make a recession shorter and milder. The fourth lesson is that tax cuts can likewise boost economic activity. Tax cuts raise after-tax income, while at the same time promoting long-term growth by enhancing incentives to work, save, and invest. Tax relief enacted in 2001 and 2002 helped lessen the severity of the recession, while the 2003 tax cut appears to have propelled the economy forward into a strong recovery. Job creation has lagged behind, even as demand has surged. Thus, the fifth lesson of the recent recession is that strong productivity growth, as was experienced in 2003, means that much faster economic growth is needed to raise employment. This productivity growth, however, is not to be lamented, since it ultimately leads to higher standards of living for both workers and business owners.

Chapter 2, *The Manufacturing Sector*, examines recent developments and long-term trends in manufacturing and considers policy responses. Manufacturing was affected by the economic slowdown earlier, longer, and harder than other sectors of the economy and manufacturing employment losses have only recently begun to abate. The severity of the recent slowdown in manufacturing was largely due to prolonged weakness in business investment and exports, both of which are heavily tied to manufacturing.

Over the past several decades, the manufacturing sector has experienced substantial output growth, even while manufacturing employment has declined as a share of total employment. The manufacturing employment decline over the past half-century primarily reflects striking gains in productivity and increasing consumer demand for services compared to manufactured goods. International trade has played a relatively small role by

comparison. Consumers and businesses generally benefit from the lower prices made possible by increased manufacturing productivity, and strong productivity growth has led to real compensation growth for workers. While the shift of jobs from manufacturing to services has caused dislocation, it has not resulted, on balance, in a shift from "good jobs" to "bad jobs." The best policy response to recent developments in manufacturing is to focus on stimulating the overall economy and easing restrictions that impede manufacturing growth. This Administration has actively pursued such measures.

Chapter 3, *The Year in Review and the Years Ahead*, reviews macroeconomic developments in 2003 and discusses the Administration forecast for 2004 through 2009. Real GDP growth picked up appreciably in 2003, with growth in consumer spending, residential investment, and, particularly, business equipment and software investment increasing noticeably in the second half of the year. The labor market began to rebound in the final five months of 2003. Inflation remained well in check, with core consumer inflation declining by the end of the year to its lowest level in decades. The improvement in the economy over the course of the year stemmed largely from faster growth in household consumption, extraordinary gains in residential investment, and a sharp acceleration of investment in equipment and software by businesses. Payroll employment bottomed out in July and increased by 278,000 over the remainder of the year. Financial markets responded favorably to the strengthening of the economy, with the total value of the stock market rising more than $3 trillion, or 31 percent, over the course of 2003.

The Administration expects the economic recovery to strengthen further in 2004, with real GDP growth running well above its historical average and the unemployment rate falling. Boosted by pro-growth policies and expansionary monetary policy, and on the foundation of the underlying strength of the free-market society in the United States, the economy is expected to continue on a path of strong, sustainable growth.

Fiscal Policy

Chapter 4, *Tax Incidence: Who Bears the Tax Burden?*, discusses the analysis of how the burden of a tax is distributed among taxpayers. This question is important to policy makers, who want to know whether the distribution of the tax burden (between rich and poor, capital and labor, consumers and producers, and so on) meets their criteria for fairness. The key result is that the economic incidence of a tax may have little to do with the legal specification of its incidence. Rather, it depends on the actions of market participants in response to the imposition of the tax.

Distributional tables showing the tax burdens borne by different income groups are an important application of incidence analysis. When used properly, distributional tables can contribute to informed decision making on the part of citizens and policy makers. Unfortunately, mainstream economic analysis suggests that these tables do not always accurately describe who bears the burden of certain taxes. This problem does not arise from bias or lack of economic knowledge on the part of the economists who prepare these tables. Instead, it reflects resource and data limitations, uncertainty about some of the economic effects of taxes, and variations in the time frame considered by the analyses. Nevertheless, the shortcomings of distributional tables can lead to misperceptions of the impact of tax changes.

An important implication of the economic analysis of incidence is that, in the long run, a large part of the burden of capital taxes is likely to be shifted to workers through a reduction in wages. Analyses that fail to recognize this shift can be misleading, suggesting that lower income groups bear an unrealistically small share of the burden of such taxes and an unrealistically small share of the gain when capital income taxes are lowered.

Chapter 5, *Dynamic Revenue and Budget Estimation*, examines how taxes affect the behavior of firms, workers, and investors and discusses the implications for the estimated effects of a tax change on revenue. Changes in taxes and spending generally alter incentives for work, investment, and other productive activity—a higher tax on an activity tends to discourage that activity. Revenue estimation is called *dynamic* if it incorporates the behavioral responses to tax changes and *static* if it does not incorporate these behavioral responses.

To make informed decisions about a policy change, policy makers should be aware of all aspects of its budgetary implications. Currently, official revenue estimates of proposed tax changes incorporate the revenue effects of many microeconomic behavioral responses. However, these estimates are not fully dynamic because they exclude the effects of macroeconomic behavioral responses. Several obstacles have prevented macroeconomic behavioral responses from being incorporated in such estimates. This chapter discusses the ongoing efforts to provide a greater role for fully dynamic revenue and budget estimation in the analysis of major tax and spending proposals. At least in the near term, it may not be practical for macroeconomic effects to be incorporated in official estimates. But estimates of these effects should be provided as supplementary information for major tax and spending proposals. Dynamic estimation of policy changes should distinguish aggregate demand effects from aggregate supply effects, include long-run effects, apply to spending as well as tax changes, reflect the differing effects of various policy changes, account for the need to finance policy changes, and use a variety of models.

Reform of entitlement programs remains the most pressing fiscal policy issue confronting the Nation. Chapter 6, *Restoring Solvency to Social Security*, examines the largest entitlement program. Social Security is a *pay-as-you-go* system in which payroll taxes on the wages of current workers finance the benefits being paid to current retirees. While the program is running a small surplus at present, deficits are projected to appear in 15 years; by 2080, the Social Security deficit is projected to exceed 2.3 percent of GDP. These deficits are driven by two demographic shifts that have been underway for several decades: people are having fewer children and are living longer. The President has called for new initiatives to modernize Social Security to contain costs, expand choice, and make the program secure and financially viable for future generations of Americans.

This chapter assesses the need to strengthen Social Security in light of its long-term financial outlook. The most straightforward way to characterize the financial imbalance in entitlement programs such as Social Security is by considering their long-term annual deficits. Even after the baby-boom generation's effect is no longer felt, Social Security is projected to incur annual deficits greater than 50 percent of payroll tax revenues. These deficits are so large that they require a meaningful change to Social Security in future years. Reform should include moderation of the growth of benefits that are unfunded and would otherwise require higher taxes in the future. However, the benefits promised to those in or near retirement should be maintained in full. A new system of personal retirement accounts should be established to help pay future benefits. The economic rationale for under-taking this reform in an era of budget deficits is as compelling as it was in an era of budget surpluses.

Regulation

Chapter 7, *Government Regulation in a Free-Market Society*, discusses the role of the free market in providing for prosperity in the United States and considers situations in which government interventions such as regulations would be beneficial. An important reason for Americans' high standard of living is that they rely primarily on markets to allocate resources. The government enables the system to work by enforcing property rights and contracts. Typically, free markets allocate resources to their highest-valued uses, avoid waste, prevent shortages, and foster innovation. By providing a legal foundation for transactions, the government makes the market system reliable: it gives people certainty about what they can trade and keep, and it allows people to establish terms of trade that will be honored by both sellers and buyers. The absence of any one of these elements—competition,

enforceable property rights, or an ability to form mutually advantageous contracts—can result in inefficiency and lower living standards. In some cases, government intervention in a market, for example through regulation, can create gains for society by remedying shortcomings in the market's operation. Poorly designed or unnecessary regulations, however, can actually create new problems or make society worse off by damaging the elements of the market system that do work.

Chapter 8, *Regulating Energy Markets*, discusses economic issues relevant to several energy markets, including natural gas, gasoline, electricity, and crude oil. While energy markets generally function well, some parts of the energy industry have characteristics associated with market failures. These could stem from the large fixed costs required to construct distribution networks for electricity and natural gas that give rise to *market power* in the form of a natural monopoly. Alternatively, the market may not function well in the presence of *negative externalities*, such as when energy producers and consumers do not fully take into account the fact that burning fossil fuels may cause acid rain or smog.

Minimizing disruptions is an important consideration in the design of regulations to address shortcomings in energy markets. Federal, state, and local regulations can have conflicting goals. If the conflicting goals are not balanced, competing regulations could lead to worse problems than the market failures the regulations attempt to address. Moreover, regulations need to be updated as markets evolve over time to ensure that their original goals still apply and that these regulations are still the lowest-cost means of meeting those goals.

The chapter also examines global trade in energy products. The United States benefits from international trade in energy products because meeting all U.S. energy needs from domestic sources would require significant and costly changes to the U.S. economy, including changes in the types of transportation fuels used by Americans. But this leads to the possibility of occasional supply disruptions. An important consideration is that the price of oil is set in global markets, so that disruptions to the supply of oil from areas that do not supply the United States affect domestic prices of oil even if U.S. imports are not directly affected. Fortunately, changes in the U.S. economy over the past three decades and the increasing sophistication of financial markets have diminished the impact of supply disruptions and temporary price changes on the United States.

Finally, the chapter considers the role for government in subsidizing research and development into new energy sources. In general, policy makers should avoid forcing commercialization of new energy sources before market signals indicate that a shift is required. One potential problem with forcing this process is that technological breakthroughs may lead to

alternatives in the future that are hard to imagine today. Premature adoption of new technologies would raise energy costs before the need arises, causing society as a whole to spend more on energy than needed.

Chapter 9, *Protecting the Environment*, discusses market-oriented approaches to safeguarding and improving the environment. While the free-market system typically promotes efficiency and economic growth, the absence of property rights for environmental "goods" such as clean air and water can lead to negative externalities that reduce societal well-being. This problem can be addressed by establishing and enforcing property rights that will lead the interested parties to negotiate mutually beneficial outcomes in a market setting. If such negotiations are expensive, however, the government can design regulations that consider both the benefits of reducing the environmental externality as well as the costs of the regulations.

Regulations should be designed to achieve environmental goals at the lowest possible cost, promoting both environmental protection and continued economic growth. Indeed, economic growth can lead to increased demand for environmental improvements and provide the resources that make it possible to address environmental problems. Some policies aimed at improving the environment can entail substantial economic costs. Misguided policies might actually achieve less environmental progress than alternative policies for the same cost. Environmental risks should be evaluated using sound scientific methods to avoid possible distortions of regulatory priorities. Market-based regulations, such as the cap-and-trade programs promoted by the Administration to reduce common air pollutants, can achieve environmental goals at lower cost than inflexible command-and-control regulations.

Reforms of Health Care and the Legal System

Chapter 10, *Health Care and Insurance*, discusses the roles of innovation, insurance, and reform in the health care market. U.S. markets provide incentives to develop innovative health care products and services that benefit both Americans and the global community. The breadth and pace of innovation in the provision of health care in the United States over the past few decades have been astounding. New treatment options, however, have also been associated with higher costs and concerns about affordability. Research suggests that between 50 and 75 percent of the growth in health expenditures in the United States is attributable to technological progress in health care goods and services. A strong reliance on market mechanisms will ensure that incentives for innovation are maintained while providing high-quality care in the most cost-efficient manner.

Health insurance plays a central role in the workings of the U.S. health care market. An understanding of the strengths and weaknesses of health insurance as a payment mechanism for health care is essential to the design of reforms that retain incentives for innovation while reining in unnecessary expenditures. Over-reliance on health insurance as a payment mechanism leads to an inefficient use of resources in providing and utilizing health care. Reforms should provide consumers and health care providers with more flexibility, more choices, more information, and more control over their health care decisions.

Chapter 11, *The Tort System*, discusses the role of the U.S. tort system and the considerable burden it imposes on the U.S. economy. The tort system is intended to compensate accident victims and to deter potential defendants from putting others at risk. Empirical evidence, however, is mixed on whether the tort system effectively deters negligent behavior. Moreover, the tort system is a costly method of providing insurance against a limited number of injuries. Research suggests that tort liability also leads to lower spending on research and development, higher health care costs, and job losses.

Ways to reduce the burden of the tort system include limits on noneconomic damages, class action reforms, trust funds for payments to victims such as in asbestos, and allowing parties to avoid the tort system contractually. The Administration has proposed a number of reforms to reduce the burden of the tort system while ensuring that people with legitimate claims can recover damages.

International Trade and Finance

Chapter 12, *International Trade and Cooperation*, discusses how growing trade helps to spur U.S. and global growth. Since the end of the Second World War, international trade has grown steadily relative to overall economic activity. Over time, countries that have been more open to international flows of goods, services, and capital have grown faster than countries that were less open to the global economy. The United States has been a driving force in constructing an open global trading system. The Administration has pursued, and will continue to pursue, an ambitious agenda of trade liberalization through negotiations at the global, regional, and bilateral levels.

New types of trade deliver new benefits to consumers and firms in open economies. Growing international demand for goods such as movies, pharmaceuticals, and recordings offers new opportunities for U.S. exporters. A burgeoning trade in services provides an important outlet for U.S. expertise in sectors such as banking, engineering, and higher education. The ability to

buy less expensive goods and services from new producers has made house-hold budgets go further, while the ability of firms to distribute their production around the world has cut costs and thus prices to consumers. The benefits from new forms of trade, such as in services, are no different from the benefits from traditional trade in goods. Outsourcing of professional services is a prominent example of a new type of trade. The gains from trade that take place over the Internet or telephone lines are no different than the gains from trade in physical goods transported by ship or plane. When a good or service is produced at lower cost in another country, it makes sense to import it rather than to produce it domestically. This allows the United States to devote its resources to more productive purposes.

Although openness to trade provides substantial benefits to nations as a whole, foreign competition can require adjustment on the part of some individuals, businesses, and industries. To help workers adversely affected by trade develop the skills needed for new jobs, the Administration has worked hard to build upon and develop programs to assist workers and communities that are negatively affected by trade.

The Administration has also worked to strengthen and extend the global trading system. International cooperation is essential to realizing the potential gains from trade. International trade agreements have reduced barriers to international commerce, and contributed to the gains from trade. A system through which countries can resolve disputes can play an important role in realizing these gains.

Chapter 13, *International Capital Flows*, discusses the economic benefits and risks associated with the transfer of financial assets, such as cash, stocks, and bonds, across international borders. Capital flows have become an increasingly significant part of the world economy over the past decade, and an important source of funds to support investment in the United States. Around $2 trillion of capital flowed into all countries in the world in 2002, with around $700 billion flowing into just the United States. Different types of capital flows—such as foreign direct investment, portfolio investment, and bank lending—are driven by different investor motivations and country characteristics. Countries that permit free capital flows must choose between the stability provided by fixed exchange rates and the flexibility afforded by an independent monetary policy.

Capital flows can have a number of benefits for economies around the world. For example, foreign direct investment can facilitate the transfer of technology, allow for the development of markets and products, and improve a country's infrastructure. Portfolio flows can reduce the cost of capital, improve competitiveness, and increase investment opportunities. Bank flows can strengthen domestic financial institutions, improve financial intermediation, and reduce vulnerability to crises.

A series of financial crises in emerging market economies, however, has raised some concerns that financial liberalization can also involve risks. In countries with weak institutions, poorly regulated banking systems, or high levels of corruption, capital inflows may not be channeled to their most productive uses. One approach to limiting the risks from capital flows when legal and financial institutions are poorly developed is to restrict foreign capital inflows. Experience suggests, however, that capital controls impose substantial, and often unexpected, costs. Instead, countries are more likely to benefit from free capital flows and minimize any related risks, if they adopt prudent fiscal and monetary policies, strengthen financial and corporate institutions, and develop sound regulations and supervisory agencies. The Administration has promoted policies to help countries reap the benefits from the free flow of international capital.

Chapter 14, *The Link Between Trade and Capital Flows*, shows that trade flows and capital flows are inherently intertwined. Changes in a country's net international trade in goods and services, captured by the current account, must be reflected in equal and opposite changes in its net capital flows with the rest of the world. The large net inflow of foreign capital experienced by the United States in recent years has funded more investment than could be supported by U.S. national saving. Corresponding to these inflows is the large U.S. current account deficit. These patterns reflect fundamental economic forces, notably strong growth in the United States that has made investment in this country attractive compared to opportunities in other countries.

An adjustment of the U.S. current account deficit could come about in several ways. Faster growth in other countries relative to the United States could increase demand for U.S. net exports. Trade flows could also adjust through changes in the relative prices of U.S. goods and services compared to the prices of foreign goods and services. Any narrowing of the U.S. current account deficit would also require reduced net capital inflows into the United States. This might occur if U.S. national saving increased, reducing the need for foreign funds to finance U.S. domestic investment, or if U.S. investment declined, so that the United States required less capital inflows. Lower investment is the least desirable form of balance of payments adjustment, however, as it could slow the expansion of U.S. productive capacity and reduce economic growth.

It is impossible to predict the exact timing or magnitude of any adjustment in the U.S. current account balance. After a large increase in the U.S. current account deficit in the 1980s, the ensuing adjustments were gradual and benign. Public policies can facilitate smooth changes in the U.S. current account and net capital flows by creating a stable macroeconomic and financial environment, promoting growth abroad, and encouraging greater saving in the United States.

Conclusion

The future of the U.S. economy is bright. This is a testament to the institutions and policies that have unleashed the creativity of the American people and their spirit of entrepreneurship. History teaches that the forces of free markets are the bedrock of economic prosperity.

In 1776, as the Founding Fathers signed the Declaration of Independence, the great economist Adam Smith wrote: "Little else is requisite to carry a state to the highest degree of opulence from the lowest barbarism but peace, easy taxes, and a tolerable administration of justice: all the rest being brought about by the natural course of things." The economic analysis presented in this *Report* builds on the ideas of Smith and his intellectual descendants by discussing the role of the government in creating an environment that promotes and sustains economic growth.

CHAPTER 1
Lessons from the Recent Business Cycle

Economic conditions in the United States improved substantially during 2003, with real gross domestic product (GDP), the most comprehensive measure of the output of the U.S. economy, expanding at an annual rate of more than 8 percent in the third quarter of the year. Based on data available through the middle of January, a further solid gain appears likely in the fourth quarter (the GDP estimate for the fourth quarter was released after this *Report* went to press). The improvement in the economy over the course of the year stemmed largely from faster growth in household consumption, extraordinary gains in residential investment, and a sharp acceleration of investment in equipment and software by businesses. Payroll employment bottomed out in July and increased 278,000 over the remainder of the year. Financial markets responded favorably to the strengthening of the economy, with the total value of the stock market rising more than $3 trillion, or 31 percent, over the course of 2003.

Despite this improvement, the U.S. economy has further to go to make up for the weakness that began showing even before the economy slipped into recession roughly three years ago. Until recently, the recovery has been slow and uneven. Employment has lagged behind gains in other areas. Strong fiscal policy actions by this Administration and the Congress, together with the Federal Reserve's stimulative monetary policy, have softened the impact of the recession and have also put the economy on an upward trajectory. The Administration's pro-growth tax policy, in particular, has laid the groundwork for sustainable rapid growth in the years ahead.

This chapter discusses the distinctive features of the recent recession and recovery, and it draws lessons for the future. The key points in this chapter are:
- Structural imbalances, such as the "capital overhang" that developed in the late 1990s, can take some time to resolve.
- Uncertainty matters for economic decisions, and was likely a factor weighing on investment in recent years.
- Aggressive monetary policy can reduce the depth of a recession.
- Tax cuts can boost economic activity by raising after-tax income and enhancing incentives to work, save, and invest.
- Strong productivity growth raises standards of living but means that much faster economic growth is needed to raise employment.

Overview of the Recent Business Cycle

The recent recession and recovery mark the seventh business cycle in the U.S. economy since 1960. This cycle shares some common features with previous business cycles. According to the National Bureau of Economic Research (NBER), the unofficial arbiter of U.S. business cycles, a *recession* is "a period of falling economic activity spread across the economy, lasting more than a few months, normally visible in real GDP, real income, employment, industrial production, and wholesale-retail sales." The recent recession, like others, has involved a downturn in economic activity of sufficient depth, duration, and breadth to be judged a recession by the NBER.

The NBER also identifies the *peaks* and *troughs* of economic activity that mark when recessions begin and end. In November 2001, the NBER determined that the economy had peaked in March 2001. However, revisions to economic data since the NBER's initial decision suggest that the peak in activity was actually months earlier (Box 1-1). In July 2003, the NBER determined that the economy had reached a trough in November 2001.

Despite the similarities between the recent business cycle and previous ones, this most recent cycle was distinctive in important and instructive ways. One noteworthy difference is that real GDP fell much less in this recession than has been typical. Chart 1-1 shows the path of real GDP over the past several years compared with the average path of the six prior recessions, with the level of real GDP at the economy's peak set equal to 100 in each case. (All of the charts in this *Report* assume that the peak for the recent recession was in the fourth quarter of 2000.) The chart shows that the decline in real GDP in the recent recession was smaller than the historical average; indeed, it was the second smallest in any recession since 1960.

Box 1-1: When Did the Recent Recession Begin?

The National Bureau of Economic Research (NBER) uses a variety of economic data to determine the dates of business-cycle peaks and troughs. This task is made more difficult because many of these data series are subject to revision. For example, on November 26, 2001, the NBER announced that a recession had begun in March 2001. Since then, the four data series that the NBER used to determine the timing of the recession have been revised. The revisions to these series suggest that the recent recession began earlier than March 2001.

The four series cited by the NBER in their decision about the recent business-cycle peak were revised as follows:

Box 1-1 — *continued*

- *Real personal income less transfers:* When the NBER dated the recession, this series showed a generally steady rise throughout 2000 and early 2001. Subsequent revisions reveal that income peaked in October 2000.
- *Nonfarm payroll employment:* The data at the time of the recession announcement showed employment growing at a substantial pace in early 2001, with 287,000 jobs added from December 2000 to its peak in March 2001. Revised data show that employment grew less than one-third of this amount in early 2001 and peaked in February 2001.
- *Industrial production:* The original data used by the NBER showed that this series peaked in September 2000. Revised data show that this peak came even earlier, in June 2000.
- *Manufacturing and trade sales:* Original data showed a peak in August 2000; the most recent data show a peak in June 2000.

Thus, the revised data show that the *latest* peak among the four series was February 2001, with some series peaking considerably earlier. Moreover, another data series, which the NBER has recently announced it will incorporate into its business-cycle dating process, also shows a peak before March 2001: *monthly GDP* reached a high point in February 2001, according to the most recently available estimates computed by a private economic consulting firm.

While some arbitrariness in determining the date on which a recession began is inevitable, revisions since the NBER made its decision for the most recent recession strongly suggest that the business-cycle peak was before March 2001. The median date of the peak for the five series discussed here is October 2000. Other data support the notion that economic activity had slowed sharply or even begun to decline by this point, including the stock market, business investment, and initial unemployment claims. For these reasons, the analyses throughout this chapter (including the charts that compare this recession to past recessions) use the fourth quarter of 2000 as the peak of economic activity and the start of the recession.

In October 2003, the NBER announced that it would defer consideration of whether the latest business-cycle peak should be revised until the results of the coming comprehensive revision of the National Income and Product Accounts were released. The major results of this revision were announced in December 2003, but the monthly manufacturing and trade sales data and some of the detail needed to estimate monthly GDP had not been released at the time this *Report* went to press.

Chart 1-1 Real GDP
Real GDP fell less in the recent recession than it typically has.

Index, level at business cycle peak = 100
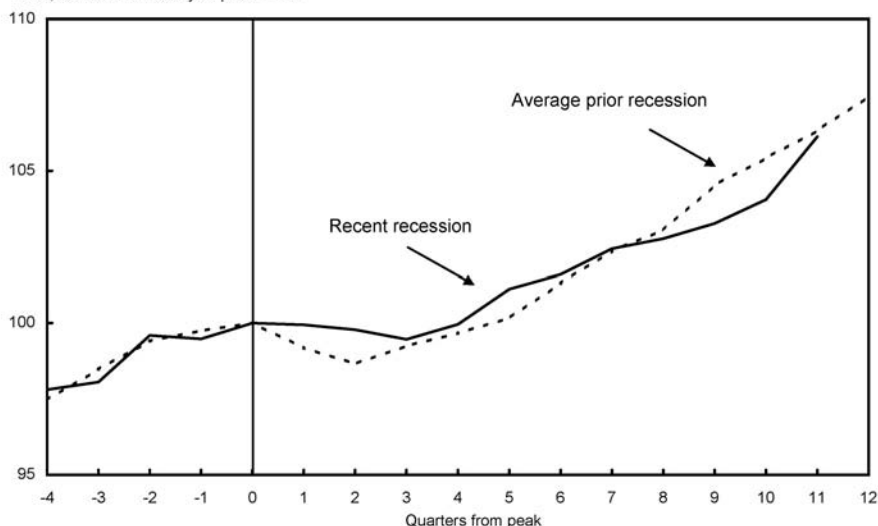
Note: Recent peak set by Council of Economic Advisers at 2000:Q4. Average based on prior recessions since 1960.
Source: Department of Commerce (Bureau of Economic Analysis).

This relatively mild decline in output can be attributed to unusually resilient household spending. Consumer spending on goods and services held up well throughout the slowdown, and investment in housing increased at a fairly steady pace rather than declining as has been typical in past recessions. In contrast, business investment in capital equipment and structures has been quite soft in this cycle. As discussed below, business spending during the past few years has likely been held down by overinvestment in the late 1990s, as well as by heightened business caution owing to terrorism and corporate scandals. As a result of these forces, investment weakened sooner and has recovered more slowly than in the typical cycle.

Another distinguishing feature of this cycle has been the weakness in labor markets relative to output. In particular, the recovery in employment—although now under way—lagged the upturn in output by a much longer period than in prior recessions. This difference was associated with unusually large productivity gains.

The balance of this chapter draws five distinctive lessons from the recent business cycle in the United States. Chapter 3, *The Year in Review and the Years Ahead,* presents details about developments over the past year and discusses the Administration's forecast.

Lesson 1: Structural Imbalances
Can Take Some Time to Resolve

Business investment in equipment and software surged in the late 1990s. Real investment increased at an average annual rate of roughly 13 percent between the fourth quarter of 1994 and the fourth quarter of 1999, compared with an average annual rate of less than 7 percent over the preceding three decades. The surge in investment was led by purchases of high-tech capital goods—computers, software, and communications equipment—which increased at an average annual rate of 20 percent over the period.

Economic theory implies that businesses invest when they believe that there are profits to be made from that investment. In the late 1990s, several developments fed a perception that the expected future return from newly installed capital would be considerably greater than the cost of this capital. Rapid advances in technology had lowered the price of high-tech capital goods dramatically throughout the 1990s and especially in the second half of the decade. For example, the quality-adjusted price index for business computers and peripheral equipment fell at an average annual rate of 22 percent between late 1994 and late 1999. In addition, rapidly growing demand for business output led firms to believe that newly installed capital would be used productively, boosting the expected return to investment.

Moreover, technological progress and legislation provided incentives for strong investment in high-tech equipment. The development of the World Wide Web enabled new and established firms to enter e-commerce, and rapidly increasing household and business access to the Internet provided a large base of potential customers for these firms. The Telecommunications Act of 1996 provided for substantial deregulation of the telecommunications industry and may have spurred investment in that sector. In addition, concern that some computer systems might be inoperable after December 1999 caused a wave of so-called Y2K-related investment. Some analysis indicates that Y2K spending alone boosted the growth rate of real equipment and software investment by more than 3½ percentage points per year in the latter part of the 1990s.

Optimism about the potential gains from new capital, and from high-tech capital in particular, was reflected not only in investment decisions but also in a sharp rise in stock prices. From late 1994 to late 1999, the Wilshire 5000—a broad index of U.S. stock prices—nearly tripled. The Nasdaq stock price index, which is heavily weighted toward high-tech industries, registered an even more dramatic ascent, increasing more than fourfold over this period. The increase in stock prices stimulated investment by reducing the cost of equity capital. In addition, the rise in stock prices fueled a consumption boom by boosting the wealth of a growing number of Americans and more

generally signaling better future economic conditions. This consumption boom encouraged further business investment.

In mid-2000, business equipment investment abruptly slowed. After rising at an annual rate of 15 percent in the first half of the year, real spending on business equipment and software inched up at about a ¼ percent annual rate in the second half. The slowdown in high-tech equipment investment was especially dramatic. For example, real outlays for computers had skyrocketed at an annual rate of 40 percent in the first half of the year, but grew at less than one-quarter of that pace in the second half. This stalling of investment preceded the downturn in the overall economy; by contrast, in the typical business cycle, investment has turned down at the same time as overall economic activity (Chart 1-2). The unusual timing of the investment slow-down in this recession is the reason that the recent business cycle has been widely viewed as an "investment-led" recession.

The sharp break in investment occurred in parallel with an apparent reevaluation of future corporate profitability among financial market partic-ipants. By the end of 2000, the Wilshire 5000 index of stock prices was down 13 percent from its peak, and analysts had substantially marked down their forecasts for S&P 500 earnings over the coming year. The movements were even more dramatic in the high-tech sector. The Nasdaq index of stock

Chart 1-2 **Real Investment in Equipment and Software**
Relative to the average prior recession, the weakness in investment in the recent recession occurred earlier, was more pronounced, and persisted longer.

Index, level at business cycle peak = 100
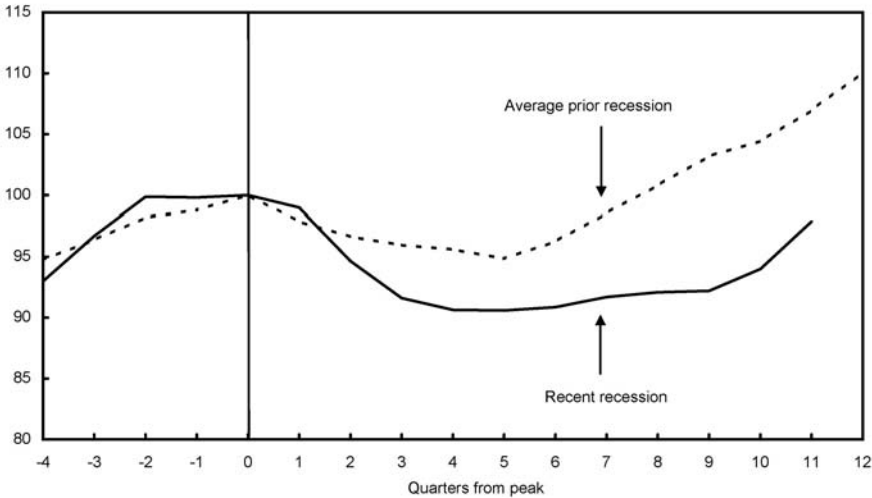
Quarters from peak

Note: Recent peak set by Council of Economic Advisers at 2000:Q4. Average based on prior recessions since 1960.
Source: Department of Commerce (Bureau of Economic Analysis).

prices dropped nearly 50 percent from its peak in March 2000 to the end of the year. The prices of technology, telecommunications, and Internet shares fell particularly sharply, along with near-term earnings estimates. The elevated valuations of many such companies also declined markedly. Indeed, the price-earnings ratio (where "earnings" are those expected over the next year) for the technology component of the S&P 500 fell from a peak of more than 50 in early 2000 to less than 35 by the end of the year.

These facts and considerable anecdotal evidence suggest that business managers and investors sharply revised downward the expected gains from new capital investment during this period. One factor that may have contributed to the downward revision is a possible slowing of the pace of technological advance—the rate at which computer prices were declining eased (from more than 20 percent in the late 1990s to about half that in 2000), and the software industry reportedly developed no new so-called "killer applications" that required or spurred purchases of new hardware. In addition, firms may have been disappointed by the response of households to e-commerce opportunities and to new communications technologies such as broadband. Finally, previous investments had not uniformly translated into higher profitability, perhaps because the true potential of new forms of capital could be realized only by changing other aspects of production processes. For example, new computer systems designed to lower inventory management costs might have required an expensive reconfiguration of warehouses.

This reassessment of the gains from capital investment also implied that existing stocks of some types of equipment exceeded the amount of equipment that firms could put to profitable use. Such an excess of the existing capital stock relative to the desired stock (often called a *capital overhang*) is one type of *structural imbalance* that can slow or reverse economic expansion. In the case of an excess supply of capital, investment would be expected to slow until the capital overhang dissipates through a combination of depreciation in the existing stock and an increase in the desired stock due to lower costs of capital or stronger final demand.

Resolving the structural imbalance that developed in the late 1990s took considerable time. Real business spending on equipment and software dropped more than 9 percent during the four quarters of 2001 and posted less than a 2 percent gain during the four quarters of 2002. The high-tech categories showed especially sharp breaks in their upward trends. In these categories, the effects of the capital overhang were likely exacerbated by a reduction in normal replacement demand following the Y2K-related investment spurt. The prolonged period of sluggishness in business investment is another distinctive feature of this business cycle. Real investment in equipment and software typically has fallen less and has recovered more quickly than it did in the current recession and recovery (Chart 1-2).

Chart 1-3 Real Exports

Real exports have also been unusually weak in recent years relative to the average prior recession.
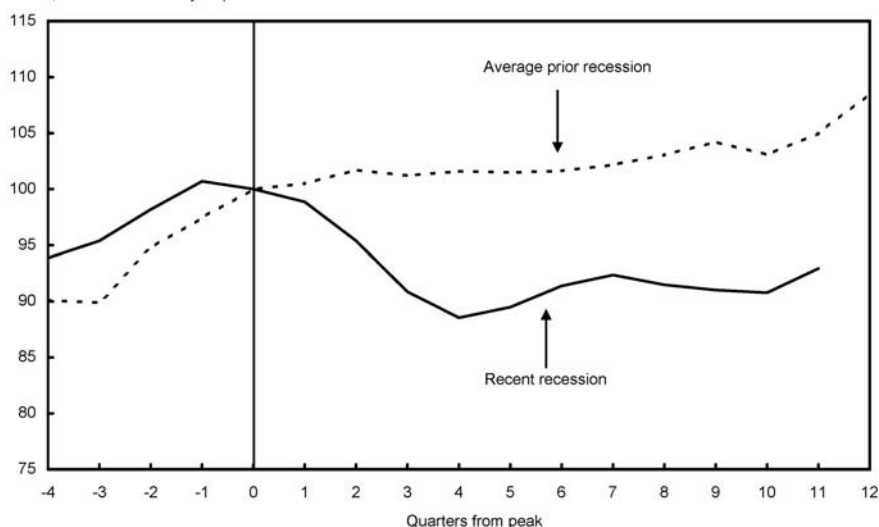
Index, level at business cycle peak = 100

Quarters from peak

Note: Recent peak set by Council of Economic Advisers at 2000:Q4. Average based on prior recessions since 1960.
Source: Department of Commerce (Bureau of Economic Analysis).

A similar structural adjustment appears to have taken place overseas, where investment demand was also weak. The global slowdown in investment hampered U.S. export growth, since capital goods traditionally account for about one-third of the value of U.S. exports. Real exports fell sharply in this recession and have recovered only a little of their lost ground. In past recessions, exports have typically leveled off but not declined (Chart 1-3). Soft investment and weak export demand led to a long period of weakness in manufacturing output, a topic discussed in the next chapter.

Several forces have more recently moved existing capital stocks into better alignment with desired stocks and thereby set the stage for a renewal of robust investment demand. Previously installed capital has depreciated, a process that occurs especially quickly for many types of high-tech equipment. Rising demand for business output and falling costs for high-tech capital (caused by ongoing technological progress) have increased firms' desired capital stocks. The elimination of capital overhangs, together with improved business confidence and reductions in tax rates on capital income discussed later in this chapter, are consistent with the marked upturn in business investment spending in the second half of 2003.

Lesson 2: Uncertainty Matters
for Economic Decisions

The U.S. economy has been hit hard in the past few years by a number of unexpected developments, including the tragic terrorist attacks of September 11, 2001, the corporate governance and accounting scandals of 2002, and the geopolitical tensions surrounding the war with Iraq in 2003. In addition to having direct effects on the economy, each of these events contributed to a climate of uncertainty that weighed on household and business confidence and thereby affected spending decisions.

The terrorist attacks have had substantial consequences for many aspects of the U.S. economy. The heightened focus on security at home, together with the determined efforts against terrorism around the world, have required increases in some types of government spending. The attacks hurt some industries directly: for example, fear of new attacks and the inconveniences associated with heightened airport security reduced air travel and tourism. Beyond these direct economic effects, the unprecedented attacks on the United States also generated uncertainty about future economic conditions.

Another setback for the economy was the series of revelations during 2002 regarding incomplete or misleading corporate financial reporting and, in some cases, wrongful conduct by corporate management. The number of financial restatements—that is, corrections to previous statements of earnings—by U.S. public corporations reached a record high in 2002. Although most of the restatements were not linked to misconduct, they raised questions about the reliability of accounting practices and the credibility of corporate financial disclosures. The combination of these concerns and allegations of misconduct by high-profile executives heightened investors' uncertainty about the quality of corporate governance and the reliability of earnings reports and projections.

In early 2003, uncertainty about the economic outlook increased during the period leading up to the war with Iraq. One source of this uncertainty was the potential effect of the conflict on the capacity for producing and transporting oil in the Persian Gulf, and thus on the future supply and price of oil. Observers were also concerned about the amount of additional government spending that would be needed to finance military operations and subsequent reconstruction, as well as the danger of retaliatory terrorist attacks on the United States. Finally, consumer confidence fell sharply in early 2003, raising concerns that the consumer demand that had supported the economy over the previous couple of years might falter. Such concerns were plausible, given that the 1990 Gulf War roughly coincided with a marked drop in consumer confidence and the start of the 1990-1991 recession.

Chart 1-4 The Wilshire 5000 Index of Stock Prices
A broad measure of stock prices moved down after the terrorist attacks of September 2001, during the period of revelations of corporate misreporting, and before the war with Iraq.

Index, January 1980 = 1,078.29

Note: The data are daily and extend through December 31, 2003.
Source: Wilshire Associates.

The uncertainties created by the three developments described above had significant effects on financial markets. Stock prices dipped noticeably in September 2001, recovered subsequently, but moved down during the summer of 2002 and fell again in early 2003 (Chart 1-4). *Risk spreads* (the difference between interest rates on corporate bonds and on comparable Treasury bonds) jumped temporarily after the terrorist attacks and rose again in late 2002 during the peak of concerns about corporate governance. Because risk spreads generally reflect the extra return investors require to hold riskier corporate assets, the rise in spreads in 2002 indicated investors' greater perceived probability of default, lesser willingness to take on risk, or both. Investor uncertainty also was reflected in measures of the expected volatility of stock prices based on option prices, which were elevated during each of the episodes noted above (Chart 1-5).

Reductions in share prices and increases in bond yields raised the cost of funding capital expenditures and thus directly discouraged business investment. Increased uncertainty likely also had *direct* effects on business decisions about investment and hiring: uncertainty may cause firms to wait until they have more information before committing to an investment. In this case, firm managers hesitate to respond to a change in demand. Anecdotal evidence from the past few years as well as some statistical analyses

Chart 1-5 Expected Near-Term S&P 500 Volatility

Expected stock price volatility was elevated after the terrorist attacks of September 2001, during the period of revelations of corporate misreporting, and before the war with Iraq.

Index
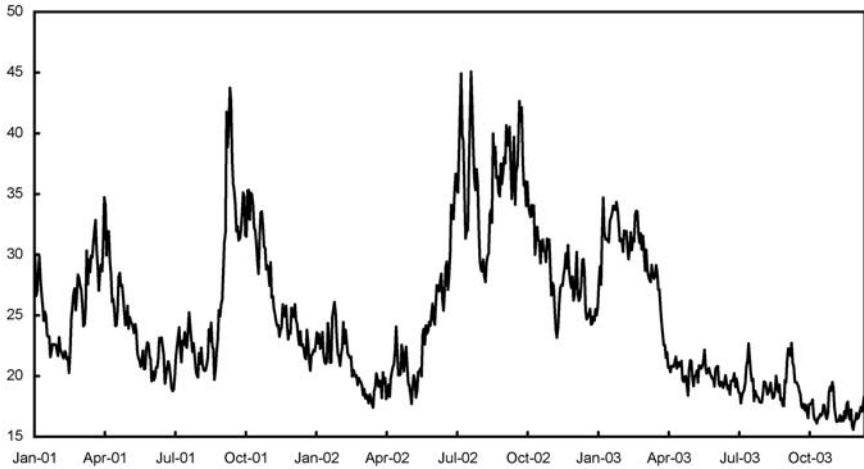
Note: The data are daily and extend through December 31, 2003.
Source: Chicago Board Options Exchange.

suggest that uncertainty has a noticeable damping effect on investment. Anecdotal evidence also suggests that uncertainty has held back hiring in the past few years.

Household spending may also have been affected by uncertainty. Economic theory and empirical evidence suggest that greater uncertainty about future economic conditions may lead households to raise saving and reduce spending. However, such effects are not immediately apparent in the recent cyclical downturn—as will be explained shortly, household spending has shown remarkable resiliency over the past few years. A possible explanation for the seeming discrepancy between this pattern and empirical work based on earlier data is that the negative effects of greater uncertainty were offset by lower taxes and the effects of lower interest rates.

While the uncertainty created by these unexpected developments has hampered the economic recovery, household and business confidence strengthened considerably during the second half of 2003. This Administration and the Congress moved swiftly to address problems with corporate governance. In March 2002, the President proposed a set of reforms aimed at a wide range of corporate governance issues, and in July 2002, Congress passed the landmark Sarbanes-Oxley Act. As concerns about corporate governance have abated, and the durability of the recovery has become more apparent, firms have begun to invest and hire.

Lesson 3: Aggressive Monetary Policy Can Reduce the Depth of a Recession

When the economy showed signs of weakening three years ago, the Federal Reserve moved decisively to reduce interest rates to stimulate the economy. During 2001, the Federal Reserve cut the Federal funds rate eleven times for a total reduction of 4¾ percentage points. When the economy failed to gain much forward momentum, the Federal Reserve reduced the funds rate another ½ percentage point in November 2002 and a further ¼ percentage point last June, to 1 percent. The decline in the Federal funds rate in this economic downturn was larger and occurred more rapidly than in previous downturns (Chart 1-6). One factor that likely contributed to the Federal Reserve's willingness to cut the funds rate so sharply was the low level of inflation. Core consumer price inflation, as measured by the 12-month change in the consumer price index excluding food and energy, was around 2¾ percent in early 2001 and fell to just over 1 percent by late last year. Thus, the Federal Reserve was able to lower the Federal funds rate and keep it low with little apparent risk of triggering an undesirably high inflation rate.

Chart 1-6 The Effective Federal Funds Rate
The decline in the Federal funds rate in the recent recession was larger and occurred earlier than in the average prior recession.
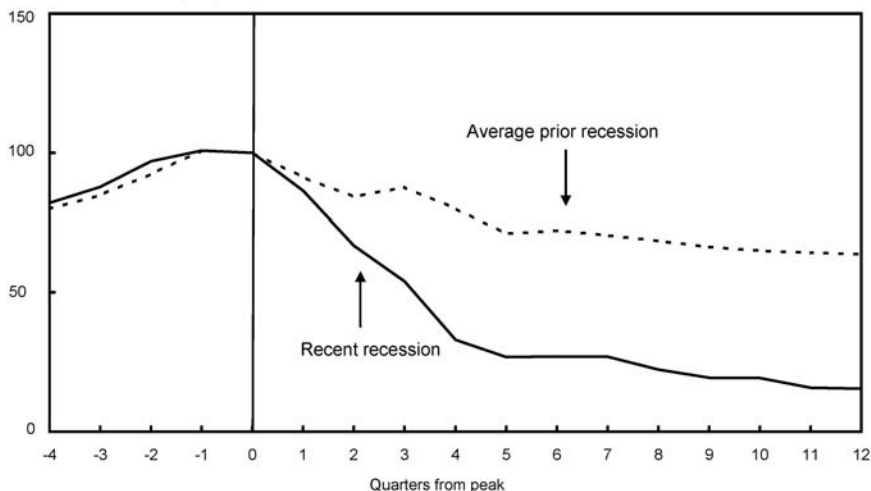
Index, level at business cycle peak = 100

Note: Recent peak set by Council of Economic Advisers at 2000:Q4. Average based on prior recessions since 1960.
Source: Board of Governors of the Federal Reserve System.

Long-term interest rates on government securities and high-grade corporate securities began falling in late 2000, likely in part reflecting an anticipated decline in the Federal funds rate in response to a weaker economic outlook. Throughout 2001, short-term and medium-term interest rates declined along with the Federal funds rate. However, long-term rates changed little, on net, because market participants apparently expected the downturn to be short-lived and believed that the Federal Reserve would soon begin raising the funds rate. Then, in 2002, persistently weak economic conditions, combined with the Federal Reserve's decisions to hold the funds rate steady for much of the year and cut it further in November, persuaded market participants that short-term rates were likely to stay low for some time. As a result, long-term rates fell substantially, on balance, in 2002. Long-term rates fluctuated in 2003, but finished the year a little above where they started.

Interest rates on fixed-rate mortgages tracked long-term government yields over this period, as they typically have. In 2003, the interest rate on 30-year fixed-rate mortgages averaged more than 2 percentage points below the average in 2000. Low and falling mortgage rates have provided strong support for housing demand over the past few years. Indeed, residential investment has increased at a fairly steady pace throughout the period of overall economic weakness—a stark contrast to the pattern in past recessions, when residential investment tended to fall sharply (Chart 1-7).

Chart 1-7 Real Residential Investment
Real residential investment has steadily increased over the past few years, in stark contrast to the considerable decline seen in the average prior recession.
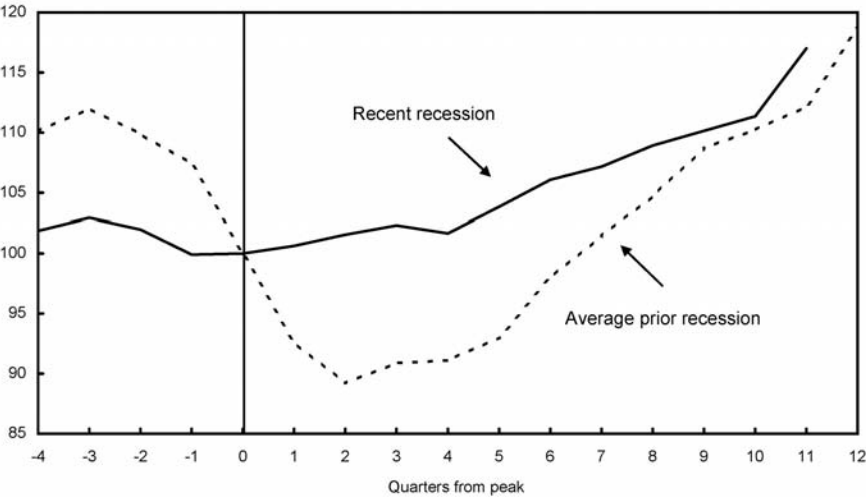
Index, level at business cycle peak = 100

Quarters from peak

Note: Recent peak set by Council of Economic Advisers at 2000:Q4. Average based on prior recessions since 1960.
Source: Department of Commerce (Bureau of Economic Analysis).

Declining mortgage interest rates have also fueled an enormous wave of mortgage refinancing. (The response has been particularly strong because technological and institutional advances in mortgage markets have reduced the costs of such transactions.) In many refinancing transactions, home-owners have "cashed out" some of their accumulated home equity by taking out new mortgages that are larger than the remaining balance on their previous mortgages. According to a survey of households, more than half of the liquefied equity funded either home renovations or household consumption and thus may have helped to sustain aggregate demand. Another substantial portion reportedly was used to pay down credit card debt, which generally carries a higher interest rate than mortgage debt and, unlike mortgage debt, is not tax-deductible. By moving from a high-cost form of debt to a lower-cost one, households have been better able to cope with their debt burdens. In particular, the transition has held down the fraction of their income committed to regular debt service payments, and thus has increased the amount of income available for spending on discretionary items.

Low long-term interest rates have also reduced the cost of funds to businesses. In some cases, this lower cost has been passed directly to households. For example, motor vehicle manufacturers made low-interest-rate loans available to car buyers in late 2001 and have generally maintained a high level of financing incentives since then. These incentives have bolstered consumer outlays for motor vehicles.

More generally, lower interest rates make it cheaper for firms to finance new investment projects. The aggressive easing of monetary policy since early 2001 has likely helped to support business investment, even though the forces discussed earlier have, on balance, caused investment to be weak.

Firms have also taken advantage of low long-term interest rates to restructure their balance sheets. Net issuance of commercial paper and net borrowing from banks were both negative in each of the past three years, while net bond issuance was strong. By issuing longer-term bonds and paying down short-term debt, businesses have substantially lengthened the overall maturity of their debt. This restructuring reduced firms' near-term repayment obligations and locked in low rates for longer periods. The strengthening of businesses' financial positions means that financial constraints are less likely to restrain a further pickup in hiring and investment.

Lesson 4: Tax Cuts Can Boost Economic Activity by Raising After-Tax Income and Enhancing Incentives to Work, Save, and Invest

The use of *discretionary fiscal policy*—explicit changes in taxes and government spending, as opposed to those that occur automatically as economic activity changes—to reduce cyclical fluctuations in the economy has fallen out of favor with many economists over the past several decades. Some have pointed to the difficulties of crafting and implementing discretionary policy quickly enough to provide stimulus while the economy is still weak rather than accentuating an upturn that is already under way. It has also been noted that a temporary reduction in taxes might be mostly saved by households and thus encourage relatively little additional spending. Moreover, some have argued that expansionary fiscal policy can push up interest rates and thereby "crowd out" interest-sensitive spending. All told, before the recent business cycle, many economists believed that monetary policy made the use of discretionary fiscal policy unnecessary to stabilize the economy.

The experience of the past three years, however, shows that well-designed and well-timed tax cuts are a useful complement to expansionary monetary policy. Over this period, three bills have made significant changes to the personal and corporate tax systems. The President came into office with proposals for permanently reducing taxes on work and saving. With the budget surplus having reached its highest level relative to GDP in half a century, the proposals were aimed predominantly at reducing tax-based impediments to long-term growth. The proposals resulted in the Economic Growth and Tax Relief Reconciliation Act (EGTRRA), which the President signed into law in June 2001. In the wake of the terrorist attacks of September 2001 and continuing softness in the economy, the Congress passed the Job Creation and Worker Assistance Act (JCWAA), which the President signed into law in March 2002. And, in early 2003, with the pace of economic growth still falling below its potential and the labor market lagging behind, the President proposed and the Congress enacted the Jobs and Growth Tax Relief Reconciliation Act (JGTRRA), which the President signed into law in May.

These three bills provided substantial short-term stimulus to economic activity and helped put the economy on the road to recovery. One source of stimulus has been the large boost to after-tax personal income stemming from lower marginal tax rates, a larger child tax credit, reduced tax rates on dividends and capital gains, and other changes in the tax law. Real after-tax income has increased much more than before-tax income over the past three

Chart 1-8 Growth in Personal Income, Before and After Taxes
Real after-tax income has increased much more than before-tax income in recent years.

Percent change in annual average

Note: Average for 2003 based on data through November. Before-tax personal income deflated by the price index for personal consumption expenditures.
Sources: Department of Commerce (Bureau of Economic Analysis) and Council of Economic Advisers.
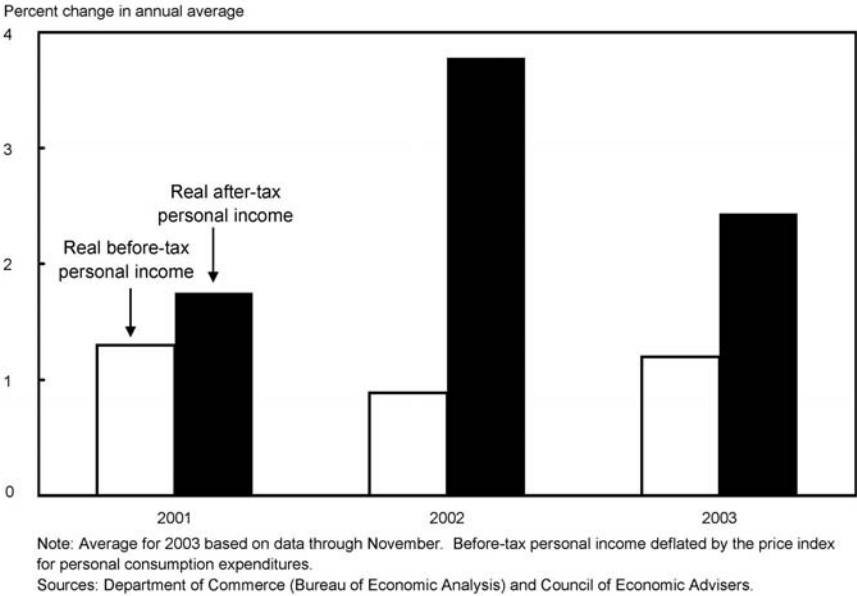
years (Chart 1-8). Over the preceding five years, average annual growth in real after-tax income was more than ½ percentage point below the growth rate of real before-tax income. Numerous studies have shown that long-term tax cuts foster higher consumer spending. Thus, the additional income provided by the tax cuts is likely to have substantially boosted aggregate demand since 2000.

The tax cuts provided further stimulus by increasing incentives for business investment. Some of these incentives came in the form of *bonus depreciation* for business investment, an expansion in the amount of expensing of investment available for small businesses. The bonus depreciation was introduced in the 2002 tax cut (JCWAA), which specified that 30 percent of the price of investments made by September 10, 2004 could be treated as an immediate expense under the corporate profits tax and the remaining 70 percent depreciated over time according to the regular depreciation schedules. Moving the depreciation closer to the time of new investment increased the present value of depreciation allowances and the net after-tax return on investment. The 2003 tax cut (JGTRRA) raised the bonus depreciation to

50 percent of the price of new equipment and extended the period of eligibility so that investments made by the end of 2004 would be covered. It also increased the cap on small-business expensing from $25,000 to $100,000 per year through 2005, effectively lowering the cost of investment for small businesses. These tax changes lowered firms' cost of capital and likely provided support for investment at a crucial time.

The tax cuts also reduced the cost of capital and increased incentives for business investment by lowering tax rates on personal capital income. The 2001 tax cut (EGTRRA) phased out the estate tax and reduced marginal tax rates on all forms of income. These steps lowered the tax burden on capital income received from corporations and also on income received through sole proprietorships, partnerships, and S corporations (corporations for which income is taxed through individual tax returns). In addition, the 2003 tax cut (JGTRRA) reduced taxes on corporate dividends and capital gains.

Altogether, these three tax bills provided $68 billion in tax stimulus in fiscal year 2001, $89 billion in fiscal year 2002, $159 billion in fiscal year 2003, and $272 billion in fiscal year 2004. However, the bills were designed not only to provide short-term stimulus, but also to encourage stronger economic growth over the long run. Lower tax rates on labor income provide an incentive to increase work effort. Lower tax rates on capital income—the reward for saving and investment—provide an incentive to do more of these activities. Investment increases the amount of capital for each worker and also increases the rate at which new technology embodied in capital can be put to use. According to one study, the cut in taxes on capital income in the 2003 tax package (JGTRRA) reduced the marginal effective total tax rate on income from corporate investment by 2 to 4 percentage points. Lower taxes on dividends and capital gains also move the tax system toward a more equal treatment of debt and equity, of dividends and capital gains, and of corporate and noncorporate capital. This move increases economic efficiency because it promotes the allocation of capital based on business fundamentals rather than a desire for tax avoidance.

In sum, the tax cuts supported by this Administration provided a substantial short-term stimulus to consumption and investment and promoted strong and sustainable long-term growth. In weighing the merits of countercyclical monetary and fiscal policy, the stimulus provided by discretionary fiscal policy may be especially important in the low-inflation, low-interest-rate environment the country now enjoys. Under these circumstances, the Federal Reserve may have less room to cut interest rates, and direct stimulus to demand from fiscal policy may be needed to ensure that the Nation's resources are fully utilized in the face of cyclical weakness.

Lesson 5: Strong Productivity Growth Raises Standards of Living but Means that Much Faster Economic Growth is Needed to Raise Employment

One distinctive feature of this recession and recovery has been the remarkably fast growth of *labor productivity*—the amount of goods and services that a worker with given skills produces from each hour of work. The late 1990s had already witnessed an acceleration of productivity growth from an average annual rate of around 1½ percent between the fourth quarters of 1972 and 1995 to a roughly 2½ percent rate between the fourth quarters of 1995 and 2000. Productivity growth then picked up further, contrary to the usual experience in which productivity growth has typically softened in the quarters surrounding business-cycle peaks. In the latest recession, productivity growth leveled off for just one quarter before beginning to rise rapidly (Chart 1-9). Since the fourth quarter of 2000, productivity has increased at an exceptional annual rate of more than 4 percent per year.

Labor productivity growth can be decomposed into the skills of the workforce *(labor quality)*, increases in the amount of capital services per worker-hour *(capital deepening)*, and increases in *total factor productivity*—a

Chart 1-9 Productivity in the Nonfarm Business Sector
Productivity has risen unusually rapidly relative to the average prior recession/recovery period.

Index, level at business cycle peak = 100
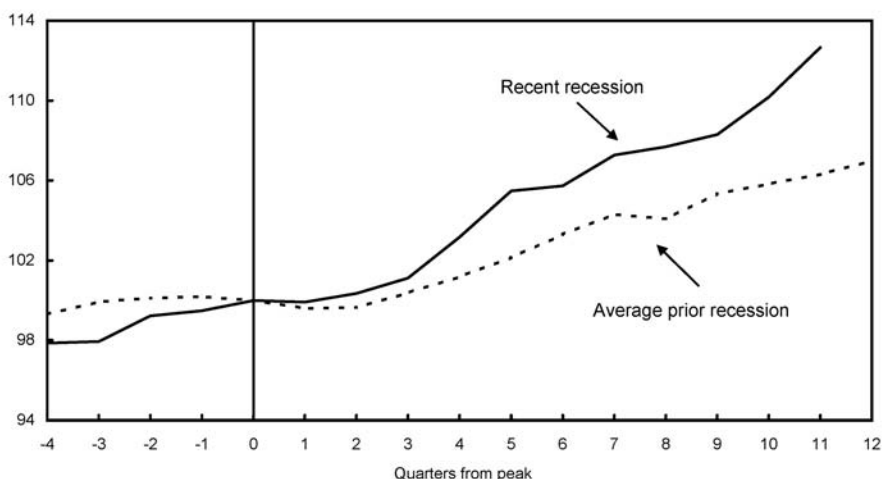
Note: Recent peak set by Council of Economic Advisers at 2000:Q4. Average based on prior recessions since 1960. Data based on productivity available as of December 3, 2003 (that is, prior to the benchmark revision of the National Income and Product Accounts).
Source: Department of Labor (Bureau of Labor Statistics).

residual category that captures the change in aggregate output not explained by changes in capital and labor inputs. According to this framework (as detailed in last year's *Report*), productivity growth stepped up in the mid-1990s partly because the rapid pace of business investment generated large increases in the amount of capital available to each worker. Yet a larger part of this acceleration owes to faster growth in the unexplained residual category of total factor productivity.

The explanation for faster productivity growth in the past couple of years is not clear (especially since the information needed to decompose productivity growth over this period is quite limited). One possibility is that weaker profits and skepticism about the return to new physical investment have encouraged firms to make better use of the resources they already had rather than investing in new technology and capacity. This effort to increase what is sometimes called *organizational capital* might involve, for example, restructuring production processes and retraining workers to take maximum advantage of new information-technology equipment installed in the late 1990s. Another possibility is that firms somehow induced extra work effort for a time because they were hesitant to hire new workers until they were more confident that increases in final demand would persist. A third possibility is that the slower recent pace of gross investment may have been accompanied by slower depreciation of the existing capital stock so that firms lengthened replacement cycles and held on to their existing equipment for longer periods. If this were the case, net investment and the growth rate of the capital stock would have been stronger than indicated by measures based on historical depreciation rates.

In the long run, productivity growth is the key determinant of growth in living standards. Without labor productivity growth, our nation's output and income would grow only at the rate at which the labor force expands; if the labor force grows proportionally with population, this would mean that income per person would be unchanged. With productivity growth, income per person increases. Indeed, U.S. average income is close to eight times as high as it was one hundred years ago, similar to the increase in productivity over this period. The recent robust gains in productivity have boosted both corporate profits and employees' compensation. Corporate profits declined sharply during the recession, but turned around and rose briskly in 2003 (based on data through the first three quarters). Average hourly earnings of production workers in private industry have risen at an average annual rate of close to 3 percent over the past three years. Moreover, productivity growth has reduced inflationary pressures by holding down growth in unit labor costs. As a result, wage gains after adjusting for inflation have been even more impressive by historical standards. In this recession, real average hourly earnings, published in the Bureau of Labor Statistics employment release, never fell below their pre-recession levels, and increased nearly 3 percent in

the eleven quarters after the recession began. The experiences in past recessions have been diverse, but many show a net decline in real hourly earnings or much weaker growth even eleven quarters after the start of the recession.

By definition, labor productivity multiplied by hours worked equals output. Thus, in an arithmetic sense, faster productivity growth generally implies that output must expand more rapidly to generate employment gains. The same principle explains why the rapid pace of productivity growth over the past couple of years has meant that gains in output occurred without gains in employment, until recently.

Indeed, the performance of employment over the past couple of years has been appreciably weaker than in past business cycles (Chart 1-10). Employment was slow to pick up in the average previous recovery, perhaps because employers delayed hiring until they became confident that the increases in demand were sustainable. However, such sluggishness typically has been short-lived (a quarter or two) and followed by vigorous expansion. In contrast, in the current business cycle, employment did not begin its recovery until nearly two years after the upturn in real GDP. The perform-ance of employment in this cycle has lagged even that of the so-called "jobless recovery" from the 1990-1991 recession. (Chart 1-10 shows data from the *establishment survey* done by the Bureau of Labor Statistics (BLS). The BLS *household survey* can show a different pattern—as it has done over the past couple of years. As discussed in Box 1-2, however, the BLS views the estab-lishment survey as a more accurate indicator of labor market conditions.)

Chart 1-10 **Total Nonfarm Employment**
The performance of employment in this recovery has lagged that in the typical recovery and even that in the "jobless recovery" of 1990-1991.
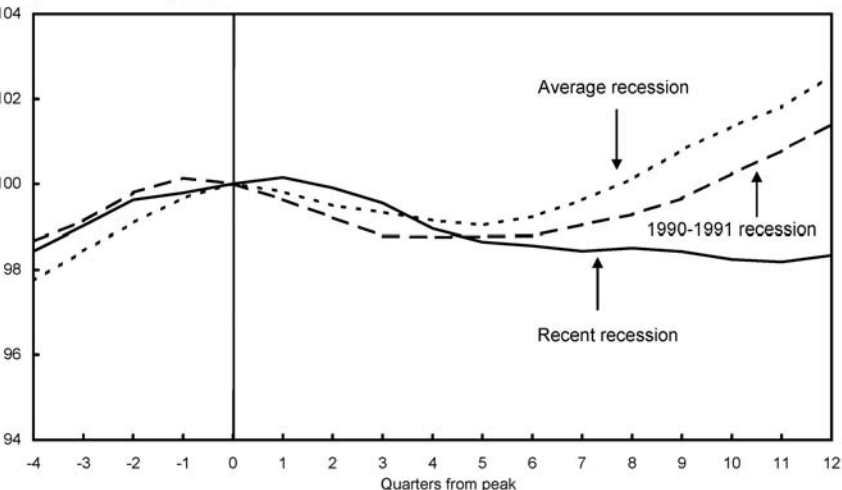
Note: Recent peak set by Council of Economic Advisers at 2000:Q4. Average based on prior recessions since 1960.
Source: Department of Labor (Bureau of Labor Statistics).

Nonetheless, one should not conclude that rapid productivity growth *causes* low employment growth. Rapid productivity growth means that output must increase faster for employment to expand, but it also means that the economy is capable of growing faster. In the long run, the faster rate of *potential output growth* is undoubtedly a good thing for living standards.

Box 1-2: Two Surveys of Employment

Everyone who works is either employed by a firm or is self-employed. Therefore, to count the total number of workers, one could ask each person whether he or she is employed, or one could ask each firm how many workers it employs. The Bureau of Labor Statistics, the agency responsible for tracking employment, uses both approaches. When the BLS asks individuals about their employment status, the results are summarized in the *household survey* of employment. When the BLS asks firms, it produces the *establishment survey* of employment.

Though both surveys ask about employment, they have some important differences that can cause their results to diverge. For example, the establishment survey obtains data from about 160,000 businesses and government agencies that represent about 400,000 worksites and employ over 40 million workers. The sample covers about one-third of all nonfarm payroll jobs in America. The household survey, in contrast, collects data from about 60,000 households, thereby directly covering fewer than 100,000 workers. The establishment survey's larger base of respondents means the calculated margin of error of its estimates is significantly smaller than that associated with the household survey estimates. In addition, the establishment survey is revised annually to match complete payroll records from the universe of establishments participating in state unemployment insurance programs, while the household survey is not.

Furthermore, definitional differences affect the scope of employment measured by the surveys. The establishment survey estimate represents the number of *payroll jobs*, or the number of jobs for which firms pay compensation, while the household survey estimate represents the number of employed persons. Because some people hold more than one job, the total number of payroll jobs can exceed the total number of employed persons. On the other hand, the household survey includes employees working in the agricultural sector, the unincorporated self-employed, unpaid family workers, workers in private households, and workers on unpaid leave from their jobs. The establishment survey excludes all of these categories because they are not reported on the nonfarm business payrolls that provide the source data for the survey.

Box 1-2 — *continued*

These differences and other factors create a gap between the household and establishment surveys' employment estimates, though they tend to display similar long-term trends. The average gap since 1990 has been about 6 percent, or 8 million workers.

While long-term trends in the two surveys are similar, over shorter periods of time their trends have sometimes diverged. This has been the case since late 2001, when employment from the two surveys has trended in opposite directions. For the first time in the two series' histories, one showed a large and sustained *decrease* in employment while the other showed a large and sustained *increase*. In particular, the establishment survey reported a decline in employment of over 1.0 million from the end of the recession in November 2001 to August 2003, while the household survey reported an increase of over 1.4 million. In every month of 2003, the establishment survey showed employment below the November 2001 level, while the household survey showed it above this level. Such a sustained string of divergence is unprecedented.

One possible explanation is that the establishment survey misses some new firms and therefore may underestimate employment at the start of an economic expansion. Past revisions to the establishment survey offer some support for this theory. For the recent data, however, this theory can explain at most the divergence since March 2003, because establishment survey data up to that point appear consistent with unemployment insurance records that cover all establishments. Another possible explanation is that the household survey results are overstated because of the way in which the survey results are extrapolated to represent the entire population. Specifically, information from the 2000 Census, together with estimates of how the population is changing over time, are used to determine how many actual U.S. households correspond to each household in the sample. If, for example, immigration has been unexpectedly low because of tighter border controls and the weaker labor market over the past few years, the estimated number of U.S. households corresponding to each household in the sample may be overstated. As a result, the estimates of total employment (and other aggregates based on the population estimates) from the household survey could be too high.

Both surveys contain valuable information about current economic developments, but, as with all economic statistics, the data from both surveys are imperfect. The Bureau of Labor Statistics has stated that the establishment survey is generally the more reliable indicator of current trends in employment. Still, the explanation for why these two surveys' results have diverged so markedly over the last few years, and what this might indicate about the economic recovery, remains a puzzle.

Conclusion

The U.S. economy is much stronger now than it was a year ago and, as will be discussed in Chapter 3, prospects for the coming year look solid. Nonetheless, the experiences of the past several years remain relevant for the future. Understanding the negative forces that weighed against the economy, as well as the policies that contributed to the recovery, can help policy makers ensure that economic activity maintains a strong upward trend in the years ahead.

CHAPTER 2

The Manufacturing Sector

The manufacturing sector was affected by the latest economic slowdown earlier, longer, and harder than other sectors of the economy and only recently have manufacturing employment losses begun to abate. Over the past several decades, the manufacturing sector has experienced substantial output growth, even while manufacturing employment has declined as a share of total employment. This chapter examines recent developments and long-term trends in manufacturing and considers policy responses.

The key points in this chapter are:

- The severity of the recent slowdown in manufacturing was largely due to prolonged weakness in business investment and exports, both of which are heavily tied to manufacturing.
- The manufacturing employment decline over the past half-century primarily reflects striking gains in productivity and increasing consumer demand for services compared to manufactured goods. International trade plays a relatively small role.
- Consumers and businesses generally benefit from the lower prices made possible by increased manufacturing productivity, and strong productivity growth has led to real compensation growth for workers. The shift of jobs from manufacturing to services has caused dislocation but has not resulted, on balance, in a shift from "good jobs" to "bad jobs."
- The best response to recent developments in manufacturing is to focus on stimulating the overall economy and easing restrictions that impede manufacturing growth. This Administration has actively pursued such measures.

Manufacturing and the Recent Business Cycle

This section looks at the characteristics and causes of the recent economic downturn with particular focus on the manufacturing sector. Output in manufacturing held up relatively well in the recent recession, but employment declined sharply. Data released over the past few months are encouraging regarding the prospects for recovery in the manufacturing sector.

The Recent Downturn in Manufacturing Output

Manufacturing output dropped 6.8 percent from its peak in June 2000 to its trough in December 2001. This was a larger decline than that for real GDP, which fell only 0.5 percent from its peak in the fourth quarter of 2000 to its trough in the third quarter of 2001. This gap is not out of line with historical experience: manufacturing output has dropped much more than real GDP during past business cycles (Chart 2-1). What is more unusual is that the recovery in manufacturing output has been far weaker than the recovery in real GDP.

As discussed in Chapter 1, *Lessons from the Recent Business Cycle*, investment demand was especially weak during the recent recession. A slowing of demand for equipment investment disproportionately hurts the manufacturing sector because nearly all business equipment involves manufactured products. The rest of final demand, in contrast, involves a mix of manufactured goods, agricultural products, services, and structures. The industries within manufacturing contributing most to the downturn in manufacturing output were those primarily associated with the production of business equipment. In particular, slower growth in production of computers and other electronics, machinery, and metals accounts for nearly two-thirds of the swing in manufacturing output from its rapid growth in the late 1990s (an annual rate of 6.9 percent) to its decline in the 18 months after

Chart 2-1 Real GDP and Manufacturing Industrial Production
Manufacturing industrial production is more volatile than real GDP.

Percent change from 4 quarters earlier
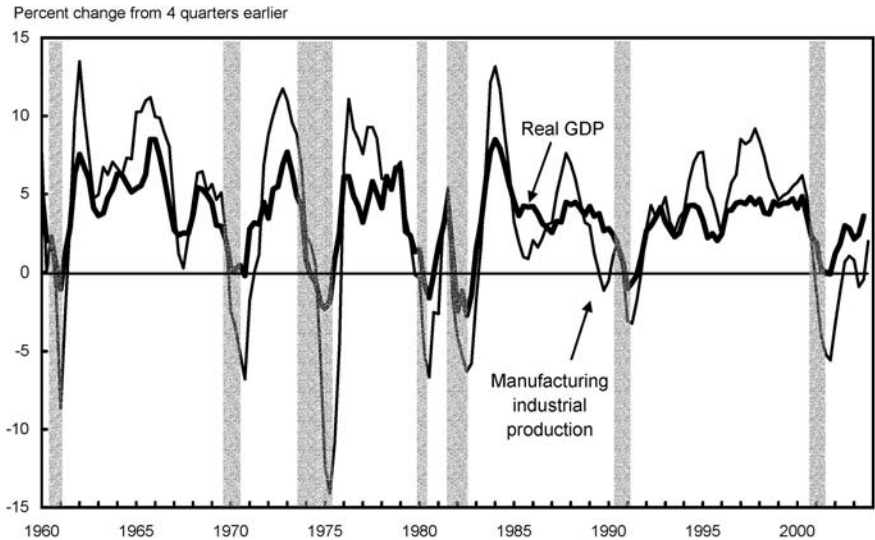
Note: Shaded areas indicate recessions. Recent peak set by Council of Economic Advisers at 2000:Q4.
Sources: Department of Commerce (Bureau of Economic Analysis) and Board of Governors of the Federal Reserve System.

mid-2000 (an annual rate of -4.6 percent). Some parts of manufacturing saw especially difficult times. The metalworking machinery industry, of which the hard-hit tool and die industry makes up 40 percent of employment, has seen its payrolls decline by almost 25 percent from mid-2000 to the end of 2003. Real production in the metalworking machinery industry fell by more than 35 percent over this period.

The timing of the manufacturing slowdown also strongly suggests a link to the decline in business investment (Chart 2-2). Manufacturing output declined substantially in the middle of 2000, months before real GDP turned downward around the fourth quarter of 2000. This pattern mirrors that of business investment in equipment and software, which also peaked in mid-2000—well before the overall economy. The prolonged period of weakness in manufacturing output also bears a notable similarity to the sluggish recovery in investment in equipment and software.

Lackluster demand for U.S. exports has been another source of weakness in the manufacturing sector over the past three years. Exports have been depressed, in part due to slow growth in other major economies. Since the fourth quarter of 2000, the average annual rates of real GDP growth in the euro area and Japan have been less than half that of the United States. Industrial supplies and capital goods make up the bulk of U.S. goods exports. Lower exports of manufactured goods can account for all of the decline in exports since 2000.

Chart 2-2 Manufacturing Industrial Production and Real Investment
Manufacturing industrial production has been low, in part reflecting low investment, since the recent recession began.

Sources: Department of Commerce (Bureau of Economic Analysis) and Board of Governors of the Federal Reserve System.
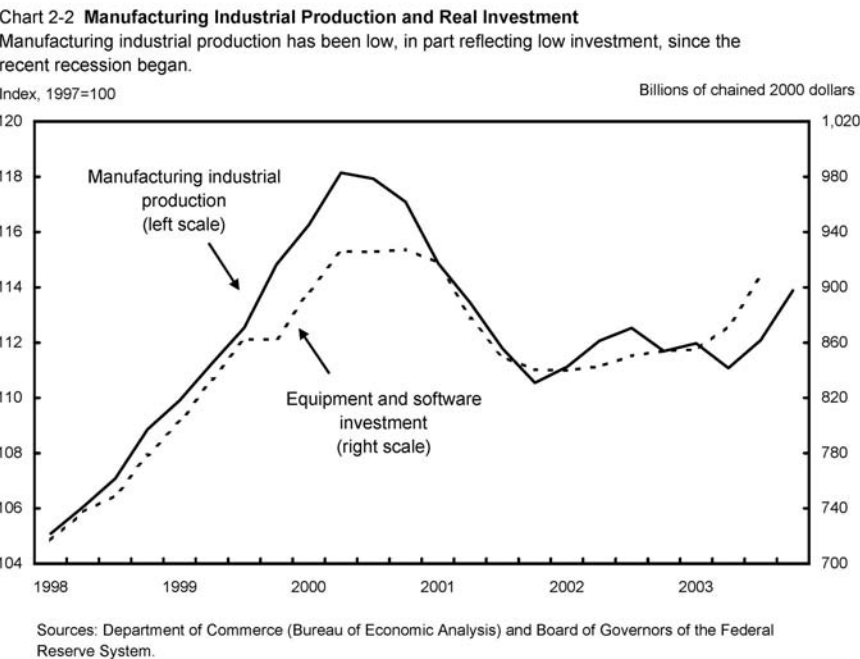
Manufacturing Employment in Recent Years

   Manufacturing employment declined more than manufacturing output during the recent downturn, just as overall employment declined more than overall output. Manufacturing employment declined 16 percent from June 2000, the peak of manufacturing production, to December 2003—a steeper decline than in recessions on average (Chart 2-3). In fact, the recent drop in manufacturing employment was the largest cyclical decline since 1960.

   As with the overall economy, the weakness of manufacturing employment relative to output during and after the recent recession has been reflected in rapid productivity growth (Chart 2-4). From the fourth quarter of 2000 through the third quarter of 2003, productivity in the nonfarm business sector and in the manufacturing sector rose more than 4 percent at an annual rate—appreciably faster than in recessions on average since 1960. This rise has allowed businesses to increase output without a corresponding increase in labor input.

Chart 2-3 **Manufacturing Employment**
Manufacturing employment was particularly hard-hit in the recent recession.

Index, level at manufacturing industrial production peak=100

Months from peak in manufacturing industrial production
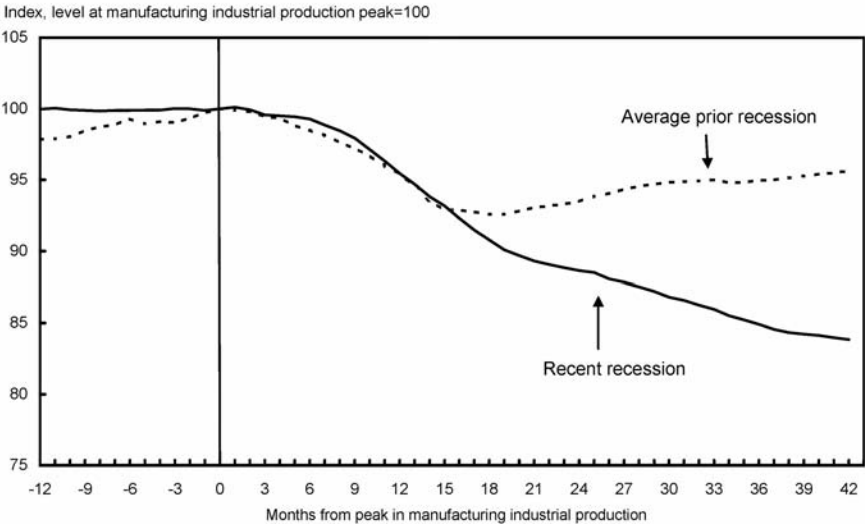
Note: Average based on prior recessions since 1960.
Sources: Department of Labor (Bureau of Labor Statistics) and Board of Governors of the Federal Reserve System.

Chart 2-4 Productivity in Manufacturing
Manufacturing productivity grew more quickly during the recent recession and recovery than in the average prior recession.
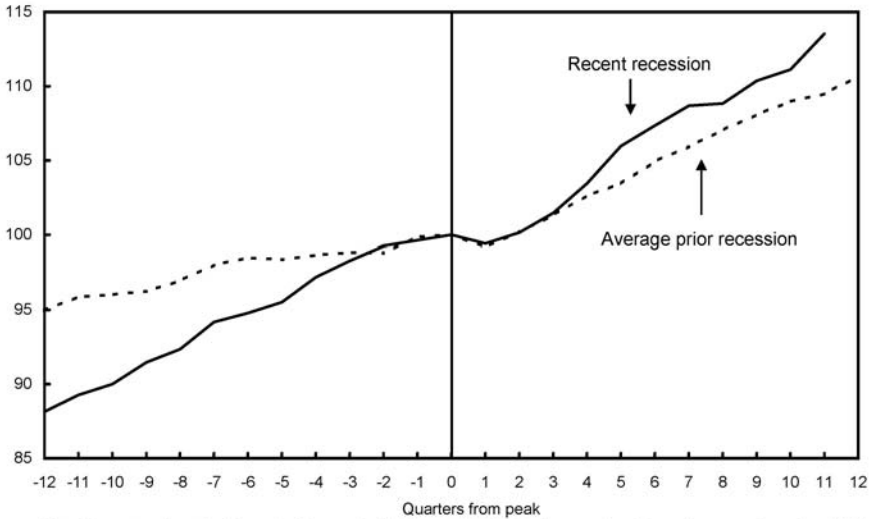
Index, level at business cycle peak=100

Note: Recent peak set by Council of Economic Advisers at 2000:Q4. Average based on prior recessions since 1960.
Source: Department of Labor (Bureau of Labor Statistics).

Signs of Recovery in the Manufacturing Sector

Data for the second half of 2003 suggest a noticeable firming in the manufacturing sector. Orders and shipments of capital goods began to increase around the middle of 2003. Industrial production rose at an average annual rate of 5.9 percent during the second half of the year, the largest six-month gain since the first half of 2000. In addition, the new orders index from the Institute of Supply Management's monthly survey of purchasing managers rose to its highest level in two decades, indicating widespread optimism that activity is picking up. Moreover, some of the factors that have historically affected firms' production decisions support a further strengthening—the cost of capital is low by the standards of the last decade and manufacturers' profits are well above their levels of two years ago.

Although manufacturing employment fell throughout 2003, recent developments hint at improving employment conditions for the sector as a whole. To be sure, some industries continue to lag—for example, textiles, apparel, printing, and petroleum and coal industries have seen employment fall substantially more than overall manufacturing employment since mid-2003. More

broadly, however, the rate of decline in overall manufacturing employment eased noticeably in the fourth quarter of 2003, with the smallest quarterly loss in three years. In addition, the rise in temporary-help services since the spring of 2003 is consistent with a future rebound in permanent employment. The temporary-help sector supplies a substantial share of its workers to the manufacturing sector, and over the past decade has tended to lead movements in the permanent payrolls of manufacturing firms (Chart 2-5).

Chart 2-5 **Employment in Manufacturing and Temporary-Help Services**
Changes in temporary-help services employment tend to lead changes in manufacturing employment.
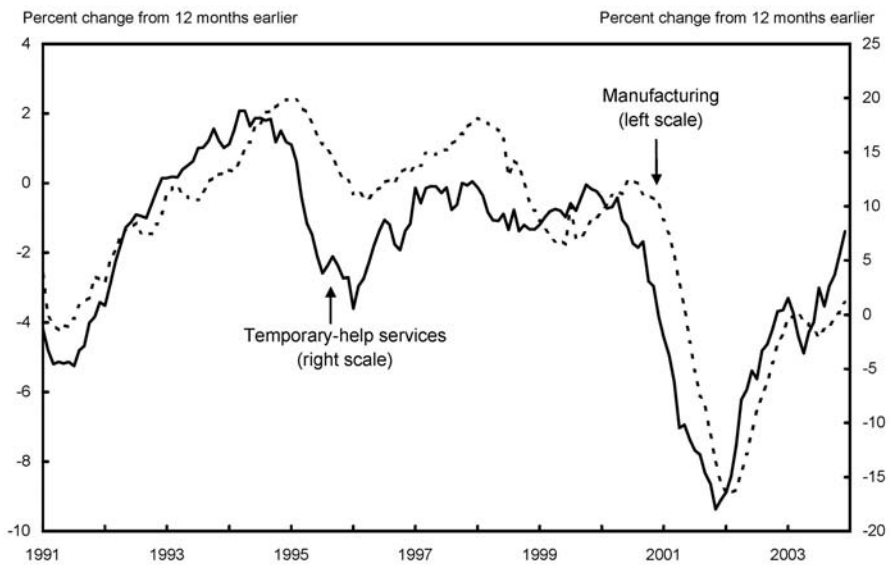
Source: Department of Labor (Bureau of Labor Statistics).

Long-Term Trends

To place the recent experience of the American manufacturing sector in perspective, this section examines the evolution of the manufacturing sector as a whole over the 50 years from 1950 to 2000 along three key dimensions: output, productivity and demand, and employment.

Manufacturing Output over the Long Term

Manufacturing output increased dramatically from 1950 to 2000, with particularly strong growth in the 1990s (Chart 2-6). *Manufacturing industrial production*, a measure of real manufacturing output, increased more than sixfold from 1950 to 2000 before declining in the recent recession. Over the same period, annual growth in manufacturing industrial production averaged 3.8 percent, faster than real GDP growth of 3.4 percent. From 1990 to 2000, manufacturing industrial production expanded at an annual rate of 4.6 percent, outpacing real GDP growth by more than a percentage point. Per capita consumption of manufactured goods has also risen: consumption of goods excluding food and fuel more than quadrupled in real 2000 dollar terms from $1,400 per person in 1950 to $6,000 per person in 2000.

Chart 2-6 Real GDP and Manufacturing Industrial Production
Manufacturing industrial production increased at a faster rate than real GDP over the period 1950 to 2000.
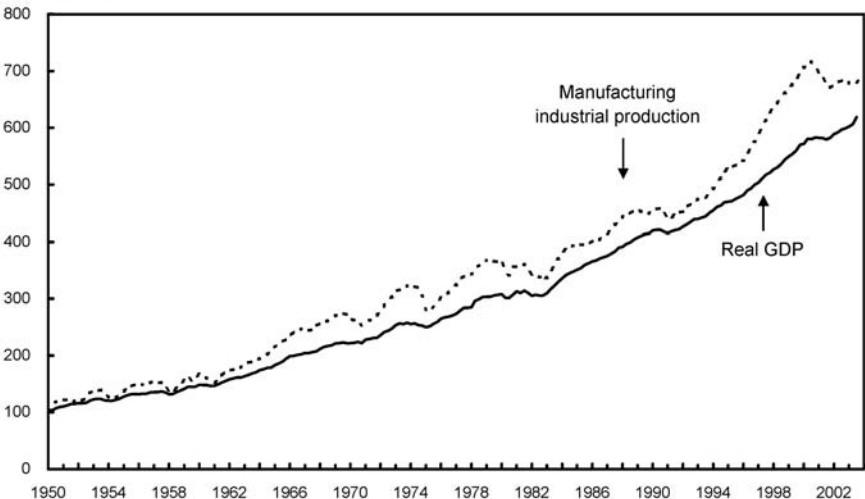
Sources: Department of Commerce (Bureau of Economic Analysis) and Board of Governors of the Federal Reserve System.

In contrast to real manufacturing output, *nominal manufacturing output* (the dollar value of manufacturing output) has grown more slowly than *nominal* GDP (the dollar value of GDP). As a result, the share of nominal GDP accounted for by manufacturing roughly halved, from 29 percent in 1950 to 15 percent in 2000 (based on GDP by industry data available when this *Report* went to press; that is, prior to the 2003 benchmark revision of the National Income and Product Accounts).

Manufacturing Productivity and Demand over the Long Term

Two factors are driving the declining share of manufacturing in U.S. nominal output. First, and most significant, productivity growth in manufacturing lowered the relative price of manufactured goods, but demand did not respond proportionately. Second, imported manufactured goods increased their market share.

Productivity, as measured by output per hour worked, has grown more rapidly in manufacturing than in the overall nonfarm business sector over the last three decades. From 1950 to 1973, manufacturing productivity grew at about the same pace as productivity overall. Over the period from 1973 to 1995, manufacturing productivity growth exceeded productivity growth overall by about 1 percentage point per year. The disparity is even wider over the period from 1995 to 2000, when manufacturing productivity grew at an annual rate nearly 2 percentage points higher than nonfarm business productivity (Chart 2-7). An hour of work in manufacturing produced about four times as much in 2000 as it did in 1950, whereas an hour of work in the nonfarm business sector produced less than three times as much in 2000 as it did in 1950.

This dramatic productivity differential has contributed to a decline in the price of manufactured goods relative to services, which in turn helps to explain the difference between the behavior of nominal and real manufacturing output. Increased labor productivity in a sector means that fewer hours are required to make a given amount of output. This reduces the cost of production and, typically, the relative price of that output. In the same way, relative prices tend to increase in sectors that have experienced less productivity growth, such as services. For example, the falling prices of computers and other electronics have contrasted sharply with the rising costs of services. This example is confirmed by the aggregate data: the average price of consumption goods relative to services fell more than 50 percent between 1950 and 2000. In contrast to the nearly ninefold increase in the prices for

Chart 2-7 **Productivity Growth**
Growth in manufacturing productivity outpaced that in nonfarm business productivity over
the periods 1973 to 1995 and 1995 to 2000.

Percent, at an annual rate
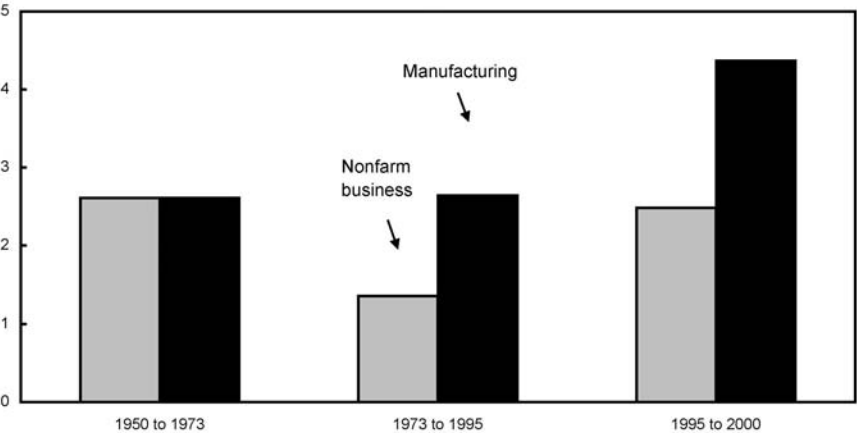
Note: Nonfarm productivity based on data available as of December 3, 2003 (that is, prior to the benchmark
revision of the National Income and Product Accounts). Manufacturing productivity based on revised estimates
beginning 1987 (released on January 7, 2004); as revised data for earlier years are not yet available, growth
rates are assumed to be the same as the previously-published SIC-based estimates.
Sources: Department of Labor (Bureau of Labor Statistics) and Council of Economic Advisers.

services, prices for *durable goods* (goods such as cars and refrigerators that are
expected to last, on average, three years or more) rose by a factor of only
2½ and prices for *nondurable goods* rose by a factor of about 5 from 1950 to
2000 (Chart 2-8). Expressed another way, to equal the buying power of
$100 worth of durable goods in 1950, a consumer would have spent $250 in
2000, while for $100 worth of services in 1950, a consumer would have
spent $890 in 2000.

   The slower growth of manufactured goods prices has increased the
purchasing power of incomes relative to what it otherwise would have been,
but the portion of this increase that Americans have allocated to manufac-
tured goods has not been large enough to maintain manufacturing's share of
nominal output. The boost to real income from the relative price decline of
manufactured goods has supported demand not only for these goods but
also for services such as health care and financial advice. That is, Americans
have used the resources made available from the relatively slow growth in
manufacturing prices to buy many things, not just manufactured goods.
Increased demand for services, combined with rising relative prices for serv-
ices, is reflected in the fact that health services and business services each

Chart 2-8 Price Level by Category of Personal Consumption Expenditures
Goods prices rose much less than services prices from 1950 to 2000.

Index, 1950=100
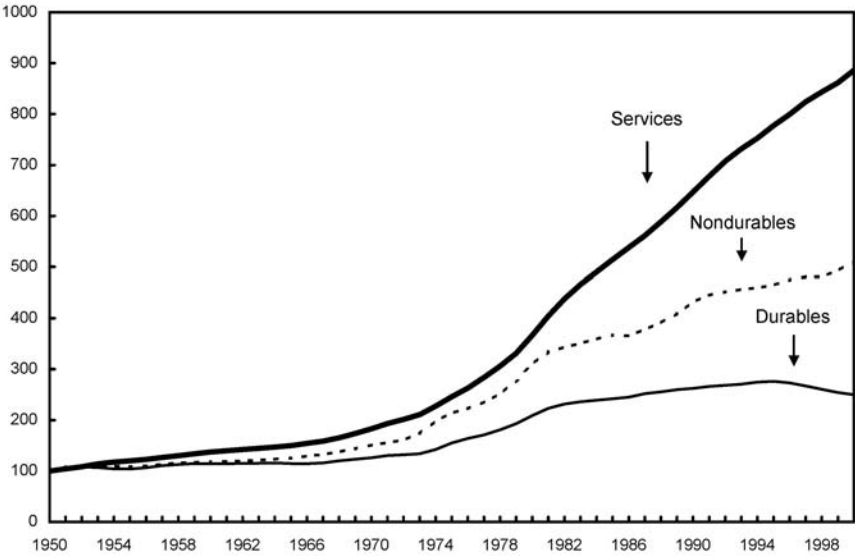
Source: Department of Commerce (Bureau of Economic Analysis).

have increased their share of total nominal output about 4 percentage points since 1950. The finance, insurance, and real estate industry has increased its share a dramatic 9½ percentage points. The opposite trend has held for manufacturing, in which relative price declines have not been fully offset by increases in demand. This explains why the share of manufacturing in total nominal output has roughly halved since 1950. (All calculations of industry share of nominal GDP are based on the pre-benchmark data available when this *Report* went to press.)

In other words, U.S. demand for manufacturing products has been relatively *price inelastic*. That is, demand has not been very responsive to price declines. For example, a family that purchased a car may have reacted to lower relative car prices (and the increased real income they create) by paying for college or hiring a home health care aide, rather than by putting those gains toward the purchase of another car. As a numerical example of inelastic demand, suppose that people buy 10 compact discs at $20 each (for a total expenditure of $200). Now suppose the price falls from $20 to $10. If people buy twice as many compact discs at $10, the value of overall sales will still be $200 (20 compact discs at $10 each). But if people increase their purchases to 15 compact discs, the value of overall sales will be only $150, a decline of 25 percent. This is similar to what has happened in manufacturing.

Productivity gains have tempered price increases, and demand has not responded strongly enough to keep nominal revenues constant as a share of nominal GDP.

A second factor that has led to a decline in manufacturing's share of GDP is that Americans are purchasing more goods from abroad. Goods purchases as a share of total domestic purchases have been declining for about 30 years. The share of domestically produced goods has fallen somewhat faster, particularly in the 1970s and 1990s. Domestically produced goods were 91 percent of overall domestic goods purchases in 1970; by 2000, they had fallen to 68 percent. In other words, imports have made up an increased share of goods bought in the United States (Chart 2-9).

Growth in exports of manufactured goods from the United States over the past several decades has offset only some of the growth in imports (Chart 2-10). As a result, net imports of nonagricultural goods (imports minus exports) have risen materially, reaching about 30 percent of manufacturing production in 2000 (based on the pre-benchmark data available when this *Report* went to press) (Chart 2-11). In relation to the overall economy, net nonagricultural goods imports have also risen, but remained below 5 percent of GDP in 2000. China has been a growing source of manufacturing imports, although this growth has not been a major factor in the increase of the U.S. trade deficit (Box 2-1).

Chart 2-9 U.S. Imports and Domestic Production of Goods
Goods produced in the United States have made up a declining share of all goods purchases in the United States since the 1960s.

Percent of total domestic purchases
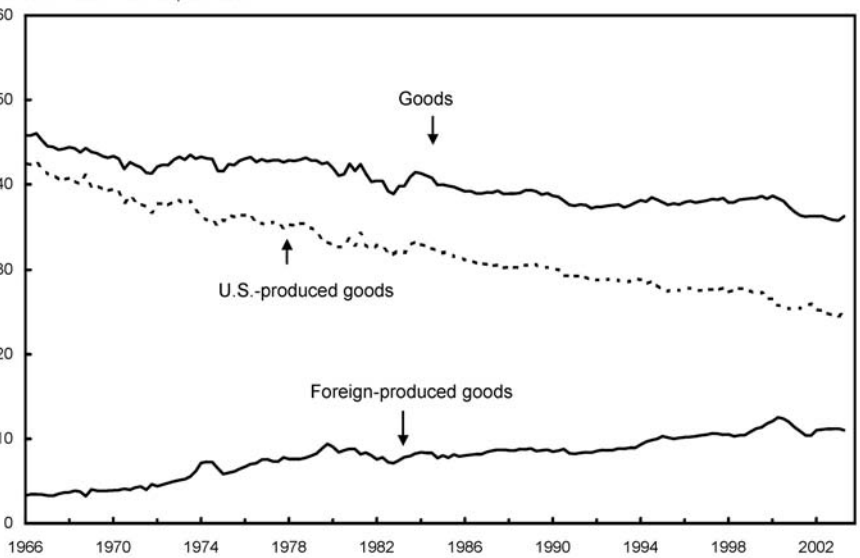
Source: Department of Commerce (Bureau of Economic Analysis).

Chart 2-10 Nonagricultural Goods Trade as a Percent of Manufacturing Output

Imports and exports have increased relative to manufacturing output.

Percent of manufacturing output

Note: Manufacturing output based on data released prior to the 2003 benchmark revision of the National Income and Product Accounts.
Source: Department of Commerce (Bureau of Economic Analysis).
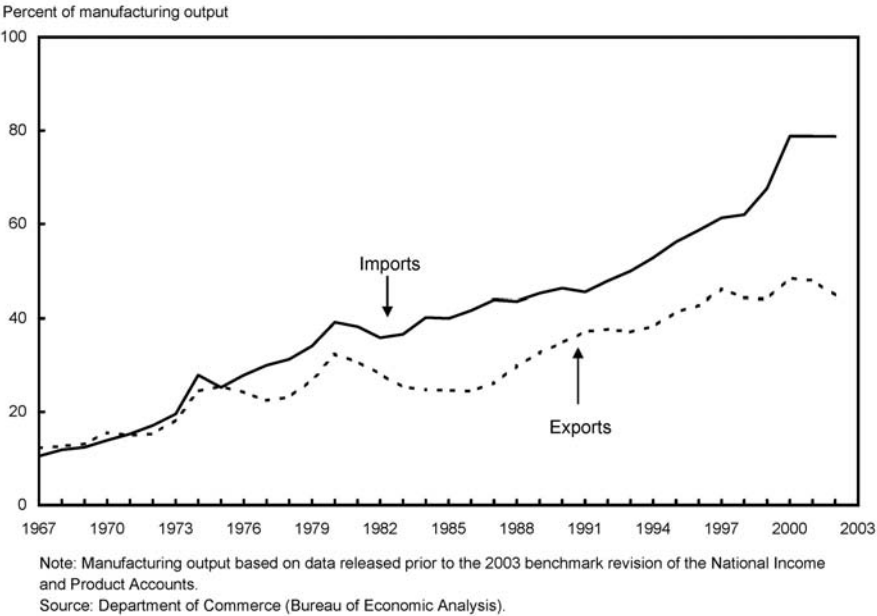
Chart 2-11 Nonagricultural Goods Net Imports as a Percent of Output

Net imports of goods have risen as a percent of manufacturing GDP over the last 30 years but remain small as a share of overall GDP.

Percent

Note: Manufacturing output based on data released prior to the 2003 benchmark revision of the National Income and Product Accounts.
Source: Department of Commerce (Bureau of Economic Analysis).
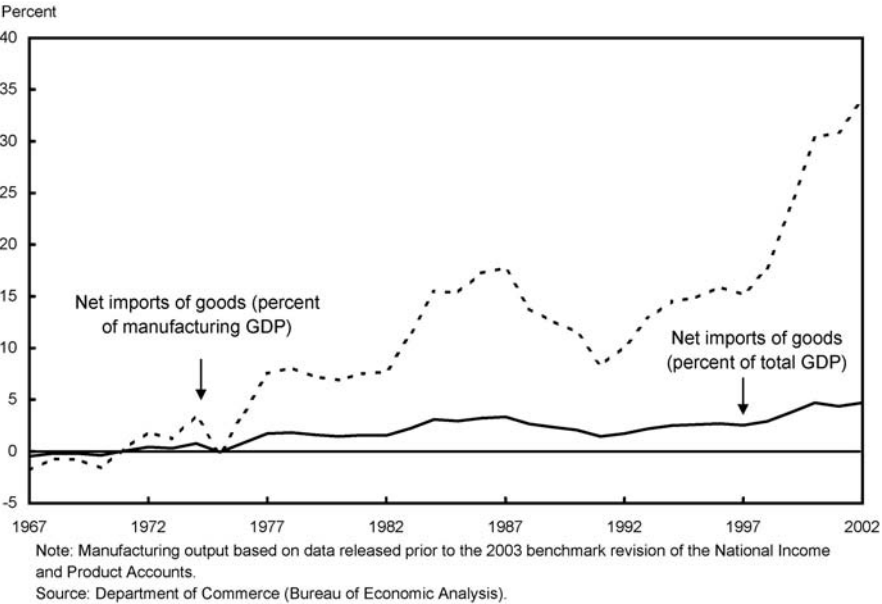
Box 2-1: China and the U.S. Manufacturing Sector

The recent decline in employment in U.S. manufacturing has coincided with a sizable increase in the overall U.S. trade deficit and a sharp increase in the U.S. bilateral trade deficit with China. In part because of the high visibility of Chinese imports, which are primarily everyday consumer goods, these events have raised concerns that imports of Chinese goods come at the expense of American manufacturing workers.

China's Trade with the World

While China's exports and imports grew quickly starting in the early 1990s, China's trade with the rest of the world has been modest until very recently (Chart 2-12). The growth in China's trade has been well balanced in that increased exports to the world have been matched by rising imports from the world. According to data from China's official statistical agency, China has had a trade deficit with the world excluding the United States for several years. China recently ran trade deficits with a number of other countries, including industrial countries such as Germany and Japan.

China's Trade with the United States

China has a significant trade surplus with the United States, its most important export market and the destination of one-quarter of all Chinese goods exports. The U.S. trade deficit with China—about $124 billion through November 2003 at an annual rate—is the single largest bilateral goods and services trade deficit for the United States. The next-largest bilateral deficit is with Japan, at $66 billion through November 2003 at an annual rate.

The U.S. trade deficit excluding China has also risen dramatically since the mid-1990s and is about 3½ times larger than the bilateral deficit with China (Chart 2-13). China's share of the overall U.S. trade deficit in goods has actually fallen since 1997—exactly the period over which trade with China grew rapidly.

Greater trade with China does not appear to have contributed to an increased overall U.S. trade imbalance, as the higher share of U.S. imports from China has been more than offset by a declining share of imports from other Asian countries. The share of U.S. imports from the Pacific Rim as a whole has fallen since the mid-1990s (Chart 2-14). Restrictions on imports from China would be expected to increase imports from other low-cost foreign producers, rather than to increase production and employment for American manufacturers. That is, any job gains from reduced Chinese imports are more likely to occur in other developing countries rather than the United States.

   U.S. exports to China have grown strongly in the last several years, with exports to China up more than 60 percent since 2000. As of the third quarter of 2003, China was the sixth-largest U.S. export market. Exports to China have grown even while exports to the rest of the world have stagnated (Chart 2-15).

The Impact of Trade with China on U.S. Manufacturing Employment

   Imports from China affect the prospects for domestic firms with which they compete, and this impact often extends to workers and communities associated with these firms. This is especially the case for firms that make items that are relatively intensive in the use of less-skilled labor, as these are goods in which China has a comparative advantage in production. This may raise the question of whether imports from China are a primary factor in the displacement of American manufacturing workers.

   A closer look at the data indicates this is not the case. The low level of U.S. imports from China before the mid-1990s suggests that declines in employment prior to that period were not due to U.S. trade with China. The data on more-recent job losses in manufacturing indicate that China is not a primary factor in these declines, either. With the exception of apparel, the largest job losses have occurred in export-intensive industries for the United States, and job losses in U.S. manufacturing have been mainly in industries in which imports from China are small. For example, the computer and electronic equipment industry accounts for 15 percent of all manufacturing job losses since January 2000, but imports from China were only 8 percent of U.S. output in 2002. Other export-intensive industries that have suffered large job losses include fabricated metal products (9 percent of manufacturing job losses and 2 percent of U.S. output), machinery (10 percent and 2 percent), and transportation equipment (12 percent and 0.4 percent).

Chart 2-12 **China's Trade in Goods**
The recent growth in China's trade has been divided fairly evenly between growth in
imports and growth in exports.

Billions of dollars
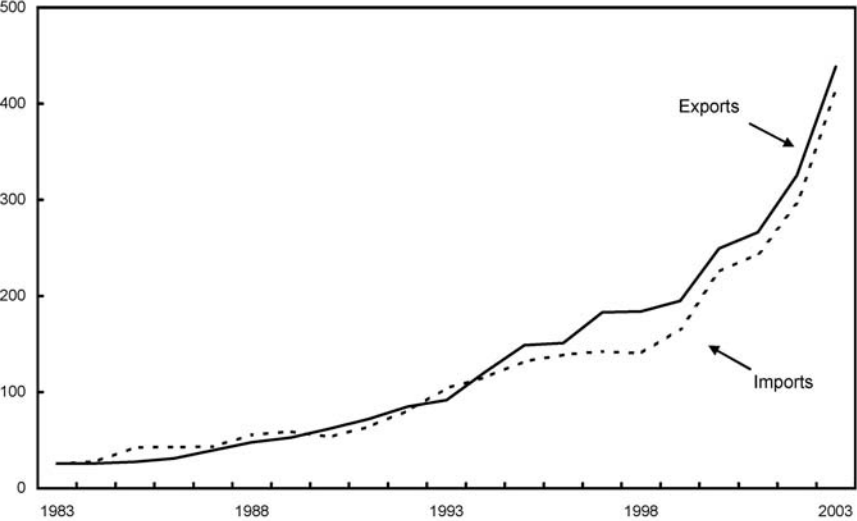
Source: China Customs Statistics.

Chart 2-13 **U.S. Trade Deficit in Goods**
The U.S. trade deficit with countries other than China has risen dramatically, while China's share of the
overall deficit has fallen since its peak in 1997.

Billions of dollars   Percent
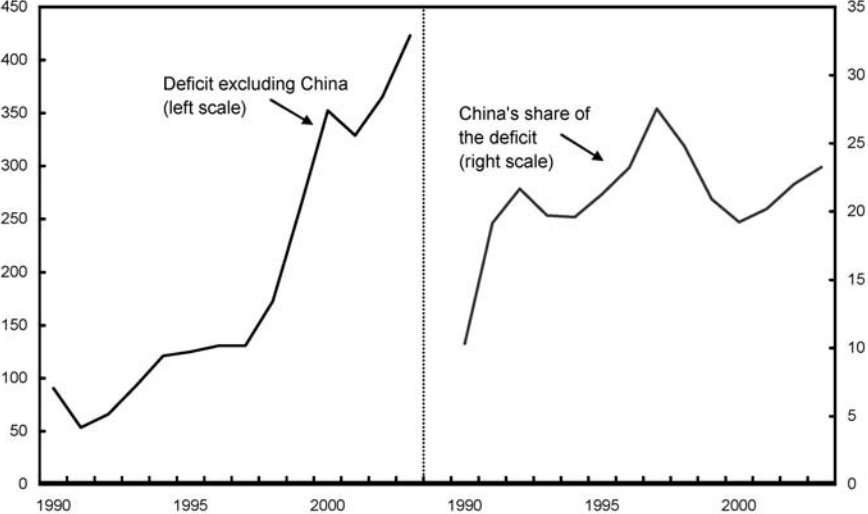
Sources: Department of Commerce (Bureau of the Census) and Council of Economic Advisers.

Chart 2-14 U.S. Imports of Goods

While the share of U.S. imports of goods from China has been increasing, the share of imports from all Pacific Rim countries combined has been falling.

Percent of U.S. imports
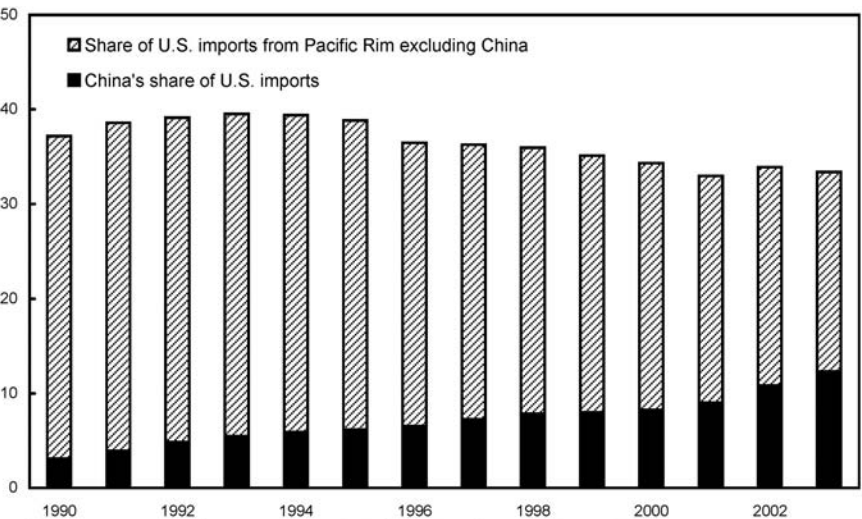
Sources: Department of Commerce (Bureau of the Census) and Council of Economic Advisers.

Chart 2-15 U.S. Exports of Goods

Exports to China have increased dramatically over the past several years compared with lackluster growth in exports to the rest of the world.
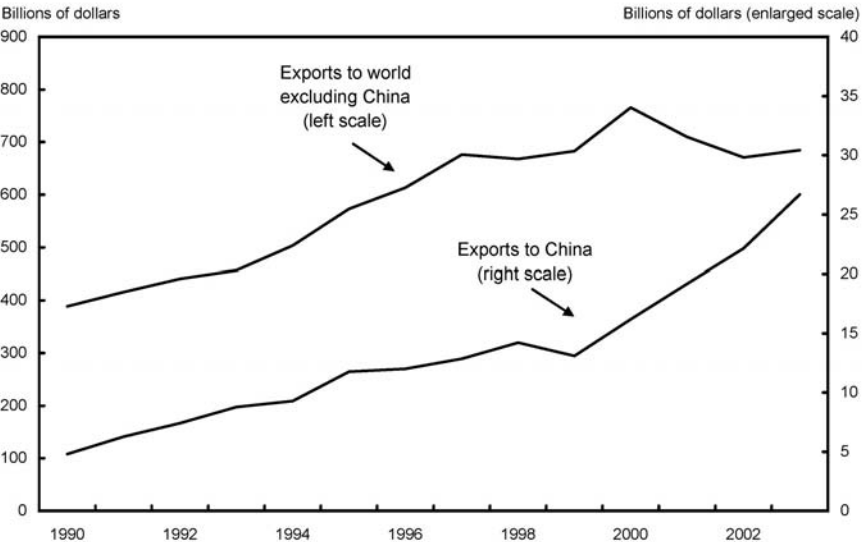
Sources: Department of Commerce (Bureau of the Census) and Council of Economic Advisers.

Manufacturing Employment over the Long Term

Employment in manufacturing as a share of total employment peaked in the early 1940s at about one-third of all farm and nonfarm workers. By 2000, it had declined to just below 13 percent (17 million out of 135 million employees). Employment in service-providing sectors (including transportation, wholesale and retail trade, finance, insurance, and real estate, and services) increased from 35 percent of payroll employment in the early 1940s to 65 percent (86 million workers) of all employees in 2000 (Chart 2-16).

The two main reasons for this shift from the manufacturing sector to service-providing sectors in the labor market are related to the explanations for the declining nominal share of manufacturing output. First, increased demand for services and relatively slow productivity growth in service-providing sectors have led to rising demand for workers in these sectors. In manufacturing, inelastic demand for manufactured goods and faster productivity growth have lowered the relative demand for manufacturing workers.

Second, manufacturing employment likely has fallen in response to the transfer of manufacturing jobs abroad. The jobs affected have generally been those involved in the production of goods requiring relatively low skills.

Chart 2-16 **Employment and Relative Productivity**
The decline in manufacturing's share of employment is largely due to rapid productivity gains.
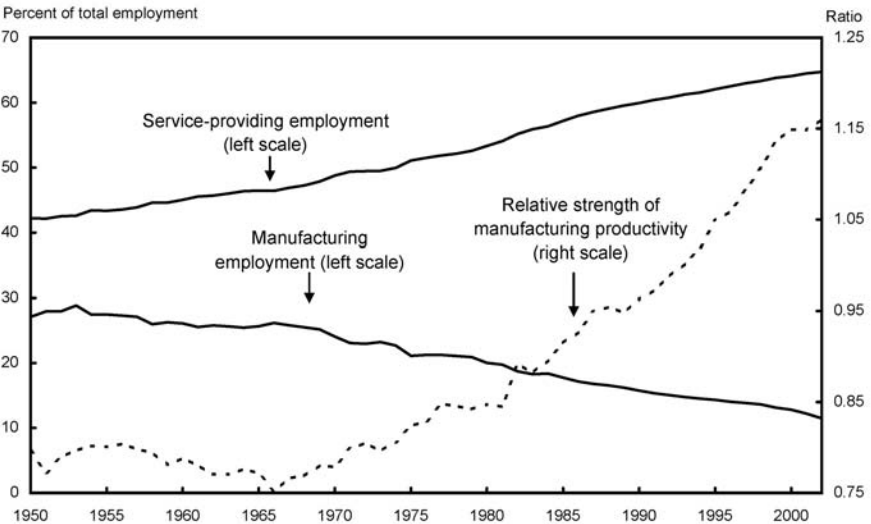
Note: Ratio of manufacturing productivity to nonfarm business productivity available as of December 3, 2003 (that is, prior to the benchmark revision of the National Income and Product Accounts); both indexed to 1992=100.
Sources: Department of Labor (Bureau of Labor Statistics) and Council of Economic Advisers.

Indeed, this is part of the explanation for the rapid growth in manufacturing productivity over the last 50 years (Chart 2-16). The relatively highly-skilled American manufacturing workforce has been increasingly focused on higher-productivity activities. This shift can be seen by looking at compensation for the industries in which employment decreased or increased the most from 1950 to 2000 (Table 2-1). With a few exceptions, employment fell dramatically in industries with relatively low-skilled jobs and rose dramatically in industries with relatively highly-skilled jobs.

This specialization is a natural outcome of the opening of economies all over the world to trade. As a result of such specialization, world efficiency increases and world output goes up as countries focus on the activities in which they are relatively more productive. All countries that participate in trade benefit from this increased output.

The effect of long-term productivity improvements on the shift to service-providing jobs is far more important than increased manufacturing imports. Two simple hypothetical exercises can help to illustrate this. In the first exercise, imagine that manufacturing productivity was fixed at its value in 1970. To match the actual amount of manufacturing output in 2000, one-third of total U.S. nonfarm employment would have been required by manufacturing, compared with the 13 percent required at 2000 productivity levels. That is, without the increase in manufacturing productivity, manufacturing's share of nonfarm employment would have *increased* 8 percentage

TABLE 2-1.— *Employment in Selected Manufacturing Industries*

Industry	Change in employment, 1950 to 2000 (percent)	Compensation per employee as percent of average for all sectors, 2000
Panel A		
Manufacturing industries with employment that grew the fastest		
Rubber and miscellaneous plastics products	213	95
Instruments and related products	207	155
Printing and publishing	102	113
Transportation equipment other than motor vehicles and equipment	83	140
Electronic equipment	77	152
Panel B		
Manufacturing industries with employment that declined the fastest		
Leather and leather products	-82	78
Tobacco products	-65	195
Textile mill products	-58	77
Apparel and other textile products	-50	67
Petroleum and coal products	-42	177

Note.—Data relate to full-time equivalent employees and include some definitional changes. Not yet available are data based on the December 2003 benchmark revision of the National Income and Product Accounts.

Source: Department of Commerce (Bureau of Economic Analysis).

points rather than *decreased* 12 percentage points from 1970 to 2000. As a second exercise, imagine that trade in manufactured goods was balanced in 2000, so that net exports were zero, but assume that the share of manufacturing employment in 1970 and productivity growth from 1970 to 2000 were their actual values. This would raise the amount of manufactured goods produced in the United States. Manufacturing employment as a share of total nonfarm employment, however, would have been only 1 percentage point higher—14 percent, compared with the actual figure of 13 percent—if there had been balanced trade in manufactured goods in 2000.

The Effects of Domestic Outsourcing and Temporary Workers on Measurement of Manufacturing Employment

The decline in manufacturing employment in the official statistics may somewhat overstate the number of actual manufacturing production jobs that have been lost. Changing business practices in the manufacturing sector have led to both the outsourcing of nonproduction work that used to be done "in house" and the increased use of temporary workers. Manufacturing firms that once employed lawyers or accountants in their legal or finance departments might now hire outside consultants to perform these services. Counting this outsourcing as a decline in manufacturing jobs is somewhat misleading, because these workers provide services whether they are working for a manufacturing firm or an outside firm.

Similarly, manufacturing firms are increasingly using temporary workers, especially during periods of uncertain demand. Such workers, previously counted as manufacturing employees, are now counted as service-sector employees in the payroll employment data, although many of them still produce manufactured goods. The way in which employment statistics capture the increased use of outsourcing and temporary workers thus overstates the shift from manufacturing to service-providing jobs.

Much of the outsourced work is taken on by industries that make up the employment category "Professional and Business Services," which includes the temporary-help services industry. The professional and business services category covers a rapidly growing sector of the labor market, so it is likely that the understatement of manufacturing employment has increased over time. Professional and business services grew from just under 3 million employees in 1950 to over 16 million employees in 2000 (Chart 2-17). Employment in subgroups of this category increased substantially in the 1990s (Chart 2-18).

Chart 2-17 Manufacturing and Professional and Business Services Employment
Employment in professional and business services has risen dramatically since 1950.

Millions of employees

Source: Department of Labor (Bureau of Labor Statistics).
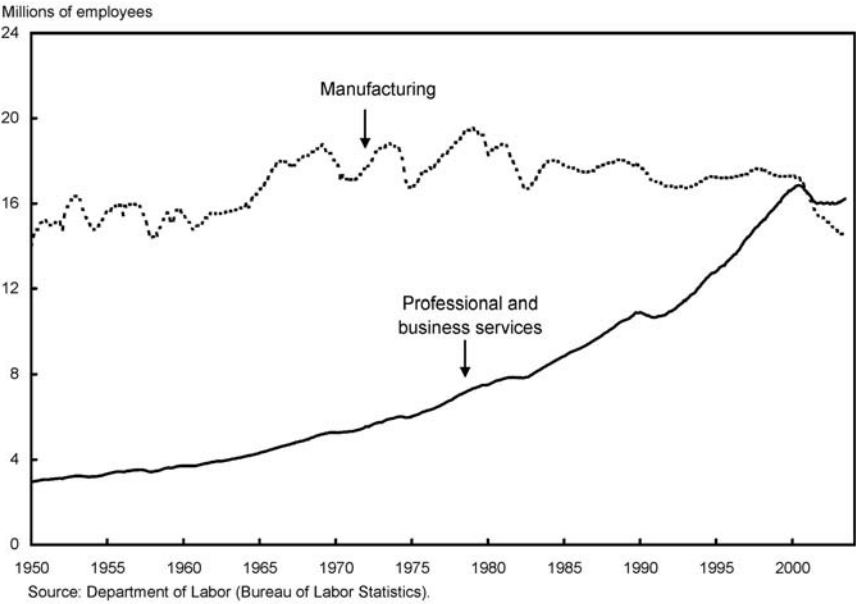
Chart 2-18 Outsourcing and Temporary-Help Services Employment
Business services employment has grown in recent years.

Millions of employees   Millions of employees

Source: Department of Labor (Bureau of Labor Statistics).
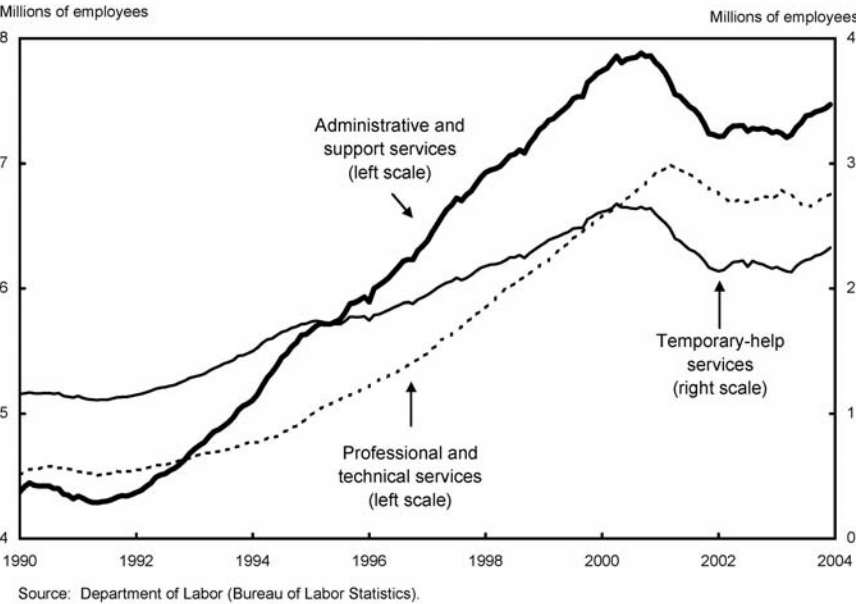
Results from academic studies can be used to estimate the understatement of employment in the manufacturing sector, bearing in mind that outsourced jobs are not necessarily comparable to permanent ones (for example, a temporary worker may receive fewer benefits than a permanent employee). One widely-cited study estimates that about one-third of all temporary-help services employees work in the manufacturing sector. If the official manufacturing employment statistics are adjusted by this amount, the decline in the level of manufacturing employment in the 1990s is eliminated.

In terms of shares of overall nonfarm employment, adjusted manufacturing shows a decline of 2.8 percentage points over the 1990s, compared with a drop of 3.1 percentage points in the reported data. If outsourcing were also included, the decline in the actual share of employment in the manufacturing sector would probably be even smaller. In other words, at least one-tenth (and perhaps as much as one-fourth) of the decline in manufacturing's share of employment over the 1990s does not reflect a loss of manufactured goods-producing jobs. Rather, it reflects how measurement conventions used to calculate employment statistics account for manufacturers' increased use of outsourced workers for tasks previously performed internally. Another example of how measurement conventions can affect, and confuse, the evaluation of the manufacturing sector is in the definition of manufacturing (Box 2-2).

Box 2-2: What Is Manufacturing?

The value of the output of the U.S. manufacturing sector as defined in official U.S. statistics is larger than the economies of all but a handful of other countries. The definition of a manufactured product, however, is not straightforward. When a fast-food restaurant sells a hamburger, for example, is it providing a "service" or is it combining inputs to "manufacture" a product?

The official definition of manufacturing comes from the Census Bureau's North American Industry Classification System, or NAICS. NAICS classifies all business establishments in the United States into categories based on how their output is produced. One such category is "manufacturing." NAICS classifies an establishment as in the manufacturing sector if it is "engaged in the mechanical, physical, or chemical transformation of materials, substances, or components into new products."

This definition is somewhat unspecific, as the Census Bureau has recognized: "The boundaries of manufacturing and other sectors… can be somewhat blurry." Some (perhaps surprising) examples of manufacturers listed by the Bureau of Labor Statistics are: bakeries, candy stores,

Box 2-2 — *continued*

custom tailors, milk bottling and pasteurizing, fresh fish packaging (oyster shucking, fish filleting), and tire retreading. Sometimes, seemingly subtle differences can determine whether an industry is classified as manufacturing. For example, mixing water and concentrate to produce soft drinks is classified as manufacturing. However, if that activity is performed at a snack bar, it is considered a service.

The distinction between non-manufacturing and manufacturing industries may seem somewhat arbitrary but it can play an important role in developing policy and assessing its effects. Suppose it was decided to offer tax relief to manufacturing firms. Because the manufacturing category is not well defined, firms would have an incentive to characterize themselves as in manufacturing. Administering the tax relief could be difficult, and the tax relief may not extend to the firms for which it was enacted.

For policy makers, the blurriness of the definition of manufacturing means that policy aimed at manufacturing may inadvertently distort production and have unintended and harmful results. Whenever possible, policy making should not be based upon this type of arbitrary statistical delineation.

Effects of the Shift to Services on Workers' Compensation

Many workers affected by the structural developments in manufacturing have experienced difficult transitions. Studies indicate that displaced workers have a significant chance of being unemployed or employed in a part-time job for some time following their job loss. Many of those who are able to find new jobs suffer earnings declines compared to previous earnings. Furthermore, workers also experience losses in earnings growth relative to what they would have had if they had remained continuously employed. Because of these effects, an often-voiced concern is that the shift toward employment in services has meant that more Americans are working in low-paying jobs.

While the shift from the manufacturing sector to service-providing sectors has been painful for many displaced from the manufacturing sector, the average effect on compensation—and in particular on new entrants into the

labor force who have chosen to work in services rather than manufacturing—has been less worrisome. Some service-providing industries pay less than some manufacturing industries, but much of the employment growth in service-providing sectors has occurred in industries with higher than average compensation. The third column of Panel A in Table 2-2 shows the total compensation per full-time equivalent employee in five service-providing industries relative to the average across industries: for example, compensation in wholesale trade in 2000 was 27 percent higher than the average (which equals 100 percent). The second column gives the change in employment from 1950 to 2000 for each industry: wholesale trade employment increased more than 4 million over this period. As Panel A reveals, four of the five service-providing industries with the largest employment increases paid compensation roughly at or above the average. Together, these five service-providing industries can explain nearly two-thirds of overall private employment growth from 1950 to 2000. Panel B of Table 2-2 shows that three of the five manufacturing industries with the highest job-loss rates paid less than the average private-sector job in 2000. For example, apparel employment fell nearly 600,000 from 1950 to 2000, and compensation of workers in the apparel industry in 2000 was only 67 percent of the average. As a result of the large increases in employment in some of these high-paying service-providing industries, the gap between compensation in service-providing sectors and manufacturing has been closing over the last couple of decades.

TABLE 2-2.— *Compensation in Selected Industries*

Industry	Change in employment, 1950 to 2000 (thousands)	Compensation per employee as percent of average for all sectors, 2000
Panel A: Service-providing		
Service-providing industries with the largest employment increases		
Retail trade	14,248	57
Business services	9,079	99
Health services	8,482	103
Finance, insurance, and real estate	5,406	158
Wholesale trade	4,259	127
Panel B: Manufacturing		
Manufacturing industries with the largest employment decreases		
Textile mill products	-715	77
Apparel and other textile products	-586	67
Primary metal industries	-491	124
Leather and leather products	-321	78
Petroleum and coal products	-91	177

Note.—Data relate to full-time equivalent employees and include some definitional changes. Not yet available are data based on the December 2003 benchmark revision of the National Income and Product Accounts.

Source: Department of Commerce (Bureau of Economic Analysis).

The Transition in Context

Individuals and communities tied to declining industries experience dislocation and distress. While many workers have made the transition from manufacturing to the service sector, the transition can be difficult. To ease it, the President has supported policies for worker retraining accounts and has extended unemployment insurance benefits when needed. The appropriate policy responses to this transition will be discussed in more detail later in this chapter. Before that, however, it is useful to place the evolution of the U.S. manufacturing sector in a broader context.

First, the shift to a relatively more service-oriented economy has involved substantial benefits for American consumers and producers. Real incomes have risen, allowing consumers to purchase more goods and services such as food, health care, transportation, and education, while measures of the quality of life and life expectancy have also increased. In addition, the growth of the service-providing sector has generated new opportunities for employment in industries such as information technology services, financial services, and entertainment.

Second, the shift of employment away from lower-productivity manufacturing toward higher-productivity manufacturing and service-providing sectors reflects economic growth and development, just as the shift away from agriculture toward manufacturing did in the last century (Box 2-3). The relative shift from manufacturing toward service-providing sectors has been shared by other advanced economies over the last few decades (Chart 2-19). Manufacturing employment declined from the mid-1990s to 2002 in a number of countries whose economies are rapidly developing, including China, Brazil, and South Korea. In fact, China, Brazil, South Korea, and Japan had steeper percentage declines in manufacturing employment over that period than the United States.

Chart 2-19 **Employment in Industry as a Percent of Total Employment**
A declining share of employment in manufacturing is common across developed economies.

Percent of total employment
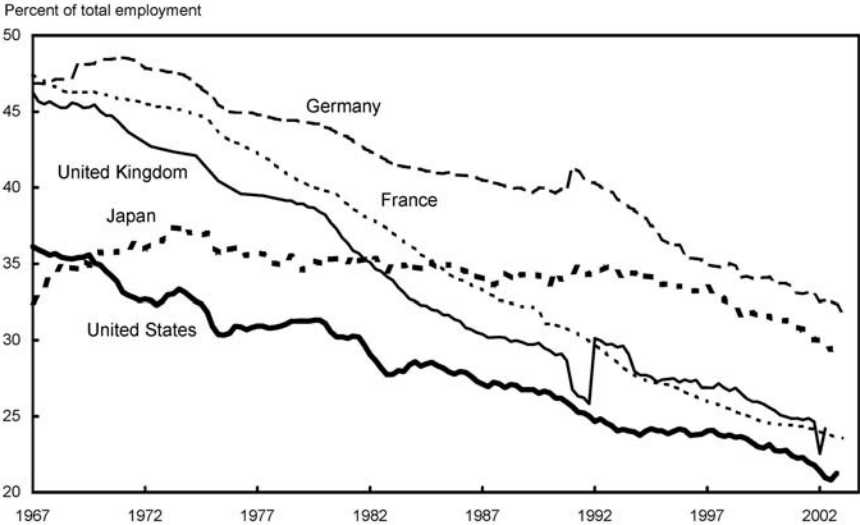
Note: Industry comprises manufacturing, mining, and construction. In 2001, manufacturing made up about two-thirds of industry in each country.
Source: Organization for Economic Cooperation and Development.

Box 2-3: The Evolution of the U.S. Agricultural Sector

The evolution of U.S. manufacturing from 1970 to 2000 mirrors, in important respects, that of U.S. agriculture from 1940 to 1970. Total real farm output increased more than 60 percent from 1940 to 1970. Over the same period, employment in farming declined nearly 6 million, or almost two-thirds of the level in 1940 (Chart 2-20). This translated into a decline in agriculture's share of total employment of 15 percentage points, from 19.4 percent in 1940 to 4.4 percent in 1970.

While the histories of agriculture and manufacturing in the United States differ in some ways, such as the prominent role of subsidies in the agricultural sector, their similarities help put the long-term story of the manufacturing sector in context.

In both sectors, a 30-year period of rapid productivity growth substantially reduced the share of the American workforce needed to meet demand for food and manufactured goods. Labor productivity in agriculture nearly quadrupled from 1940 to 1970 (Chart 2-21), a period that has been called the "second American agricultural revolution."

Box 2-3 — *continued*

This productivity boom has been attributed to the invention of new technologies, such as hybrid crop varieties, as well as the widespread application of existing technologies, such as machinery and conservation practices.

Agricultural productivity growth led to low growth in the price of food, bringing substantial benefits to American consumers and the U.S. economy as a whole and significantly improving U.S. competitiveness in world markets. Despite the mid-century expansion in the demand for agriculture's output, prices remained essentially flat. After the run-up in demand and prices during World War II and its immediate aftermath, agricultural prices increased only 4 percent from 1950 to 1970. The average price of all commodities, in comparison, increased 35 percent from 1950 to 1970 (Chart 2-22). The lack of food price inflation is mimicked by the low inflation in manufacturing in the last few decades, with a sizable benefit for American consumers in both cases.

The evolution of the agricultural sector has been good for the economy on the whole, but it meant dislocation for millions of agricultural workers—a process that continues today. Displaced farm workers faced uncertainty regarding their next job and the applicability of their skills in different sectors, just as manufacturing workers do today. The 1940s and 1950s saw the rapid growth of new industries that hired workers no longer needed on farms. Manufacturing itself likely absorbed a substantial percentage of former agricultural workers: nearly 8 million new manufacturing jobs were created between 1940 and 1970, 2 million more than the total decline in agricultural employment.

In the 1970s and 1980s, service-providing sectors likely absorbed workers not needed in manufacturing. This continued in the 1990s, as high-tech and financial services accounted for new employment growth. Looking forward, it is difficult to predict which industries will grow and require more workers. The past experience of the adjustment in agriculture suggests that market forces will continue to reshape the American workforce.

Chart 2-20 **Employment and Real Output in Agriculture**
Agricultural employment declined dramatically from 1940 to 1970, while real agricultural output increased substantially.

Millions of workers   Index, 1996=100
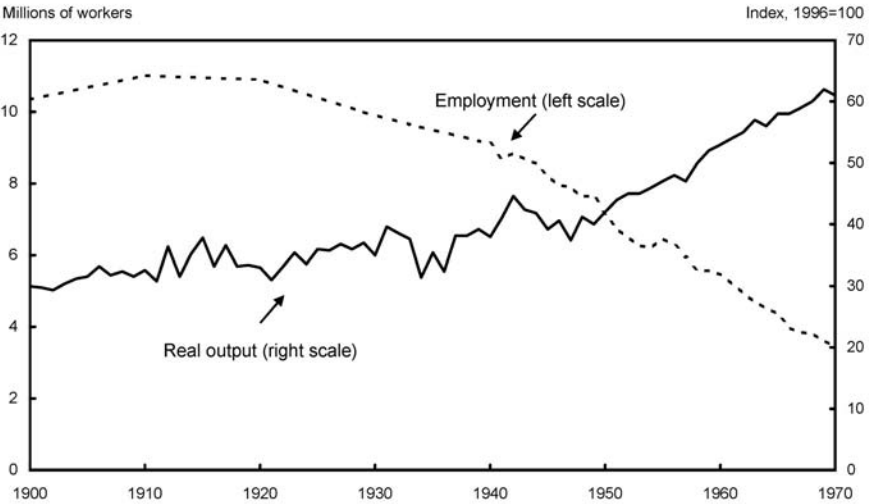
Sources: Department of Commerce (Bureau of the Census and Bureau of Economic Analysis) and Council of Economic Advisers.

Chart 2-21 **Agricultural Productivity**
Agricultural productivity on farms surged in the mid-20th century

Index, 1996=100

Sources: Department of Commerce (Bureau of the Census) and Council of Economic Advisers.
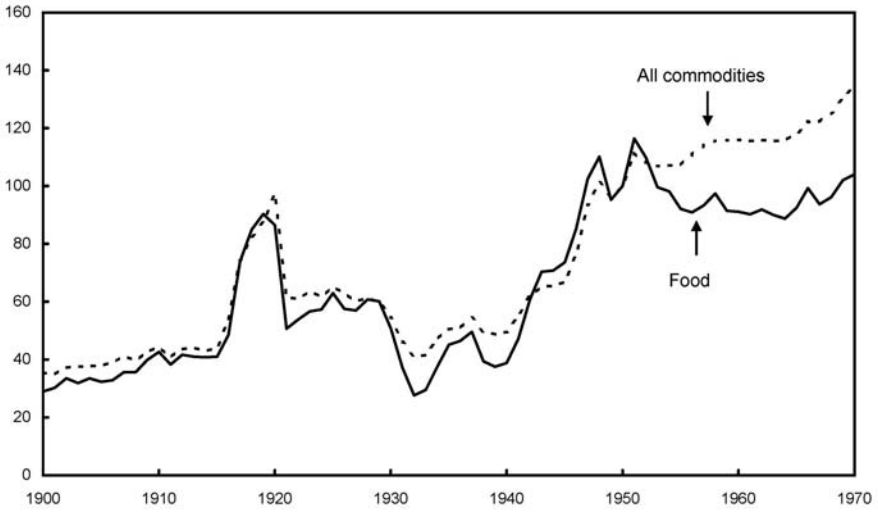
Source: Department of Labor (Bureau of Labor Statistics).

The Role of Policy

Markets operating free from government intervention will, in most cases, best allocate the Nation's resources across sectors. It is generally a mistake to target government assistance to a particular sector at the expense of other sectors, and manufacturing is no exception. That said, government policy can play a positive role. Policies targeted toward general education and training, such as the President's landmark education reforms and proposed funding to help displaced workers train for new opportunities, will help people adapt to ongoing structural changes. The President's Jobs for the 21st Century plan will support students and workers by improving high school education and strengthening post-secondary education and job training.

The short-run performance of the manufacturing sector is closely tied to fluctuations in overall economic activity. Policies that increase aggregate output and economic growth will help to improve the near-term outlook for the manufacturing sector. This Administration put forward a six-point plan for the U.S. economy in September 2003. The plan would help the manufacturing sector along with the overall economy, and it includes the following components:

Making Tax Relief Permanent

The Administration has undertaken several important fiscal measures to strengthen growth, including the 2001 tax relief program, the March 2002 stimulus package, and the May 2003 Jobs and Growth Act. These policies have already contributed to the current recovery in manufacturing. The President has proposed making provisions of the 2001 and 2003 tax cuts permanent. These include measures that lower the cost of capital and thereby encourage business investment. Capital investment makes up a relatively large share of manufacturers' costs, so a lower cost of capital provides a particularly important benefit to manufacturers. Moreover, manufacturers produce capital goods, so increased investment demand particularly benefits manufacturing firms.

Making Health Care Costs More Affordable and Predictable

The President's proposals aim to reduce frivolous litigation, help individuals save for future health expenses, and allow small businesses to pool together to purchase health coverage. Health care costs as a share of total compensation are one-third higher in manufacturing than in service-providing industries. The President's proposals will help manufacturers reduce the burden of increasing health care costs.

Reducing the Burden of Lawsuits on the Economy

The President seeks to address the burden that lawsuits impose on American businesses. For example, estimates suggest that roughly 60 companies entangled in asbestos litigation have gone bankrupt primarily because of asbestos liabilities, displacing between 52,000 and 60,000 workers.

Ensuring an Affordable, Reliable Energy Supply

Initiatives include modernizing the electricity grid and streamlining the process of acquiring permits for natural gas exploration. This is vital for manufacturing, which makes up about 15 percent of nominal GDP but accounts for around one-quarter of energy use in the United States.

Streamlining Regulations to Ensure that they are Reasonable and Affordable

Research has shown that manufacturing bore about 30 percent of the costs of regulation in the United States in 2000—nearly double its share of nominal output.

Opening International Markets to American Goods and Services

This has become particularly important for the manufacturing sector. While exports accounted for about one-sixth of American manufacturing production in 1970, they made up nearly half by 2002.

Conclusion

The manufacturing sector in the United States has undergone significant change in the last half-century. Productivity and real output in manufacturing have risen dramatically, and faster than in the economy as a whole. Productivity improvements have boosted real income in the United States. However, because Americans have spent much of their real income gains on services rather than manufactured goods, manufacturing's share of employment has declined. In the recent recession, manufacturing output and employment were hit particularly hard. The President's policies, aimed at stimulating the overall economy, easing restrictions that impede manufacturing growth, and ensuring that workers have the skills they need to be competitive, address the short-term difficulties of the sector and ensure its long-term health.

The Year in Review and the Years Ahead

The U.S. economy made notable progress in 2003. The recovery was still tenuous coming into the year, as continued fallout from powerful contractionary forces—the capital overhang, corporate scandals, and uncertainty about future economic and geopolitical conditions discussed in Chapter 1, *Lessons from the Recent Business Cycle*—still weighed against the stimulus from expansionary monetary policy and the Administration's 2001 tax cut and 2002 fiscal package. However, the contractionary forces dissipated over the course of 2003, and the expansionary forces were augmented by the Jobs and Growth Tax Relief Reconciliation Act (JGTRRA) that was signed into law at the end of May. The economy now appears to have moved into a full-fledged recovery.

This chapter reviews the economic developments of 2003 and discusses the Administration's forecast for the years ahead. The key points in this chapter are:

- Real GDP growth picked up appreciably in 2003. Growth in consumer spending, residential investment, and, particularly, business equipment and software investment appear to have increased noticeably in the second half of the year.
- The labor market began to rebound in the final five months of 2003.
- Core consumer inflation declined to its lowest level in decades.
- The Administration's forecast calls for the economic recovery to strengthen further this year, with real GDP growth running well above its historical average and the unemployment rate falling. Looking further ahead, the economy is expected to continue on a path of strong, sustainable growth.

Developments in 2003 and the Near-Term Outlook

After rising 2.8 percent during the four quarters of 2002, real GDP expanded at an average annual rate of 4.4 percent during the first three quarters of 2003. The economy appears to have gained momentum as the year went on, with annualized real GDP growth averaging close to 2½ percent during the first half of the year and more than 8 percent in the third quarter. The available data suggest solid further growth in the fourth quarter, though not as spectacular as in the third quarter. (The *Report* went to print before GDP data for the fourth quarter were available.)

The Administration expects real GDP to grow at an annual rate of 4.0 percent during the four quarters of 2004, a figure that is close to the latest Blue Chip consensus economic forecast (as of January 10, 2004). The unemployment rate, which peaked at 6.3 percent in June 2003, is projected to fall to 5.5 percent by the fourth quarter of 2004.

The pace of real GDP growth during the first three quarters of 2003 was supported by robust gains in consumption, residential investment, and defense spending. Inventory investment, in contrast, declined over the first three quarters of last year. In 2004, the composition of GDP growth is expected to shift away from government spending and toward business fixed investment and net exports. Evidence of emerging momentum in investment accumulated over the course of 2003: businesses began to hire, build inventories, and increase shipments of nondefense capital goods. In addition, expected faster growth among our trading partners and the recent decline in the exchange value of the dollar make U.S. exporters well positioned for expansion.

Much of the growth of private demand during 2003 was attributable to the effects of expansionary fiscal and monetary policy designed to counteract the lingering effects of the stock market decline, the capital overhang, worries about geopolitical developments, and concern about accounting scandals. Much stimulus remains in the pipeline in the form of refunds on 2003 tax liabilities this spring and the ongoing effects of the current low interest rates. The fiscal stimulus will not disappear suddenly. The reduction in the tax withholding schedule included in the 2003 fiscal package (JGTRRA) only began in July 2003, and households are still adjusting to these lower tax rates. Moreover, tax refunds in the first half of 2004 are expected to be higher than usual: the tax cuts were retroactive to January 2003, but last year's withholding changes generally did not capture tax savings on income earned in the first half of the year. In addition, because of the 2002 and 2003 tax cuts, businesses will be able to cut their tax liabilities by expensing 50 percent of their equipment investment (rather than depreciating the new capital) through the end of 2004. The lower tax rates, higher tax refunds, and investment expensing included in the Jobs and Growth Tax Relief Reconciliation Act are expected to reduce tax collections by about $146 billion in 2004, up from about $49 billion (or about $98 billion at an annual rate) in the second half of 2003.

Consumer Spending

Consumer spending increased briskly in 2003. Real personal consumption expenditures increased at an average annual pace of 3 percent during the first half of the year, and then surged at an annual rate of 6.9 percent in the third quarter. Data on retail sales and motor vehicle purchases through December

and services outlays through November are consistent with consumer spending remaining at a high level in the fourth quarter. As a result, real consumption growth in the second half of the year likely ran noticeably above that in the first half.

The pickup in spending growth in the second half of the year corresponded to an increase in the rate of growth of household income. After rising at an annual rate of 3.6 percent in the first half, real disposable personal income (that is, inflation-adjusted household income after taxes) jumped at an annual rate of 6.3 percent in the third quarter, boosted by tax relief, and appears to have held steady in the fourth quarter.

Wages and salaries increased moderately in the second half of the year, bolstered by the emerging recovery in the labor market. Moreover, the personal tax cuts included in the 2003 fiscal package (JGTRRA) meant that U.S. households were able to keep substantially more of their earnings. The reduction in withholding and the advance rebates of the child tax credit added $37 billion to disposable income (not at an annual rate) in the second half of the year.

Other factors also likely contributed to the strengthening of consumer spending over the course of 2003. The robust performance of equity markets and solid gains in home prices bolstered wealth. Household wealth (net financial resources plus the value of nonfinancial assets such as cars and homes) increased $2¼ trillion during the first three quarters of 2003, and it probably rose substantially further in the fourth quarter given the solid increase in broad indexes of stock prices in the last few months of the year. Consumer sentiment was depressed early in the year by the prospect of war with Iraq. Sentiment jumped in April and May following the successful resolution of major combat operations and then was little changed until November, when it picked up noticeably. By the end of the year, household sentiment was somewhat higher than it had been at the end of 2002 and much higher than it was just prior to the war with Iraq.

All told, consumption grew in line with household after-tax income during 2003. Personal saving as a fraction of disposable personal income averaged 2.3 percent in 2002 and remained at this level, on average, in the first three quarters of 2003. Swings in personal saving have contributed to movements in national saving in recent years (Box 3-1).

Growth of real consumption is expected to be lower than that of real GDP in coming years. As explained in Box 3-1, the relative flatness of the personal saving rate over the past couple of years is likely the result of offsetting forces. On the one hand, capital losses associated with the decline in the stock market from March 2000 to March 2003 probably tempered consumption (with some lag) and, in turn, caused the personal saving rate to increase. On the other hand, personal saving was likely depressed by the

Box 3-1: Personal Saving and National Saving

One important influence on the personal saving rate (the saving of the household sector divided by its after-tax income) over the past 10 years has been changes in households' wealth, although the past couple of years appear to have been an exception.

Driven by movements in stock prices, the ratio of household wealth to personal income climbed dramatically in the second half of the 1990s, peaked in early 2000, and then retreated substantially over the next two years. Economic theory suggests that increases in wealth tend to raise household spending, and decreases in wealth tend to lower household spending. This "wealth effect" often produces a negative correlation between household wealth and the personal saving rate because personal saving is defined in the national accounts as the difference between income excluding capital gains and spending (Chart 3-1).

Empirical studies suggest that an additional dollar of wealth leads to a permanent rise in the level of household consumption of about two to five cents, with the adjustment occurring gradually over a period of one to three years (the range depends on the exact specification—for example, one study found that including the components of wealth separately produces lower estimates). Such estimates of the wealth effect can explain the behavior of personal saving in the second half of the 1990s fairly well. For example, assuming that a dollar of wealth leads to an increase in consumption of three cents and that adjustment lags are typical, one would predict that the rise in wealth in the late 1990s would have caused the saving rate to decline by 4 percentage points between the end of 1994 and the end of 2000— close to the actual decline in the saving rate (ignoring the quarter-to-quarter volatility in the series).

The wealth effect also suggests that the (net) fall in wealth after 2000 would have caused a rebound in the personal saving rate of more than 2 percentage points. In fact, however, the personal saving rate has not risen materially. One potential explanation for the divergence is that households have raised consumption in anticipation that the labor market recovery will continue and, in turn, bolster income. Some of the additional consumption may have been funded through the wave of cash-out home mortgage refinancing enabled by the combination of low interest rates and technological advances that have made such transactions easier. Another possibility is that the availability of low-interest-rate loans on cars and other items has spurred households to replace cars and other durable goods earlier than they otherwise would have.

Box 3-1 — *continued*

Direct saving by households represents only part of the total saving done in the United States. Corporations also save in the form of retained earnings—the difference between after-tax profits and dividends. Most of the year-to-year variation in retained earnings stems from profits because dividend payments tend to have a fairly smooth upward trend over time. Profits rose in the early and mid-1990s, boosted by brisk productivity growth. After peaking as a share of GDP in 1997, profits fell over the next few years, owing to the 1998 global financial crisis, a catch-up of wages to productivity gains, and the economic slowdown. Retained earnings as a share of GDP also trended lower over this period. During the first three quarters of 2003, both profits and retained earnings picked up.

National saving is the sum of private saving (that is, the saving of households and corporations) and government saving (equal to the Federal budget surplus plus the state and local government budget surpluses). The saving of state and local governments tends to make a small positive contribution to government saving, but in the past few years, deteriorating fiscal conditions in states and localities have pushed their overall saving into slightly negative territory. Saving of the Federal government has declined sharply since 2000, as the recession and tax cuts have pulled down revenue, and homeland security and national defense expenditures have increased.

National saving rose (as a fraction of GDP) during the 1990s, but has fallen sharply since 2000 (Chart 3-2). As a fraction of GDP, it now stands at the low end of its range since World War II. Although both government saving and private saving are above their historic lows, the fact that they are both fairly low at the same time has led to the low level of national saving.

National saving is important because it represents the portion of our country's current income that is being set aside for investment in new capital. In particular, national saving plus the net capital inflow from abroad equals domestic investment. Greater saving and investment today boost future national income. To increase national saving, the President supports raising Federal saving by restraining Federal spending. He has also proposed Lifetime Savings Accounts and Retirement Savings Accounts, which are designed to increase incentives for households to save.

Chart 3-1 Wealth-to-Income Ratio and Personal Saving Rate

The "wealth effect" produces a negative correlation between household wealth and personal saving.

Ratio

Percent of disposable income

Source: Department of Commerce (Bureau of Economic Analysis) and Board of Governors of the Federal Reserve System.
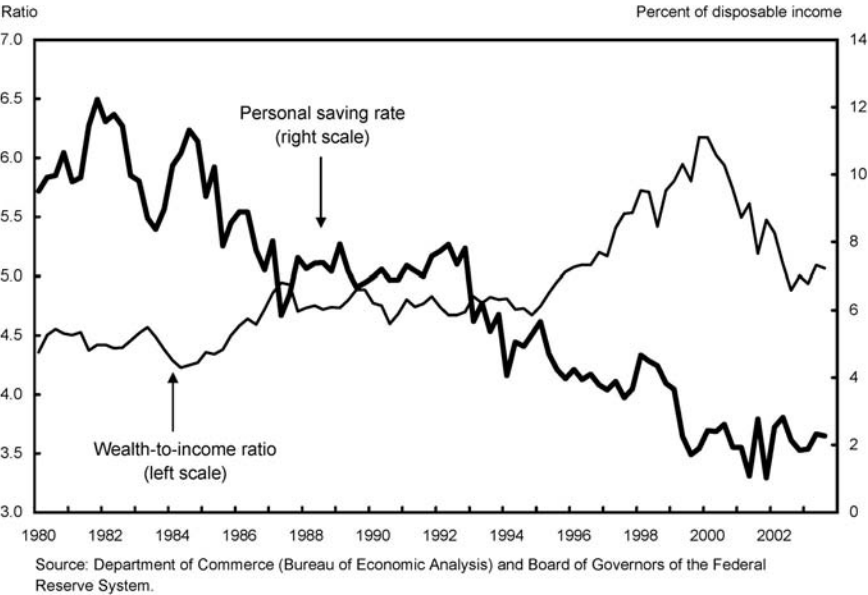
Chart 3-2 National Saving Rate

Net national saving is at an extremely low level by historical standards, reflecting both low private saving and low government saving.

Percent of GDP

Source: Department of Commerce (Bureau of Economic Analysis).
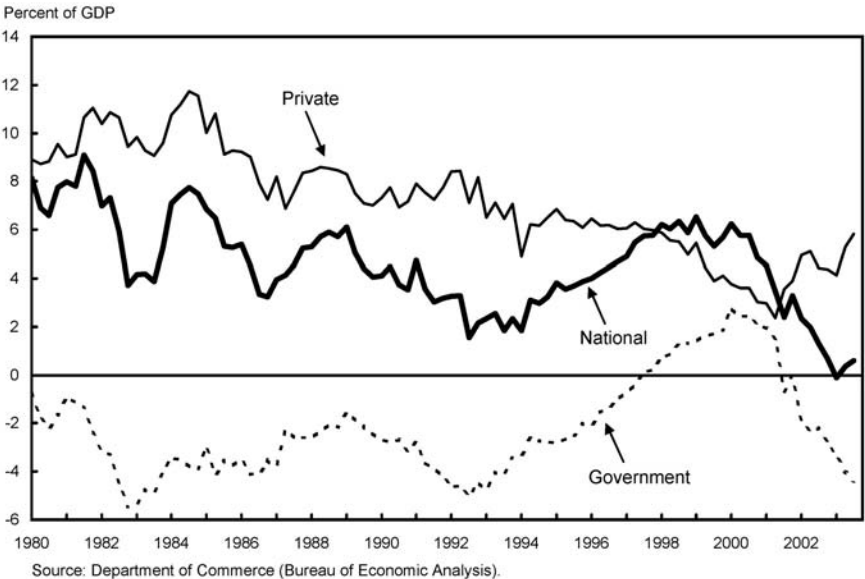
boost to consumption from low interest rates (both directly through the availability of low-interest-rate loans on durable goods and indirectly through the funds made available by cash-out mortgage refinancings). As interest rates and incomes rise over the course of the next several years, the transitory forces boosting consumption growth should dissipate, and as a result, real consumption is expected to grow more slowly than real GDP over the forecast period.

An increase in corporate contributions for defined-benefit pension plans is likely to boost the saving rate in the near term from what it might be otherwise. The Pension Benefit Guarantee Corporation (PBGC) has estimated that corporate contributions to defined-benefit plans will increase sharply above 2003 levels. Indeed, rapid increases have already begun, according to separate data included in the Employment Cost Index. The contributions raise personal income, but because these funds are not placed in the hands of employees until retirement, they seem unlikely to affect current-year consumption. As a result, they should increase the personal saving rate.

Residential Investment

The housing sector continued to show remarkable vigor in 2003, with real residential investment climbing at an average annual rate of more than 10 percent in the first three quarters of the year. Housing starts moved above the already high 2002 level to 1.8 million units in 2003, the largest number of starts since 1978. In addition, sales of both new and existing single-family homes rose to record levels.

Some of the strength in housing demand reflected the same gains in after-tax income and wealth that bolstered real consumer spending. The low levels of mortgage interest rates were another important driving force. The interest rate on new fixed-rate 30-year mortgages slipped from an average of 6½ percent in 2002 to an average of 5¾ percent in 2003. This level is the lowest in the 32 years for which comparable data are available. Indeed, according to data from the Michigan Survey Research Center, consumers' assessments of home-buying conditions remained very positive in 2003, largely because of low mortgage interest rates. As a result of the very favorable conditions in the housing sector, the U.S. home-ownership rate climbed to 68.2 percent in the third quarter of 2003—equal to its highest level on record.

During 2004, real residential investment is expected to slip lower as housing starts edge down to levels determined by long-run demographics.

Business Fixed Investment

Real business fixed investment (firms' outlays on equipment, software, and structures) turned around in 2003, posting an annualized gain of 6.2 percent during the first three quarters of the year after declines of 10.2 percent during the four quarters of 2001 and 2.8 percent during the four quarters of 2002. The acceleration during the year was noteworthy, with real investment rising at an annual rate of 12.8 percent in the third quarter and indications of further growth in the fourth quarter, compared with an average annual pace of 3.1 percent in the first half of the year. The improvement from 2002 to 2003, as well as the pickup over the course of 2003, largely reflected a strengthening in real purchases of equipment and software.

Within the equipment and software category, the largest increases occurred for certain high-tech items. Real outlays for computers increased at an annual rate of 43 percent in the first three quarters of 2003, and real spending on communications equipment, which had performed particularly poorly during the recession, rose almost 15 percent. Shipments data suggest that spending in these categories remained strong in the fourth quarter. Meanwhile, real investment in software continued its solid upward trend, rising 12 percent during the first three quarters of the year. Outlays for transportation equipment were held down by further large declines in purchases of aircraft in the first three quarters of the year. Finally, the available data suggest that real spending on equipment outside of the high-tech and transportation categories posted a solid gain over the course of 2003.

The increased momentum in business purchases of capital goods in 2003 likely reflects the factors mentioned in Chapter 1. First, with capital overhangs probably behind them, firms were poised to take advantage of further declines in prices of high-tech goods stemming from continued technological advances. Second, striking gains in productivity and falling unit labor costs bolstered corporate profits. Third, the cost of capital was held down by a number of factors, including falling prices for high-tech capital goods, but also by low interest rates, rising stock prices, and the investment incentives introduced in the Job Creation and Worker Assistance Act of 2002 (JCWAA) and expanded in the 2003 fiscal package (JGTRRA).

The Administration expects the recovery in real business investment in equipment to strengthen further this year, reflecting the acceleration in output, continued low interest rates, and the investment incentives provided by the 2002 and 2003 tax cuts. Fixed investment in equipment and structures tends to be related to the pace of growth in output (along with the cost of capital), and so the pickup in real GDP growth from 2.8 percent during the four quarters of 2002 to 4.4 percent during the first three quarters of 2003 is projected to lead to an increase in investment during 2004.

One reason for the development of a capital overhang was the lowered business expectations of the future level of output that developed just prior to the past recession. As these projections fell, the demand for investment also fell. In contrast to that period, current projections of 2004 output have been rising since mid-2003 and are expected to lead to increased demand for capital goods in the initial quarters of the forecast.

Growth in equipment investment in 2004 should be further boosted as firms pull forward spending in anticipation of the expiration of the period when businesses are able to expense (rather than depreciate) 50 percent of the value of their equipment investment. The flip side of some investment being advanced into 2004 is that investment may grow more slowly in 2005. Even so, the growth of equipment investment in 2005 is projected to be solid.

Despite the emerging recovery in spending on equipment and software, business demand for structures remained soft in 2003. High overcapacity seems to have offset the impetus imparted by low interest rates and higher cash flow. In the office sector, vacancy rates rose substantially for the third consecutive year. Vacancy rates moved still higher in the industrial sector and now stand at extremely elevated levels. The good news is that the substantial declines in total spending on structures seem to have abated. Indeed, real investment in nonresidential structures was approximately flat over the first three quarters of 2003, in contrast with a plunge of more than 25 percent during the preceding two years. Strength in oil and gas drilling and an increase in construction of general merchandise stores during the year have offset continued softness in some other sectors.

The forces that shape the outlook for business structures—the growth of output and the cost of capital—are much the same as for business equipment. However, they operate with a longer lag because of the time it takes to plan and build these structures. Investment in business structures is projected to post a small gain during 2004.

Business Inventories

Businesses began 2003 with lean inventories following a massive liquidation in 2001 and little restocking during 2002. Inventory investment was substantially negative over the first three quarters of 2003, as increases in production lagged those in final demand. The reasons for this slow response of production are unclear. Firms may have been surprised by the strength of final demand, or they may simply have been waiting for compelling evidence that a sustainable recovery was under way.

The net decline in inventories during the first three quarters of 2003 left stocks in their leanest position relative to final sales of goods and structures in at least 50 years. (This lean position results, at least in part, from efficiencies generated by just-in-time inventory-management techniques.) Stockbuilding

seems to have begun in September, however. Inventory investment appears likely to have made a positive contribution to GDP growth in the fourth quarter of 2003, and the contribution is projected to remain noticeably positive through the first half of 2004. Inventory investment is expected to plateau thereafter at a level that keeps stocks in line with rising sales throughout 2004 and 2005.

Government Purchases

Real Federal spending (consumption expenditures and gross investment) climbed at an annual rate of 8 percent during the first three quarters of 2003. The available data suggest that 2003 as a whole likely saw the largest increase in more than 30 years. The gain during the first three quarters was led by an annualized rise of 10 percent in real defense spending largely related to military operations in Iraq. Real nondefense spending rose at an average annual pace about 4 percent. This gain was less than half as large as the gain during the four quarters of 2002, when outlays were stepped up considerably for homeland security.

The defense supplemental appropriations for FY 2004, signed in November, allows for some further near-term growth in government purchases. Defense spending is projected to fall during FY 2005, and as a result, overall Federal spending is projected to edge down.

Like the Federal government, the governments of states and localities saw their tax receipts decelerate during the economic slowdown. Budgets have also deteriorated because of rising health care costs and increased demand for security-related spending. With many of these governments subject to balanced-budget rules, they have taken a variety of measures to address their fiscal imbalances, including drawing on accumulated reserves (so-called "rainy day funds"), raising taxes, and restraining spending. Real expenditures of state and local governments were little changed during the first three quarters of 2003, in contrast with an average annual gain of around 3 percent over the preceding five years. With state and local governments still under pressure, their real expenditures are projected to increase slowly during the coming year. Eventually, their fiscal situations should be improved by increases in tax revenue resulting from the strengthening of the economy.

Exports and Imports

The U.S. current account deficit as a share of GDP was little changed, on net, during the first three quarters of 2003, averaging about 5 percent. The deficit on trade in goods and services as a share of GDP also moved in a narrow range during the first three quarters. U.S. net investment income (the income paid to U.S. investors in foreign endeavors less that paid to foreign

investors in U.S. projects) was roughly flat, as both receipts from abroad and payments to foreign investors rose somewhat during the first three quarters of 2003. Real imports of goods and services have likely been restrained in recent quarters by the decline in the value of the dollar. Real imports rose at an annual rate of 1 percent during the first three quarters of 2003, a substantially slower pace than during the four quarters of 2002. Real imports of capital goods (other than autos) rose solidly, as would be expected given the recovery in U.S. investment. Real oil imports increased at a faster pace (on an annual basis) than during the four quarters of 2002. Real services imports fell markedly in the first half of 2003 but turned up in the third quarter.

America's major trading partners have recovered from the global slowdown somewhat more slowly than has the United States. For example, an index of real GDP for our G-7 trading partners increased at an average annual pace of less than 2 percent during the first three quarters of 2003. As a result, foreign demand for U.S. exports was lackluster in the first half of 2003. Real exports picked up sharply around the middle of 2003, increasing at an annual pace of 10 percent in the third quarter. The increase was led by a gain in real exports of capital goods. Even so, the level of exports remained well below its peak in 2000.

Prospects for exports over the next two years look better. Growth among the non-U.S. OECD countries is projected by the OECD Secretariat to rise 2.6 percent during the four quarters of 2004, up from a pace of 1.6 percent during 2003. Growth is expected to rise further to 2.8 percent in 2005. The expected growth in foreign markets should support growth in U.S. exports. In addition, the effect will likely be augmented by a rise in the U.S. market share of world exports owing to the effects of the 23 percent decline in the value of the dollar against major currencies from its peak in early 2002 through the end of 2003. The effect of the recent dollar decline on exports will likely take a couple of years to be fully felt.

Real imports are projected to increase along with domestic output, but the growth of real imports is likely to be slowed by the recent decline in the dollar's value relative to other currencies. On balance, real imports are projected to grow at about the same pace as GDP, on average, during the next two years. Nominal imports will increase faster than real imports because import prices will rise in reaction to the recent dollar decline. Even so, the current account deficit, which rose to about 5 percent of GDP in the first three quarters of 2003, is projected to edge up in 2004 and decline thereafter.

Overall, real net exports are expected to be approximately flat during the next year and are likely to make a positive contribution to real GDP growth thereafter. Over the next six years, the returns to foreign owners of U.S. capital are likely to grow faster than the returns to U.S. owners of foreign capital, a legacy of a long period of strong foreign investment in the United

States during the past decade. As a result, real gross national product (GNP), which includes these net foreign returns to capital, is expected to grow slower than real gross domestic product (GDP).

The Labor Market

Nonfarm payroll employment fell an average of 50,000 workers per month in the first seven months of 2003, before increasing 35,000 in August, 99,000 in September, and an average of 48,000 per month in the fourth quarter. The strengthening was experienced in most sectors. Job gains in professional and business services stepped up appreciably from the modest upward pace seen earlier in the year. Construction employment began to expand in the second quarter after two years of modest job losses, and the quarterly averages of employment in the wholesale trade, transportation, and utilities industries turned up at the end of the year. The manufacturing sector continued to shed jobs through year-end, though the pace of decline slowed, and the factory workweek climbed more than 0.5 hour, on balance, in the final five months of 2003.

The unemployment rate increased in the first half of 2003, reaching a peak of 6.3 percent in June, before falling during the second half of the year. In the fourth quarter, the unemployment rate averaged 5.9 percent, the same as it had been a year earlier. Because the labor force is constantly expanding, employment must be growing moderately just to keep the unemployment rate steady. For example, if the labor force is growing at the same rate as the population (about 1 percent per year), employment would have to rise 110,000 a month just to keep the unemployment rate stable, and larger job gains would be necessary (and are expected) to induce a downward trend in the unemployment rate.

Looking ahead, temporary-help services employment—a leading indicator for the labor market—suggests substantial further employment growth. Average growth in temporary-help services employment over a six-month period has a striking positive correlation with growth in overall employment over the subsequent six months (Chart 3-3). Statistical analysis suggests that an increase of one job in temporary-help services corresponds to a subsequent rise of seven jobs in overall employment. Employment in temporary-help services has expanded 194,000 since last April, suggesting robust growth in overall employment this year. The unemployment rate is projected to fall to 5.5 percent by the fourth quarter of 2004.

Chart 3-3 Growth in Temporary-Help Services and Overall Employment, 1990-2003
Growth in temporary-help services employment tends to lead growth in overall nonfarm employment.

6-month percent change in total nonfarm payroll employment

6-month percent change in temporary-help employment
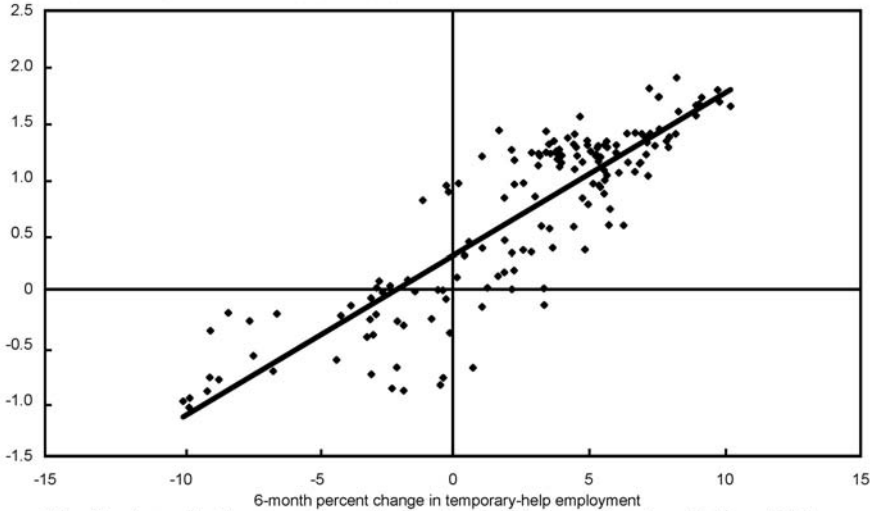
Note: The change in total employment covers the 6-month period following the 6-month period during which the change in temporary-help employment is measured.

Source: Department of Labor (Bureau of Labor Statistics).

Productivity, Prices, and Wages

The consumer price index (CPI) increased 1.9 percent over the 12 months ended in December 2003, a little below the 2.4 percent rise experienced during the same period the previous year. Consumer energy prices fluctuated markedly over the course of 2003, but ended the year 6.9 percent above their level at the end of 2002. The core CPI (which excludes food and energy) rose only 1.1 percent during 2003, considerably below the 1.9 percent increase of the previous year. This deceleration likely stems from the slack in labor and product markets last year. In addition, unit labor costs were held down by an impressive performance for labor productivity, with output per hour in the nonfarm business sector rising at an annual pace of about 6 percent during the first three quarters of the year following an increase of roughly 4½ percent during the four quarters of 2002. This pace of productivity growth is well above the annual average of just over 2 percent experienced since 1960.

Hourly compensation of workers appears to have picked up a little last year. During the 12 months of 2003, the employment cost index (ECI) for private nonfarm businesses moved up 4 percent following a 3.2 percent gain during the previous year. The wages and salaries component of the index rose 3.0 percent during 2003, slightly below the 2.7 percent increase recorded for 2002. The benefits component of the ECI, however, surged 6.4 percent over the 12 months of 2003, much faster than the 4.7 percent

pace during 2002. The increase in benefits was especially large in the first quarter of 2003, led by a jump in contributions to defined-benefit pension plans as employers began making up for losses in the value of pension fund assets. Employer-paid health premiums rose 10.5 percent during 2003, roughly the same pace as in 2002.

Core CPI inflation is expected to continue at a low level in 2004, and overall inflation is expected to be even lower as energy prices retreat further. Overall CPI inflation is projected to fall to 1.4 percent during the four quarters of 2004—close to the past year's pace of core inflation. With the unemployment rate expected to average 5.6 percent for the year as a whole (above our estimated 5.1 percent midpoint of the range of rates consistent with stable inflation) the level of slack—although less than in 2003—is still projected to hold down inflation during 2004. Also keeping inflation in check is the recent rapid pace of—and solid near-term prospects for—productivity growth. Offsetting this effect is the somewhat higher pace of import-price inflation (resulting from the recent dollar decline) and the quicker pace of GDP growth. Over the next five years, CPI inflation is expected to edge up, eventually flattening out at 2.5 percent, a level that is identical to the consensus forecast.

The path of inflation as measured by the GDP price index is similar, but a bit lower throughout the projection period. Inflation as measured by the GDP price index is projected fall to 1.2 percent during the four quarters of 2004, the same as the 1.2 percent pace of the core GDP price index during the first three quarters of 2003. GDP price inflation is projected to increase slowly thereafter—roughly parallel to the rise in CPI inflation.

The wedge between the CPI and the GDP measures of inflation has important implications for the Federal budget and budget projections. A larger wedge reduces the Federal budget surplus because cost-of-living adjustments for Social Security and other indexed programs rise with the CPI, whereas Federal revenue tends to increase with the GDP price index. For a given level of nominal income, increases in the CPI also cut Federal revenue because they raise income tax brackets and affect other inflation-indexed features of the tax code. Of the two indexes, the CPI tends to increase faster in part because it measures the price of a fixed market basket. In contrast, the GDP price index increases less rapidly than the CPI because it reflects the choices of households and businesses to shift their purchases away from items with increasing relative prices and toward items with decreasing relative prices. In addition, the GDP price index includes investment goods, such as computers, whose relative prices have been falling rapidly. Computers, in particular, receive a much larger weight in the GDP price index (0.8 percent) than in the CPI (0.2 percent).

During the eight years ended in 2002, the wedge between inflation in the CPI-U-RS (a version of the CPI designed to be consistent with current methods) and the rate of change in the GDP price index averaged 0.5 percentage point per year. With the core CPI and the core GDP price index both increasing at about a 1¼ percent pace during the past year, inertia suggests that the near-term wedge will be only about 0.2 percentage point in 2004. The wedge is expected to widen eventually to its recent mean of 0.5 percent by 2009.

Financial Markets

Stock prices skidded early in the year, but rallied in March and have been on a solid uptrend since then. During the 12 months of 2003, the Wilshire 5000 index—a broad measure of stock prices—rose 29 percent. An increase of this magnitude has not been seen since 1997. High-tech stocks did even better; for example, the Nasdaq index, which is heavily weighted toward high-tech industry, rose 50 percent during 2003. Nearly two-thirds of the rise in broad measures of stock prices occurred after the President signed the 2003 tax cut (JGTRRA) in late May; the Act reduced marginal tax rates on dividends and capital gains and thus likely contributed to the robust performance of stock prices.

Following a large decline in 2001, and a smaller one in 2002, the interest rate on 91-day Treasury bills fell an additional 29 basis points in 2003 and ended the year at 0.9 percent. These reductions reflected the Federal Reserve's efforts to stimulate the economy, leaving real short-term rates (that is, nominal rates less expected inflation) slightly negative. Following market-based expectations of interest rates (derived from rates on Eurodollar futures), the Administration does not expect real rates this low to persist once the recovery becomes firmly established, and nominal Treasury bill rates are projected to increase gradually. Long-term interest rates fell sharply last spring and then rebounded in the summer. For the year as a whole, long-term Treasury rates were about unchanged, but corporate interest rates dropped a bit as the spread over Treasury rates narrowed. The Administration projects that the yield on 10-year Treasury notes, which averaged 4.3 percent in December 2003, will edge up gradually next year, consistent with the path of short-term Treasury rates.

The Long-Term Outlook

The economy could well grow faster than in the projection presented here, as the long-run benefits from the full reductions in marginal tax rates are felt. These should lead to higher labor force participation than would

occur otherwise, more entrepreneurial activity, and greater work effort by highly productive individuals. The Administration, however, chooses to adopt conservative economic assumptions that are close to the consensus of professional forecasters. As such, the assumptions provide a prudent and cautious basis for the budget projections.

Growth in Real GDP and Productivity over the Long Term

The economy continues to display supply-side characteristics favorable to long-term growth. Productivity growth has been remarkable, and inflation remains low and stable. As a result of stimulative fiscal and monetary policies, real GDP is expected to grow faster than its 3.1 percent potential rate during the next four years. The Administration forecasts that real GDP growth will average 3.7 percent at an annual rate during the four years from 2003 to 2007—in line with the consensus projection. Because this pace is somewhat above the assumed rate of increase in productive capacity, the unemployment rate is projected to decline over this period. In 2008 and 2009, real GDP growth is projected to continue at its long-run potential rate of 3.1 percent, and the unemployment rate is projected to be flat at 5.1 percent (Table 3-1).

The growth rate of the economy over the long run is determined by its supply-side components, which include population, labor force participation, productivity, and the workweek. The Administration's forecast for the contribution of different supply-side factors to real GDP growth is shown in Table 3-2.

TABLE 3-1.—*Administration Forecast*[1]

Year	Nominal GDP	Real GDP (chain-type)	GDP price index (chain-type)	Consumer price index (CPI-U)	Unemploy-ment rate (percent)	Interest rate, 91-day Treasury bills (percent)	Interest rate, 10-year Treasury notes (percent)	Nonfarm payroll employ-ment (millions)
	Percent change, fourth quarter to fourth quarter				Level, calendar year			
2002 (actual)	4.3	2.9	1.3	2.2	5.8	1.6	4.6	130.4
2003	5.8	4.2	1.5	2.0	6.0	1.0	4.0	130.1
2004	5.2	4.0	1.2	1.4	5.6	1.3	4.6	132.7
2005	4.9	3.4	1.4	1.6	5.4	2.4	5.0	136.3
2006	5.0	3.3	1.6	1.9	5.2	3.3	5.4	138.6
2007	5.2	3.3	1.8	2.2	5.1	4.0	5.6	140.6
2008	5.2	3.1	2.0	2.5	5.1	4.3	5.8	142.5
2009	5.2	3.1	2.0	2.5	5.1	4.4	5.8	144.4

[1] Based on data available as of December 2, 2003.

Sources: Council of Economic Advisers, Department of Commerce (Bureau of Economic Analysis), Department of Labor (Bureau of Labor Statistics), Department of the Treasury, and Office of Management and Budget.

TABLE 3-2.—*Accounting for Growth in Real GDP, 1960-2009*[1]

[Average annual percent change]

Item	1960 Q2 to 1973 Q4	1973 Q4 to 1990 Q3	1990 Q3 to 2003 Q3	2003 Q3 to 2009 Q4
1) Civilian noninstitutional population aged 16 or over	1.8	1.5	1.2	1.1
2) Plus: Civilian labor force participation rate	.2	.5	-.1	-.1
3) Equals: Civilian labor force [2]	2.0	2.0	1.1	1.0
4) Plus: Civilian employment rate [2]	.0	-.1	.0	.2
5) Equals: Civilian employment [2]	2.0	1.9	1.1	1.2
6) Plus: Nonfarm business employment as a share of civilian employment [2][3]	.1	.1	-.1	.6
7) Equals: Nonfarm business employment	2.1	2.0	1.0	1.8
8) Plus: Average weekly hours (nonfarm business)	-.4	-.4	-.1	.0
9) Equals: Hourall persons (nonfarm business)	1.7	1.7	.9	1.8
10) Plus: Output per hour (prods of uctivity, nonfarm business) ...	2.8	1.4	2.5	2.1
11) Equals: Nonfarm business output	4.6	3.1	3.4	3.9
12) Plus: Ratio of real GDP to nonfarm business output [4]	-.3	-.2	-.4	-.5
13) Equals: Real GDP	4.2	2.9	3.0	3.4

[1] Based on data available as of December 2, 2003.
[2] Adjusted for 1994 revision of the Current Population Survey.
[3] Line 6 translates the civilian employment growth rate into the nonfarm business employment growth rate.
[4] Line 12 translates nonfarm business output back into output for all sectors (GDP), which includes the output of farms and general government.

Note.—The periods 1960 Q2, 1973 Q4, and 1990 Q3 are business cycle peaks.
Detail may not add to totals because of rounding.

Sources: Council of Economic Advisers, Department of Commerce (Bureau of Economic Analysis), and Department of Labor (Bureau of Labor Statistics).

The Administration expects nonfarm labor productivity to grow at a 2.1 percent average annual pace over the forecast period, virtually the same as that recorded during the 43 years since the business-cycle peak in 1960. The projection is notably more conservative than the roughly 4½ percent average annual rate of productivity growth since the output peak in the fourth quarter of 2000. After such an extraordinary surge, a period of slower productivity growth is likely as firms shed their hesitancy to hire. In addition, the slower pace of productivity assumed in the forecast reflects the Administration's view that in the absence of a good explanation for the recent acceleration, it is wiser to base the productivity forecast on longer-term averages.

In addition to productivity, growth of the labor force (also shown in Table 3-2) is projected to contribute 1.0 percentage point per year to growth of potential output on average through 2009. Labor force growth results from growth in the working-age population and changes in the labor force partic-ipation rate. The Bureau of the Census projects that the working-age population will grow at an average annual rate of 1.1 percent through 2009—roughly the same pace as during the years between 1990 and 2003. The last year in which the labor force participation rate increased was 1997,

so the long-term trend of rising participation appears to have come to an end. Since then, the participation rate has fallen at an average 0.2 percent annual pace—although some of the decline in 2001 and 2002 probably resulted from the recession-induced decline in job prospects. In 2003, the baby-boom cohort was 39 to 57 years old, and over the next several years the boomers will be moving into older age brackets with lower participation rates. As a result, the labor force participation rate is projected to edge down an average of 0.1 percent per year through 2009. The decline may be greater, however, after 2008, which is the year that the first baby boomers (those born in 1946) reach the early-retirement age of 62.

In sum, potential real GDP is projected to grow at a 3.1 percent annual pace, slightly above the average actual pace since 1973 of 3.0 percent. Actual real GDP growth during the six-year forecast period is projected to be slightly higher, at 3.4 percent, because the civilian employment rate (line 4 of Table 3-2) makes a small (0.2 percentage point) and transitory contribution to growth through 2007 as the unemployment rate falls. This contribution then ends as the unemployment rate stabilizes at 5.1 percent.

Interest Rates over the Long Term

The gradual increase in the interest rate on 91-day Treasury bills is projected to continue through 2009. The rate is expected to reach 4.4 percent by 2009, at which date the real interest rate on 91-day Treasury bills will be close to its historical average. The projected path of the interest rate on 10-year Treasury notes is consistent with that on short-term Treasury rates. By 2008, this yield is projected to be 5.8 percent, 3.3 percentage points above expected CPI inflation—a typical real rate by historical standards. By 2009, the projected term premium (the difference between the 10-year interest rate and the 91-day rate) of 1.4 percentage points is in line with its historical average.

The Composition of Income over the Long Term

A primary purpose of the Administration's economic forecast is to estimate future government revenue, which requires a projection of the components of taxable income. The Administration's income-side projection is based on the historical stability of the long-run labor and capital shares of gross domestic income (GDI). During the first three quarters of 2003, the labor share of GDI was on the low side of its historical average. From this jump-off point, it is projected to rise to its long-run average and then remain at this level over the forecast period. (The income share projections are consistent with data available through December 2, 2003. They exclude any effects of the later comprehensive revision to the National Income and Product Accounts.) The labor share consists of wages and salaries, which are taxable, employer contributions for employee pension and insurance funds (that is,

fringe benefits), which are not taxable, and employer contributions for government social insurance. The Administration forecasts that the wage and salary share of compensation will decline while employer contributions for employee pension and insurance funds grow faster than wages. This pattern has generally been in evidence since 1960 except for a few years in the late 1990s. During the next five years, the fastest growing components of employer contributions for employee pension and insurance funds are expected to be employer-paid health insurance and contributions for defined-benefit pension plans.

The capital share (the complement of the labor share) of GDI is expected to fall before leveling off at its historical average. Within the capital share, a near-term decline in depreciation (an echo of the decline in short-lived investment during 2001 and 2002) helps boost corporate economic profits, which in the third quarter 2003 were noticeably above their post-1973 average of about 8 percent of GDI. The share of corporate economic profits in GDI is projected to be bolstered in 2004 by the strong recent productivity growth together with stable gains in hourly compensation, and an expected decline in depreciation. From 2005 forward, the profit share is expected to slowly decline back to its historical average of about 8 percent. The projected pattern of book profits (known in the national income accounts as "profits before tax") reflects the 30 percent expensing provisions of the Job Creation and Worker Assistance Act of 2002 and the 50 percent expensing provisions of the Jobs and Growth Tax Relief Reconciliation Act of 2003. These expensing provisions reduce taxable profits from the third quarter of 2001 through the fourth quarter of 2004. The expiration of the expensing provisions increases book profits thereafter, however, because those investment goods expensed during the three-year expensing window will have less remaining value to depreciate thereafter. The share of other taxable income (the sum of rent, dividends, proprietors' income, and personal interest income) is projected to fall, mainly because of the delayed effects of past declines in long-term interest rates, which reduce personal interest income during the projection period.

Conclusion

The Administration's policies have been a key force shaping recent economic developments and the prospects for economic growth in coming years. The policies are designed to enhance U.S. economic growth, not just maintain it. The remaining chapters of this *Report* illustrate the ways in which pro-growth economic policies can improve economic performance by striking a balance between encouragement and regulation of firms, by reducing barriers to trade, and by reducing tax-based disincentives to economic activity.

Tax Incidence:
Who Bears the Tax Burden?

The study of *tax incidence* is the economic study of which taxpayers bear the burden of a tax. This question is of considerable importance to policy makers, who want to know whether the distribution of the tax burden (between rich and poor, capital and labor, consumers and producers, and so on) meets their criteria for fairness.

Distributional tables showing the tax burdens borne by different income groups are an important application of incidence analysis. The Joint Committee on Taxation (JCT) and the Department of the Treasury prepare distributional tables for the existing tax system and for some proposed and adopted tax changes. The Congressional Budget Office (CBO) prepares such tables for the existing tax system. In addition to these official analyses, some private groups also publish distributional tables.

When used properly, distributional tables can contribute to informed decision making on the part of citizens and policy makers. Unfortunately, mainstream economic analysis suggests that these tables do not always accurately describe who bears the long-run burden of certain taxes. This problem does not arise from bias or lack of economic knowledge on the part of the economists who prepare these tables. Instead, it reflects resource and data limitations, uncertainty about some of the economic effects of taxes, and variations in the time frame considered by the analyses. Nevertheless, the shortcomings of distributional tables can lead to misperceptions of the impact of tax changes.

This chapter discusses some of the ways in which distributional tables can be improved. The key points in this chapter are:

- The actual incidence of a tax may have little to do with the legal specification of its incidence. Official distributional tables recognize this fact in many contexts, but not in all of them.
- In the long run, a large part of the burden of capital taxes is likely to be shifted to workers through a reduction in wages. Analyses that fail to recognize this shift can be misleading, suggesting that higher income groups bear an unrealistically large share of the long-run burden of such taxes.

To begin, it is useful to review the basic economic principles of tax incidence and apply them to different types of taxes.

Theory of Tax Incidence

One crucial finding in the study of tax incidence is that the *economic incidence* of a tax (the identity of the person who bears the burden of the tax) can be completely different from its *statutory* or *legal incidence* (the identity of the person upon whom the law officially imposes the tax). In other words, the person who is legally responsible for paying the tax may not be the one who actually bears the burden of the tax. As explained below, the incidence of a tax depends upon the law of supply and demand, not the laws of Congress.

Another crucial principle is that only people can pay taxes. Businesses and other artificial entities cannot pay taxes. Although the corporate income tax is legally imposed on firms that are organized as corporations, the actual burden of the tax can fall only on people—perhaps the firm's owners, or its employees, or its customers—but certainly not on a legal artifact such as a corporation. Similarly, although the estate tax is legally imposed on the estate, the burden of the tax can fall only on people—perhaps the decedent who left the estate, perhaps the heirs, perhaps other people—but not the estate, which is merely a legal construct established to sort through the ownership of the decedent's assets.

It is simplest to first discuss the incidence of a simple *excise tax*, a tax levied on a specific good or service. As explained below, the key insights from this analysis can be extended to apply to other types of taxes.

Incidence of an Excise Tax

Consider a tax on apples. Suppose that when there is no tax, the price of apples is $1. Now, suppose that the government imposes a 10-cent excise tax on apples and that the producers are legally responsible for paying this tax. Do producers actually bear the economic burden of the tax?

The answer depends on what happens to the price of apples. If the price remains unchanged, producers bear the economic burden (the economic incidence of the tax is the same as the legal incidence). Consumers pay $1, the same as before, and suffer no burden. Producers, after collecting $1 from the consumers, must pay 10 cents to the government, so they clear only 90 cents. Alternatively, if the price rises by the amount of the tax, from $1 to $1.10, consumers bear the burden. Although they do not send any money to the government, they pay 10 cents more per apple than they did without the tax. The producers bear no economic burden, even though they are legally responsible for paying the tax. After collecting $1.10 from consumers and sending 10 cents to the government, they still clear $1, as they did without the tax. In this case, economists say that the producers *shift* the burden of the tax to consumers. To consider another possibility, if the price of apples rises by 5 cents, to $1.05, consumers and producers share the

burden equally. Consumers bear a 5-cent burden because they pay $1.05 for each apple, compared to the $1 that they paid without the tax. Producers bear a 5-cent burden because they clear only 95 cents per apple, compared to the $1 they cleared without the tax: they collect $1.05 from consumers, but send 10 cents to the government.

As these examples show, the division of the tax burden between consumers and producers depends on what happens to the price of apples. When prices are free to adjust, they are likely to be determined by the law of supply and demand. If the price of apples was $1 with no tax, then the number of apples consumers wanted to buy at that price must have equaled the number of apples that producers wanted to sell at that price.

What happens when the 10-cent excise tax is imposed? It depends on how responsive consumers and producers are to changes in the prices they pay or receive. The relevant questions are: How many fewer apples do producers sell if the amount they clear per apple declines? How many fewer apples do consumers buy if the amount they pay per apple rises?

For example, suppose that producers are four times more responsive to price changes than consumers. Then, producers face a price change that is one-fourth as large as that faced by consumers. The 10-cent tax causes the price to rise from $1 to $1.08, putting an 8-cent burden on consumers and a 2-cent burden on producers. At that price, the number of apples consumers want to buy falls by the same amount as the number that producers want to sell. Alternatively, if consumers were four times more responsive than producers, then producers would bear 8 cents of the burden and consumers would bear only 2 cents.

The group that is less responsive bears more of the burden of the tax. The group that is more responsive escapes much of the burden because it responds to the tax, abandoning the taxed activity when threatened with a tax burden. The price-responsiveness of each group depends upon its flexibility. Do producers have good alternatives (in the form of other industries in which they can produce)? Do consumers have good alternatives (in the form of other products they can buy)?

The answers vary across products, types of producers (such as workers and owners of capital), and time frames. If the excise tax applied only to Granny Smith apples, consumers could switch to other, untaxed, kinds of apples. If it applied to all apples, consumers would have somewhat less flexibility. Some workers may have skills specific to the apple industry. Other workers may be more flexible because their skills are more general; they could avoid bearing the tax burden by finding a job in another industry. The owners of capital employed in a taxed industry may bear a significant short-run burden because the buildings and equipment in the industry may be designed specifically for its use and the owners may have little ability to move those

resources elsewhere. In the long run, though, capital can leave the taxed industry: as buildings and equipment depreciate in the taxed industry, *new* buildings and equipment are constructed in other industries.

A similar logic applies if the product is subsidized rather than taxed. The group that is more responsive receives the smaller benefit because the subsidy prompts new members of that group to enter the market and compete away the benefits of the subsidy. Conversely, the group that is less responsive receives the greater benefit from the subsidy because little entry occurs.

Because the incidence of an excise tax depends upon the relative flexibility of consumers and producers, the burden may not always fall where the Congress intends. When the Congress imposed a "luxury" tax on yachts in 1991, for example, it intended the wealthy purchasers of yachts to bear the burden. Such purchasers, however, may be quite responsive to price because there are many alternative goods that they can purchase (expensive cars and jewelry, for example). If this is so, then a significant part of the burden of a yacht tax may fall on workers in the industry, who may be less well-off than owners of yachts. Indeed, after the tax was introduced, production and employment in the boat industry fell, leading some observers to claim that workers were bearing much of the burden of the tax. Although the validity of this claim cannot be conclusively determined (the industry's decline may have been caused by the 1990-1991 recession rather than the tax), the Congress responded to these concerns by repealing the tax in 1993.

Legal Incidence Is Unimportant

As long as prices can freely adjust, the economic incidence of a tax does not depend on the legal incidence. Suppose that, in the above example, the government imposes the 10-cent excise tax on apple consumers rather than apple producers. Consumers then must make the tax payment to the government, in addition to the price they pay to producers.

Because producers are four times more price-responsive than consumers, the price received by producers must still fall by 2 cents and the price paid by consumers must still rise by 8 cents. Despite the legislative change, that is still the only outcome that keeps the number of apples producers want to sell equal to the number that consumers want to buy. If the tax is legally imposed on producers, they shift 8 cents of the burden to consumers. If it is legally imposed on consumers, they shift 2 cents of the burden to producers.

Given that the price can freely adjust, it should not be surprising that the final outcome is unchanged. It is irrelevant whether the tax collector stands next to consumers and takes 10 cents from them when they buy an apple or stands next to producers and takes 10 cents from them when they sell an apple. It does not matter whether the consumer puts a dime in a bowl marked "taxes" or hands the dime to the producer who puts it in the same bowl.

Applied Distributional Analysis of Excise Taxes and Subsidies

The legal incidence of Federal excise taxes is sometimes placed on consumers, sometimes on manufacturers, and sometimes on other producers or importers. In most cases, this legal incidence rightly receives little attention. In accordance with the economic theory of tax incidence, the JCT and Treasury economists preparing distributional tables uniformly ignore the legal incidence of conventional excise taxes. The JCT generally allocates excise tax burdens to consumers. Treasury follows a similar, but more elaborate, approach.

These approaches are reasonable, since consumers are likely to bear much of the long-run burden of most excise taxes. In the long run, most producers are flexible, or price-responsive, because they can switch to other industries. Consumers are likely to have less flexibility, except in special cases where there are good substitutes for the product being taxed.

The theory of incidence also applies to more-subtle excise subsidies, such as those included within the individual income tax. The income tax law grants tax reductions for purchasers of various products—for example, an itemized deduction for medical expenses, a credit for electric cars, and the Hope and Lifetime Learning credits for the costs of higher education. The economic benefits of these provisions are likely to be divided between consumers and producers, with the greater benefit going to the group that is less price-responsive. The long-run benefits are likely to go largely to consumers, because they are likely to be less price-responsive than producers. Official distributional analyses generally allocate these income tax reductions to the consumers.

The basic insight that tax burdens fall more heavily on groups that are less flexible can be applied to a wide range of taxes. The remainder of this chapter applies this framework to payroll taxes, taxes on capital, and estate and gift taxes.

Payroll Taxes

The largest Federal payroll tax, earmarked to finance Social Security and Medicare Part A, is imposed at a 15.3 percent rate on the first $87,900 of earnings and at a 2.9 percent rate on earnings above that amount. A much smaller Federal payroll tax, earmarked to finance unemployment compensation, is imposed at a 0.8 percent rate on the first $7,000 of earnings. The legal incidence of the Social Security-Medicare tax is divided equally between employers and employees. The legal incidence of the Federal unemployment compensation tax is placed entirely on employers.

With a payroll tax, the product being taxed is labor and its price is the wage rate. Applying the insights obtained from the analysis of excise taxes, the relevant question is whether firms' demand for labor or workers' supply of labor is more responsive to changes in the wage rate. In the long run, it is likely that firms are more responsive, or flexible, particularly in a global economy in which they can relocate abroad. This conclusion implies that employees bear most of the payroll tax burden, a result supported by empirical studies. In other words, wages paid to employees are lower by an amount roughly equal to the employers' part of the payroll tax. In accord with this conclusion, official distributional analyses generally assign the full burden of payroll taxes to employees. The primary controversy in this area concerns whether the distributional analysis should also include the Social Security benefits that are financed by the payroll tax (Box 4-1).

Much of the individual income tax is also imposed on labor income. Based on the above discussion, the burden of the individual income tax on labor, like that of payroll taxes, should also fall on workers. Official distributional analyses generally allocate the individual income tax on labor income to workers.

Some taxes on and subsidies to labor income are more subtle. The income tax laws deny firms their normal business-expense deductions for some payments of labor income. For example, under certain circumstances, firms cannot deduct salaries greater than $1,000,000 per year paid to senior executives or some "golden-parachute" payments made to executives in connection with corporate takeovers. Because of this denial of deductibility, the firm pays a tax on these labor income payments, in addition to the regular tax on its owners' net income. This tax operates as an additional payroll tax legally imposed on employers, although of a much narrower scope than the payroll taxes discussed above. On the other hand, the income tax laws allow firms to claim tax credits for some other payments of labor income. Examples include the work opportunity tax credit, the welfare-to-work credit, the empowerment zone employment credit, and the Indian employment credit. (The work opportunity and welfare-to-work credits expired on December 31, 2003, but may be reinstated by future legislation.) In economic terms, these credits are subsidies to labor.

The fact that these taxes and subsidies are implemented as changes in the employer's (rather than the employee's) income tax does not change their economic incidence. The fact that they apply only to employees in specific jobs or in specific locations or to those receiving specific forms of compensation, however, may change their incidence. Because employees can, to some extent, change their jobs, locations, and forms of compensation, the flexibility of the employee may be greater than was assumed in the discussion of general taxes on labor income. As a consequence, the division of the burdens

Box 4-1: Social Security and Transfer Payments in Distributional Tables

In addition to collecting taxes, the government makes transfer payments to households. The net burden that the fiscal system imposes on households is better measured by looking at tax payments minus transfer payments received rather than by looking at tax payments alone. Official distributional tables, however, usually show only tax payments. They do not tabulate the distribution of transfer payments, except sometimes the refundable tax credits that are administered through the individual income tax, such as the Earned Income Tax Credit. For example, if a household has $20,000 of wage income, pays $5,000 in taxes, and receives transfer payments of $2,000, the distributional table would report that the household bears a $5,000 tax burden, overlooking the fact that its net burden imposed by the fiscal system is only $3,000. In some tables, transfer payments are included in the income measure that is used to classify households into different income groups—in this example, the household might be classified as having income of $22,000 rather than $20,000. But, the transfer payments are not netted against the taxes in measuring the household's burden.

This practice induces a potential political bias because policy makers receive "distributional credit" for helping the poor only if they do so through the tax system rather than through transfer payments.

The omission of government benefits from distributional tables may provide a misleading picture of Social Security. Official distributional tables generally show that the Social Security payroll tax imposes a smaller burden, as a fraction of income, on high income groups than on lower and middle income groups. However, if the analysis were expanded to include the Social Security benefits financed by the payroll tax, it would likely reveal that high income groups bear a larger net burden, as a fraction of income, than some other groups. Thus, distributional tables might be more accurate if these benefits were included in some manner. One possibility would be to treat the present value of the future benefits accrued by a worker each year as an offset to his or her payroll tax liability.

or benefits between employees and firms is not clear. Official distributional analyses generally allocate the burdens and benefits of these provisions in the same manner as firms' other income tax payments. (As discussed below, these analyses differ in their treatment of the corporate income tax.)

Taxes On Capital Income

The Federal tax system imposes taxes on capital income. Capital income generated by corporations is generally subject to the corporate income tax. Capital income received by individuals is generally subject to the individual income tax.

Many observers view capital income taxes as highly progressive because capital income is highly concentrated. However, economic analysis suggests that capital income taxes are particularly likely to be shifted, especially in the long run. Taxes imposed on owners of capital in one sector of the economy may be shifted to the owners of capital in other sectors. More importantly, capital income taxes may be partly shifted to workers through a reduction in wages. The extent of shifting differs across time horizons because savers (who provide capital and earn capital income) are more flexible in the long run than in the short run.

Shifting Across Sectors

Even if a tax is imposed on capital income in one sector of the economy, it is likely that owners of capital in all sectors bear the same economic burden in the long run. To see why, note that if capital is mobile across sectors, after-tax rates of return must be equalized across sectors, after adjustment for risk. Suppose that an economy contains two sectors and that, when there are no taxes, capital earns a 6 percent rate of return in each sector. Now, suppose that a 50 percent tax is imposed on capital income in one sector, while no tax applies in the other sector. In the very short run, capital in the taxed sector earns an after-tax return of only 3 percent, while capital in the tax-exempt sector earns an after-tax return of 6 percent. At this point, only the owners of capital in the taxed sector bear the burden.

This state of affairs cannot continue. Owners of capital in the taxed sector will move their money out of that sector and begin investing in the tax-exempt sector. As they do so, two things happen. First, the before-tax rate of return rises in the taxed sector as capital becomes more scarce. Second, the before-tax rate of return falls in the tax-exempt sector as capital becomes more plentiful. This movement continues until investors are indifferent between the two sectors, which happens when after-tax rates of return are once again in balance. For example, after a certain amount of capital has relocated, the before-tax rate of return in the taxed sector may rise from 6 to 8 percent, while the before-tax rate of return in the tax-exempt sector may fall from 6 to 4 percent. At this point, investors in both sectors earn the same 4 percent after-tax rate of return. Because all investors initially earned 6 percent and now earn 4 percent, they all bear the same burden from the tax, even though the tax legally applies to only one sector.

For example, the corporate income tax is likely to be shifted across sectors. This tax applies only to the corporate sector, but the above analysis suggests that the burden is shared by owners of capital in both the corporate and noncorporate sectors. Similarly, tax provisions that apply to only a single industry are likely to ultimately affect owners of capital in all industries.

Shifting to Workers

Shifting across sectors may not be the most important way in which the burden of capital income taxes is shifted. In the long run, much of the burden of capital income taxes (whether imposed at the firm or individual level) is likely to be shifted to workers. The reason is that such taxes reduce investment, which diminishes the capital stock. With a smaller capital stock, the before-tax rate of return to capital is higher, offsetting part of the burden that the owners of capital would otherwise bear. Also, workers are less productive because they have a smaller capital stock to work with and earn lower real wages. Part of the tax burden is therefore shifted to workers.

In accordance with the insights obtained by studying the incidence of excise taxes, owners of capital bear less of the burden if the supply of capital is more responsive to changes in its after-tax rate of return. This responsiveness, and hence the extent to which capital income taxes are shifted, depends upon several factors, including the amount of time that has elapsed since the tax was imposed, the willingness of consumers to substitute between current and future consumption, and the extent to which capital can escape the tax by relocating abroad.

The time frame is very important. The shifting of the tax burden to workers is likely to occur slowly because it takes time for large changes in the capital stock to occur. In the short run, the tax causes little change in the capital stock, because most of the capital on hand was already in existence when the tax was adopted. With little change in the capital stock, very little of the burden is shifted from owners of capital to workers. Over time, however, the tax has a greater impact on the capital stock as it discourages the accumulation of new capital. As a result, more of the tax burden falls on workers and less falls on owners of capital.

Under certain assumptions, the entire burden of the capital income tax is shifted to workers in the long run, although owners of capital bear much of the burden in the short run. A textbook model of economic growth, called the Ramsey model, provides an illustration of this effect. (The Appendix to Chapter 5, *Dynamic Revenue and Budget Estimation*, explains the basic features of this model.) Using plausible values for the key inputs to the Ramsey model demonstrates that the economy adjusts only gradually to a capital tax increase. Initially, 100 percent of the burden of a capital tax increase is borne by the owners of capital, since they have already invested

in the capital currently in place. Five years after the tax increase, about a quarter of the tax burden has shifted to workers. Ten years after the tax increase, workers have taken on over 40 percent of the burden. It takes 50 years for the burden to shift nearly completely—by that time, capital owners bear only 6 percent of the burden and workers bear 94 percent.

If consumers are more willing to substitute between present consumption and the future consumption made possible by their savings, saving is more responsive to the after-tax rate of return and more of the capital income tax is shifted. The responsiveness of saving to the after-tax rate of return also depends on consumers' planning horizons. The Ramsey model assumes that consumers consider the impact of their saving decisions on their descendants. If, instead, consumers plan only for their own lifetimes, saving is less responsive to changes in its after-tax rate of return and less of the capital income tax burden is shifted to workers.

International capital flows also play a role. If the tax applies only to capital located in the United States and capital is mobile across international boundaries, the tax is more likely to be shifted to workers. The above example assumes that there are no international capital flows; incorporating such flows would increase the speed at which the tax is shifted.

Empirical work provides some evidence that capital income taxes are shifted to some extent: studies find that the before-tax return to capital income is higher when the tax rate on capital income is higher. However, the picture is not entirely clear, because other factors may cause tax rates and before-tax rates of return to move together.

The belief that a large portion of the capital income tax burden is shifted in the long run is common in the economics profession. In a 1996 survey, public finance economists were asked to state "the percentage of the current corporate income tax in the United States that is ultimately borne by capital." The average response was 41 percent, and three-quarters of the respondents gave answers of 65 percent or less. This survey indicates that the average public finance economist believes that more than half of the tax is eventually shifted from the owners of capital to workers or other groups.

Because labor income is more evenly distributed across taxpayers than capital income is, recognizing that part of the burden of capital income taxes is shifted to workers reveals that high income taxpayers bear a smaller share of the burden than is often assumed. Chart 4-1 classifies households by their levels of total income and tabulates the share of national labor income and national capital income earned by different groups. The chart shows, for example, that the 10 percent of households with the highest total incomes receive 37 percent of labor income and 62 percent of capital income. If half of capital taxes are shifted to workers in the long run, the fraction of the burden falling on this high-income group is reduced from 62 percent to 49 percent; if all capital taxes are shifted to workers in the long run, the high-income share of the burden falls to 37 percent.

Percent of total tax burden
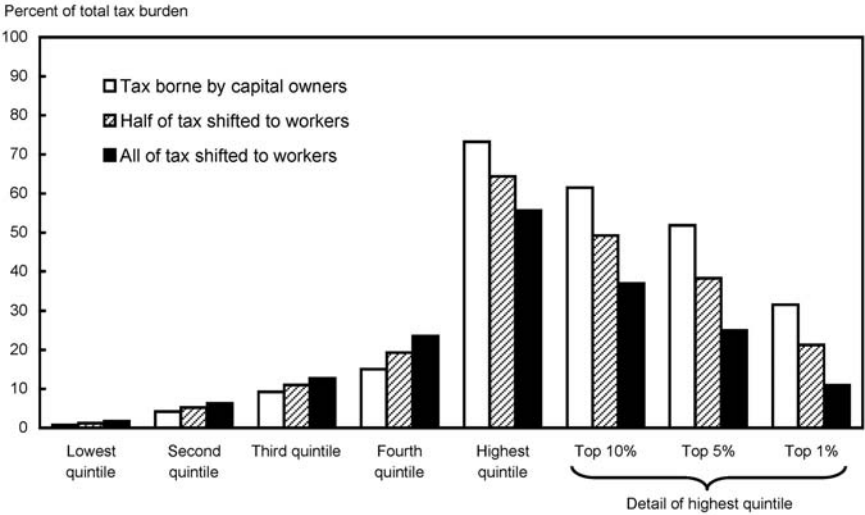
Note: Quintiles defined with respect to family economic income. Treasury model for 1989 extended to 2000.
Source: Julie-Anne Cronin, "U.S. Treasury Distributional Analysis Methodology," Department of the Treasury, Office of Tax Analysis, OTA Paper 85, September 1999.

Applied Distributional Analysis and the Choice of Time Frame

Official distributional analyses differ in their treatment of the corporate income tax. The JCT previously distributed the burden to owners of corporate capital, but now does not distribute it on the grounds that the incidence of the corporate income tax is uncertain. The CBO and Treasury now distribute the corporate tax burden to owners of all capital. None of these analyses currently recognizes the shifting of the tax to workers. The CBO previously presented analyses that allocated half of the burden to workers and Treasury did the same in its January 1992 corporate integration study. Official analyses generally allocate individual income taxes on capital income to the persons who bear the legal incidence of the taxes.

The time frame plays a key role in how tax incidence is treated. When the JCT adopted its former practice of allocating the corporate income tax to corporate capital, it stated that its analysis was intended to refer to the very short run, when little shifting of any kind would occur. Similarly, Treasury has justified allocating the burden to owners of all capital by stating that this is the most reasonable assumption for incidence over a 10-year horizon. These analyses serve the useful objective of informing policy makers of how the current tax burden is divided between current workers and current owners of capital.

Nevertheless, presenting estimates only for short time frames leaves an incomplete picture. If a tax change is intended to be permanent, it is important to also inform policy makers how its long-run burden will be divided between future workers and future owners of capital. Answering that question requires additional distributional tables that recognize the significant shifting to workers that is likely to occur in the long run.

Estate and Gift Taxes

Capital can also be subject to estate and gift taxes when its ownership changes hands due to an inheritance or gift. The lessons from the analysis of capital income taxes can therefore be applied to estate and gift taxes.

The estate and gift taxes apply on a cumulative basis to an individual's lifetime gifts and to the estate the individual bequeaths at his or her death. An individual may make up to $11,000 of gifts to any recipient per year, without counting them against the lifetime total. Bequests to surviving spouses are exempt, as are gifts and bequests to charitable organizations.

The taxes apply only when lifetime gifts plus the estate exceed an exemption amount, which was $675,000 in 2001. Under the laws in place at the beginning of 2001, the exemption amount was scheduled to increase to $1,000,000 starting in 2006. The taxes applied at rates of up to 55 percent.

The tax law adopted in June 2001 provides for further reductions in the estate and gift taxes for 2002 through 2009. This law increases the exemption amount to $1 million for 2002 and 2003 and gradually increases it to $3.5 million for 2009. The law reduces the top tax rate from 2002 to 2009, with top rates of 50 percent in 2002 and 45 percent in 2007 through 2009. For 2010, the law completely repeals the estate tax, but retains the gift tax with a top rate of 35 percent. It also increases, in some cases, the capital gains taxes paid by heirs who sell property that they inherit.

Because the 2001 tax law is scheduled to expire at the end of 2010, the estate and gift taxes are scheduled to return in 2011, at the levels specified by the previous laws. The President has proposed permanently extending the provisions of the 2001 tax law that are in effect in 2010, including the repeal of the estate tax.

The issue of who benefits from estate tax repeal has been a prominent one in the debate over repeal. Treasury allocates the burden of estate and gift taxes to the decedents (the individuals who have died) and donors. The JCT used to do

the same, but has now stopped distributing them due to uncertainty about the taxes' incidence. The CBO's recent distributional analyses have not included estate and gift taxes. Allocating the estate tax burden to decedents supports the common view that the tax is highly progressive, since (at the current exemption amount) the tax applies to only the largest 2 percent of estates.

It is virtually certain, however, that little of the economic burden of the estate tax is borne by the decedents. The burden of the estate tax is borne by them only if the tax prompts them to reduce their lifetime consumption and accumulate a larger estate, so that the tax can be paid without reducing the after-tax bequests left to their heirs. In other words, the estate tax must reduce lifetime consumption and promote estate accumulation for it to be borne by the decedents.

This condition is unlikely to hold. Because the estate tax makes estate building less attractive, it probably reduces the size of bequests. Empirical research confirms that the estate tax reduces the amount that decedents accumulate and pass on to their heirs. As a first step, it would make more sense to distribute the burden of the tax to heirs rather than to decedents.

Despite what one might expect, the heirs of wealthy decedents are not always wealthy. Economists have found that the correlation between the long-term labor earnings of successive generations is around 0.4 or 0.5. The correlation between long-term incomes (which includes the inheritances themselves) or between long-term consumption levels of successive generations has been estimated to be around 0.7. (Correlation is a number, ranging from -1 to 1, that measures the strength of the relationship between two variables. A correlation of 0.4 or 0.7 indicates that one variable tends to increase when the other increases, but that the relationship is not perfect.) Some bequests are left to grandchildren or nephews and nieces where the correlation between the incomes of decedents and the incomes of heirs may be even lower. Because heirs can be less wealthy than decedents, recognizing that the estate tax burden is more likely to fall on the former reveals that less of the burden is borne by the very wealthy.

A more important point, however, is that the reduction in estate building induced by the tax is likely to take the form of a reduction in capital accumulation. Because the estate and gift taxes are taxes on capital, part of their long-run burden is likely to be shifted to workers through a reduction in wage rates, as discussed above. Part of the burden is therefore likely borne by ordinary workers who never receive a bequest or taxable gift.

Conclusion

Distributional analysis can be a useful tool for policy makers. It is important, however, to recognize the limitations of existing analyses. Current analyses can be misleading, particularly with respect to the estate and gift taxes and other capital taxes. These taxes are likely to be shifted substantially to workers in the long run, reducing the extent to which their burden falls on high-income groups.

Dynamic Revenue and Budget Estimation

A central conclusion of the study of taxation is that taxes affect behavior and distort the choices of firms, workers, and investors. In particular, a higher tax on an activity tends to discourage that activity relative to others. These behavioral responses to a tax change can, among other things, alter the revenue effect of the tax change, a topic that is the focus of this chapter. Revenue estimation is called *dynamic* if it incorporates the revenue implications of behavioral responses to tax changes and *static* if it does not incorporate these revenue implications. Like changes in taxes, changes in government spending can encourage or discourage certain behavior; budget estimates are dynamic if they incorporate the budgetary implications of these behavioral responses.

If policy makers are to make informed decisions about policy changes, all significant effects should ideally be included in estimates of the policy's budgetary implications. Several obstacles have prevented macroeconomic behavioral responses from being incorporated in such estimates. This chapter discusses the ongoing efforts to provide a greater role for fully dynamic revenue and budget estimation in the analysis of major tax and spending proposals.

The key points in this chapter are:

- Currently, official revenue estimates of proposed tax changes are not fully static because they incorporate the revenue effects of many microeconomic behavioral responses. These estimates are not fully dynamic, however, because they exclude the effects of macroeconomic behavioral responses.
- Changes in taxes and spending generally alter incentives for work, investment, and other productive activity. These macroeconomic behavioral responses have revenue and budgetary implications.
- Steps have recently been taken to provide more information about the revenue effects of macroeconomic behavioral responses. At least in the near term, it may not be practical for macroeconomic effects to be incorporated in official estimates. But estimates of these effects should be provided as supplementary information for major tax and spending proposals.
- Dynamic estimation of policy changes should distinguish aggregate demand effects from aggregate supply effects, include long-run effects, apply to spending as well as tax changes, reflect the differing effects of various policy changes, account for the need to finance policy changes, and use a variety of models.

Revenue Estimation and
Microeconomic Behavioral Responses

To frame the issues, it is useful to begin with a simple example of how a tax change can affect behavior and how the behavioral response then alters the revenue impact of the tax change.

An Example of Revenue Implications of Microeconomic Behavioral Responses

Consider an excise tax on apples (similar to that discussed in Chapter 4, *Tax Incidence: Who Bears the Tax Burden?*). If the current tax rate is 25 cents per apple and 1,000 apples are produced and consumed at this tax rate, tax revenue is $250. Now, suppose the tax rate is cut to 20 cents per apple. If apple output and consumption don't change, total tax revenue falls to $200, a decrease of $50. Therefore, a purely static estimate of the revenue loss would be $50.

The actual change in tax revenue is likely to be different, however, because consumers and producers respond to the tax rate change. The tax rate drives a wedge between the price paid by consumers (including the tax) and the price that producers receive (net of the tax). When the tax rate is reduced, this wedge is reduced, meaning that consumers are likely to pay a lower price and producers are likely to receive a higher price. For example, at the 25-cent tax rate, consumers might pay $1.13 per apple and producers might receive 88 cents per apple; at the 20-cent tax rate, consumers might pay $1.10 and producers might receive 90 cents. The lower price paid by consumers induces them to consume more apples and the higher price received by producers induces them to produce more apples. As explained in Chapter 4, the changes in the two prices must be such that consumers' desired increase in consumption equals producers' desired increase in production.

Suppose that 1,100 apples are produced and consumed at the lower tax rate. (The actual increase in the quantity of apples depends on how responsive consumers and producers are to their respective prices; the increase is larger when both groups are more responsive.) Tax revenue is then $220 (20 cents per apple times 1,100 apples), not $200. Thus, the tax cut lowers revenue by $30, not $50. Of the $50 static revenue loss, $20 is "paid for" by the increase in apple production and consumption caused by the tax cut. In other words, 40 percent of the tax cut "pays for itself" through this revenue feedback.

Conversely, increasing the tax from 25 cents to 30 cents yields $50 of additional revenue if the quantity of apples remains at 1,000. However, the quantity of apples is likely to fall as the tax rate increases. With the higher tax, consumers pay a higher price and producers receive a lower price,

prompting a decline in the desired levels of apple production and consumption. If the quantity of apples falls from 1,000 to 900, revenue rises from $250 to $270, so that the revenue gain from the tax rate increase is $20 rather than the $50 that would occur with no behavioral response.

The actual revenue effects of such a tax change may be more complex than this discussion suggests. As the quantity of apples changes, the quantities of other items produced and consumed also change. If those items are also subject to taxes, changes in those quantities also impact revenue. In any event, behavioral responses to a tax change can alter its revenue impact.

Incorporation of Microeconomic Behavioral Responses in Revenue Estimation

The insight that microeconomic behavioral responses to tax changes affect revenue has been incorporated into the official revenue-estimation process.

The staff of the Joint Committee on Taxation (JCT) prepares the official revenue estimates for thousands of proposed tax changes submitted by members of Congress each year. Similarly, the Department of the Treasury prepares official revenue estimates for tax changes proposed by the President and some changes considered by the Congress. Official estimation of the revenue effect of a tax change is commonly called *scoring*. Each revenue estimate presents the estimated change in revenues in the current fiscal year and up to 10 subsequent fiscal years, a period referred to as the "10-year window."

In preparing their estimates, JCT and Treasury economists routinely include the effects of microeconomic behavioral responses to tax changes. For example, when excise taxes change, JCT and Treasury estimates reflect how much sales of the taxed item are expected to change. So, official revenue estimates for the hypothetical apple tax change described above would reflect an estimate of the change in the quantity of apples. For changes in the tax treatment of a particular type of business investment, the revenue estimates reflect shifts between that type of investment and other types.

Changes in the capital gains tax rate provide another example of how behavioral changes play a prominent role in the scoring process. Economic theory and statistical studies have established that capital gains taxes deter *realization* of capital gains—the sale of assets that have risen in value. A cut in the capital gains tax rate, therefore, is likely to spur an increase in capital gains realizations. Put simply, investors are likely to sell their assets to take advantage of the lower tax rate on any gains they have already accrued. This increase in realizations will mean that the capital gains tax will be applied to a larger tax base, partially offsetting the cut in the tax rate itself. Indeed, depending on the timing and structure of the rate cut, it may actually raise revenue immediately after enactment. JCT and Treasury estimates recognize these effects.

Economists' understanding of—and data on—human behavior is incomplete. This makes it difficult to determine the exact magnitude of behavioral responses to tax changes and their exact impact on tax revenues. Nevertheless, a revenue estimate that ignores such behavioral responses will be inaccurate. By taking account of microeconomic behavioral responses, the JCT and Treasury produce estimates that are likely to be more accurate than strictly static estimates.

Macroeconomic Behavioral Responses to Policy Changes

Despite advances in making revenue estimates more dynamic, the incorporation of behavioral responses has been subject to one fundamental limitation. The official revenue estimates assume that macroeconomic aggregates, such as total investment, total labor supply, and GDP, are not affected by tax and spending changes. Because the estimates ignore these potentially important effects, they are not fully dynamic.

Lowering taxes on labor and capital income strengthens incentives to work and invest and is likely to spur increases in these activities. Additional work and investment boosts national income, which increases the tax base and thus partially offsets the revenue loss from lower tax rates.

As an example, suppose that the current income tax rate is 25 percent and that total national income is $1,000. Total tax revenue is $250. Now, suppose the tax rate is cut to 20 percent. If total income did not change, total tax revenue would be $200. The lower tax rate, however, is likely to encourage work and saving, boosting total income. If income rises to $1,100, total tax revenue will be $220, not $200. Thus, the tax cut lowers total tax revenue by $30, not $50. In other words, 40 percent of the tax cut ($20 of the $50 static revenue loss) "pays for itself."

Popular attention is often focused on the possibility that an income tax rate cut could stimulate so much additional income that it would *fully* pay for itself. Most economists believe that, starting from current U.S. tax rates, such an outcome is unlikely for a broad-based income tax change. It is important to realize, however, that any behavioral response alters the size of the revenue loss from a tax cut, even if it does not transform the loss into a revenue gain.

Official scoring of income tax rate changes already includes a number of microeconomic behavioral responses. The scoring takes into account a

variety of ways in which the rate cut may raise taxable income, such as a shift from tax-exempt fringe benefits to taxable wages. However, the estimates do not recognize that lower income taxes can encourage greater labor supply and capital accumulation, and thereby raise total income in the economy.

The exclusion of macroeconomic behavioral responses from official revenue estimation is not due to ignorance of these responses or disagreement about their existence. Instead, it reflects the judgment that accurately including these effects is impractical, due to the controversy about their magnitudes and the complexity of modeling them. Uncertainty about the correct model of the economy and the size of behavioral responses to tax changes, and disagreement about the appropriate time frame for revenue projections has made consensus difficult to achieve.

Ultimately, it may be possible for macroeconomic effects to become part of the official scoring process for those tax and spending proposals that are likely to have significant macroeconomic effects. However, given the time and resource constraints facing revenue estimators and the lack of consensus about these issues, that goal is not likely to be feasible in the near future. To promote informed policy making, though, it is essential that fully dynamic revenue estimates (incorporating macroeconomic as well as microeconomic effects) be presented as supplementary information for major tax and spending proposals.

Recently, estimators have taken major steps in precisely this direction. In November 1997, the JCT compiled and published estimates of the macroeconomic effects of fundamental tax reform prepared by nine sets of economists using nine different economic models. The JCT subsequently began developing its own macroeconomic models and formed a blue-ribbon panel of academic and private-sector economists to further explore dynamic revenue estimation. In January 2003, the House of Representatives adopted Rule XIII.3(h)(2), which requires the JCT to prepare analyses of the macroeconomic effects of major tax bills before such bills can be considered by the House. In May 2003, the JCT prepared such an analysis of the Ways and Means Committee's version of the Jobs and Growth tax bill. In December 2003, the JCT published a description of the methodology it used for this analysis. The Congressional Budget Office (CBO) provided a similar analysis of the President's 2004 Budget in March 2003 and provided a more detailed description of its analysis in a July 2003 technical document. Private organizations have also prepared dynamic analyses of proposed tax changes that reflect macroeconomic behavioral responses

User's Guide to Dynamic Revenue and Budget Estimation

Recent work suggests six guidelines for dynamic revenue and budget estimation.

Guideline 1: Dynamic Estimation Should Distinguish Aggregate Demand Effects and Aggregate Supply Effects

Tax cuts can affect output through two different channels, by changing aggregate demand and by changing aggregate supply. Any aggregate demand effects are likely to be concentrated in the first few years. Aggregate supply effects are likely to occur over a longer time period.

Changes in Aggregate Demand

In the short run, tax cuts may push an underperforming economy back toward its potential by raising consumers' disposable incomes and, thus, their demand for goods and services. Tax cuts may also increase firms' demand for investment goods. These effects increase the aggregate demand for goods and services.

The extent of the increase in aggregate demand depends upon how the tax cut is financed. If the tax cut is accompanied by reductions in government spending, little or no stimulus to aggregate demand is likely to occur. If the tax cut is financed by borrowing, then aggregate demand is more likely to be stimulated.

The net effect of a tax cut on aggregate demand also depends upon the reaction of the Federal Reserve. If taxes are cut in an under-performing economy, the Federal Reserve may perceive less need for interest-rate reductions. In such a case, the boost to aggregate demand from a tax cut would, at least in part, be offset by the reduced stimulus provided by the Federal Reserve.

The Federal Reserve is less likely to offset the aggregate demand stimulus from tax cuts, however, in a low-interest-rate economy, because interest rates cannot go below zero. Under these circumstances, the fiscal stimulus provided by a tax cut may reinforce, rather than replace, monetary stimulus. This case seems relevant for the 2003 tax cut; the Federal Reserve's target for the Federal funds interest rate was 1.25 percent from November 6, 2002, to June 25, 2003, and has since remained at 1 percent.

Aggregate demand effects are primarily relevant in the short run. These effects tend to fade over time as prices and wage rates adjust and the economy returns to its normal level of output. The bulk of the aggregate demand stimulus from a policy change is likely to be felt within a few years.

Changes in Aggregate Supply

Tax cuts can raise after-tax returns to work and investment, encouraging both activities. This effect increases aggregate supply because it increases the amount of goods and services that the economy is capable of producing.

Tax changes can also improve the long-run allocation of resources, allowing greater output to be produced with a given set of resources. One way to do this is to make tax rates more uniform across different types of income, as exemplified by the reduction in dividend and capital gains tax rates adopted in the Jobs and Growth Tax Relief Reconciliation Act of 2003. This provision reduced the tax burden on investment in the corporate sector, which was more heavily taxed than investment in the noncorporate sector. Over time, this tax reduction is expected to make the allocation of resources more efficient, leading the economy to allocate more resources to the corporate sector.

Because some commercially available forecasting models tend to emphasize short-run aggregate demand effects, it may be necessary to develop models that place greater emphasis on long-run aggregate supply effects. In their recent dynamic analyses, the CBO and the JCT used a mix of models with varying emphases on short-run and long-run effects. The time frame over which tax revenues are estimated should be long enough to fully capture the longer-run, supply-side effects, which leads us to the second guideline.

Guideline 2: Dynamic Estimation Should Include Long-Run Effects

While official revenue scoring is confined to a 10-year "window," it is important that dynamic revenue estimation provide some information for a longer horizon. Presenting the dynamic revenue estimates as supplementary information, rather than as part of the official revenue estimate, facilitates the use of a longer horizon.

The longer horizon is necessary because exclusive use of the 10-year horizon skews the emphasis given to different macroeconomic effects. As discussed in guideline 1, aggregate demand effects are likely to be fully realized within the 10-year window but have little long-run importance. In contrast, aggregate supply effects may not fully materialize within the 10-year window: Although changes in labor supply may occur relatively quickly, changes in the capital stock occur more slowly.

Economic analysis indicates that, in a closed economy, such capital accumulation takes place over a period of decades. Consider an example in which the government cuts taxes slightly on capital income, starting from a 25 percent marginal tax rate. If standard parameter values are assumed for a leading model of economic growth (the Ramsey growth model used in Chapter 4 and described in the Appendix to this chapter), only 42 percent

of the long-run increase in the capital stock is put in place within the first 10 years. That is, more than half of the increased capital stock accumulates outside the conventional 10-year window. In fact, only two-thirds of the increase takes place within 20 years. In an open economy, international capital flows may allow the capital stock to adjust somewhat more quickly. Still, this analysis indicates the importance of considering a long time horizon when estimating a tax cut's effect on aggregate supply.

A longer time horizon would give adequate emphasis to the tax cut's aggregate supply effects. It would also permit policy makers to accurately compare the fundamental consequences of different types of tax and spending changes. This leads to the third and fourth guidelines.

Guideline 3: Dynamic Estimation Should Be Applied to Spending Changes as well as Tax Changes

The logic behind dynamic estimation applies to spending as well as tax changes. As discussed more fully in guideline 4, spending programs can also affect aggregate demand and aggregate supply. To be sure, dynamic budget estimation for spending changes can be even more difficult, and has been less common, than dynamic revenue estimation for tax changes. Nevertheless, including macroeconomic effects only for tax changes would lead to an unbalanced and misleading comparison of policies.

Guideline 4: Dynamic Estimation Should Reflect the Differing Macroeconomic Effects of Various Tax and Spending Changes

Not all types of taxes or spending programs would be expected to have the same effects on the economy. Moreover, the policies that may have the most beneficial effects in the short run (because they provide a powerful boost to aggregate demand) can, in some cases, be the least beneficial, or even harmful, in the long run (because they fail to boost aggregate supply by increasing work and investment).

In the short run, the most immediate stimulus may be provided by an increase in government purchases of goods and services, an increase in transfer payments, or by tax cuts designed to boost consumer spending. These policies, however, are generally not the best ways to boost aggregate supply. Although some spending programs, such as infrastructure construction and education, may increase economic growth, others, particularly some transfer payments, would be expected to reduce aggregate supply by weakening incentives to work and invest.

In the long run, the strongest boost to aggregate supply is likely to come from tax cuts designed to boost investment. Tax cuts on capital income are

most likely to have this effect. Their short-run revenue feedback may be small (because their aggregate demand effects may be limited), but their long-run revenue feedback may be large.

Consider again the example discussed above, in which the government cuts capital taxes slightly, starting from a 25 percent marginal tax rate. Using the same assumptions as before (as detailed in the Appendix to this chapter), increased growth resulting from the tax cut significantly moderates its revenue loss. This is particularly true in the very long run (after 50 years or so), as the economy settles back toward equilibrium. At that point, the reduction in tax revenues is about half of what conventional scoring would indicate. The estimated revenue feedback is so large for two reasons: the policy being considered is a reduction in the capital tax rate and the estimate refers to the very long run.

It is also possible for some tax cuts to reduce the incentive to work and save. In this case, the revenue loss from the tax cut is *larger* than it would have been without macroeconomic behavioral changes. A prime example is an increase in a tax credit or deduction that is phased out as income rises. Such a phase-out is another form of a higher marginal tax rate on income. For example, married couples filing jointly lose $5 of tax credits per child for each $100 of additional income above $110,000. For a couple with two children, this phase-out increases their effective marginal income tax rate by 10 percentage points. Given the existence of the phase-out, any increase in the size of the credit lengthens the interval of income to which this higher marginal tax rate applies. As with any increase in marginal income tax rates, parents in this income bracket have less incentive to work, save, or otherwise increase their income. This outcome does not mean that increasing the child credit is a bad idea. The President proposed such increases in 2001 and 2003 and advocates making the increases permanent. As long as the income phase-out remains in place, however, revenue estimates for increases in the credit should include the revenue lost because of the reduction in the parents' work and saving.

Dynamic estimation is not accurate unless it includes all of the policy changes that are required to support the tax change. This leads to our fifth guideline.

Guideline 5: Dynamic Estimation Should Account For the Need to Finance Policy Changes

Because the government is a going concern, it need never actually pay off its outstanding debt. Nevertheless, government debt cannot indefinitely grow faster than national income. One implication of this constraint is that, over the entire time frame of the economy's existence, the present value of tax revenue must equal the present value of noninterest government

spending. (Present value takes into account the time value of money—the fact that monetary sums can earn interest over time.) In the long run, there can be no "unfunded tax cuts" or "unfunded spending increases." If current tax revenue is reduced while current spending is left unchanged or if current spending is increased while current revenue is left unchanged, government debt increases. Servicing this debt requires that future taxes be higher, future spending be lower, or both.

To be sure, it is infeasible for estimators to accurately predict which adjustments future Congresses and Presidents may adopt. Nevertheless, to avoid analyzing economically impossible policy specifications, dynamic revenue estimates should recognize that some such adjustments must occur. To ensure comparability across proposals, it may be best to adopt a few stylized assumptions about the nature of the financing. A few benchmark cases could then be considered; perhaps one in which the debt service is financed with reductions in government purchases, one in which it is financed with reductions in transfer payments, and one in which it is financed with higher income tax rates.

How a tax cut is financed can alter its effects in the short run and the long run. If current revenues are reduced through a tax cut while current spending is held fixed, the government's budget deficit will get larger. Such a deficit-financed tax cut has a strong positive effect on aggregate demand because consumers are likely to spend part of the tax cut and there is no offsetting reduction in government spending. It may do somewhat less to boost aggregate supply, however, if the deficit raises interest rates and, as a result, lowers investment. This effect is often called *crowding-out*, because a government's deficit spending reduces, or crowds-out, the amount of savings available for private firms to use for funding investment. On the other hand, if current spending is reduced along with current revenue, the aggregate demand effects of the tax cut are muted, because the spending cuts lower aggregate demand. The boost to aggregate supply is greater, however, because no crowding-out occurs.

To maximize the aggregate supply impact of the recent tax cuts, the President has stressed the need to restrain government spending.

Guideline 6: Dynamic Revenue Estimation Should Use a Variety of Models Until Greater Consensus Develops

One challenge facing estimators is that different models yield different results. Comparing the results from different models is the best way to resolve differences between possible approaches and to test the sensitivity of results to changes in assumptions. To improve our ability to distinguish

among models, a set of models could be applied to clearly defined and relatively simple hypothetical policies. This would allow the different models' results to be compared and would make it easier to attribute any variation among their results to differences in their assumptions. As mentioned above, the JCT did such an exercise in 1997 when exploring the possible effects of fundamental tax reform. The CBO and the JCT also used a variety of models in their dynamic analyses in 2003. Presenting dynamic revenue estimates as supplementary information rather than as part of the official revenue estimates facilitates the use of a variety of models.

One reason that dynamic revenue estimation is subject to so much uncertainty is that fiscal changes may have important effects that are left out of standard models of economic growth. For example, standard growth models take the rate of technological progress as given. Some research, however, has suggested that technological progress may be a by-product of capital accumulation; if so, changes in capital income taxes can alter the rate of technological progress. As another example, standard models take the economy's equilibrium level of unemployment as given. Yet some research has indicated that the equilibrium unemployment rate depends on productivity growth, which can also be influenced by changes in capital taxation. Incorporating such nonstandard effects into dynamic revenue estimation is undoubtedly a formidable challenge, but if initial results on these effects are confirmed by future research, this challenge should not be avoided.

Conclusion

Fully dynamic revenue estimation that incorporates macroeconomic behavioral changes is an important step forward in applying economic insights to policy analysis. Significant progress has been made on this front; continued progress is essential to sound policy making.

Appendix: The Model Used in the Capital-Tax Example

The model underlying the capital-tax example is the growth model developed by Frank Ramsey in 1928. It is a leading textbook model, and most of its assumptions are standard among models of economic growth. For instance, output is produced by combining capital and labor, and productivity growth increases how much output a given amount of capital and labor can produce. Consumers maximize their welfare by deciding how much of their income to save. Businesses maximize profit and compete when hiring workers and selling products. Over the long term, the saving rate determines the capital stock and, thus, the level of output in the economy. The Ramsey

model allows consumers to choose their saving rate, while simpler models impose a constant saving rate estimated from historical data.

Unlike some other models, the Ramsey model assumes that consumers are members of families comprised of an infinite number of generations and that they care about the well-being of their descendants. This means that consumers consider the effects of their choices on their children and subsequent generations. Some critics of the Ramsey model view this assumption as unreasonable. However, the results presented in the text do not change substantially if we assume that people care less about each successive generation and, for generations far enough into the future, hardly consider their welfare at all.

In the Ramsey model, the long-run equilibrium for the economy can be described by two relationships. Firms invest in capital equipment until the value of the output produced by the last unit of capital equipment just equals the interest rate—the cost borne by the firm to invest. The interest rate is, in turn, determined by consumers' choices about their consumption and savings. These choices depend on the growth rate of technology, the discount rate (a measure of how much consumers prefer having a dollar today compared to a dollar in one year), and consumers' flexibility with regard to spending in different time periods. To solve the model, we must make an assumption about how the government finances policy changes in the long run. In the capital-tax example, we assume that the government adjusts transfer payments accordingly.

Knowing this long-run equilibrium allows us to calculate the impact of a cut in tax rates on tax revenue taking into account the aggregate dynamic effects that this chapter has described. In particular, we can summarize the difference between a dynamic analysis and a static analysis with a few key parameters, or inputs, to the model. We assume that the tax rates on labor and capital income are each 25 percent, capital's share of total income is one-third, and the elasticity of substitution between capital and labor is one. Then, if dynamic effects are considered, a capital tax cut reduces tax revenue in long-run equilibrium by half as much as a static analysis would indicate.

Restoring Solvency to Social Security

Much of the Federal government's budget is dedicated to *entitlement programs*, in which expenditures are determined not by discretionary budget allocations but by the number of people who qualify. Reform of entitlement programs remains the most pressing fiscal policy issue confronting the Nation. With projected expenditures of $478 billion in 2003, Social Security is the largest entitlement program and an appropriate place to begin. Social Security is designed as a *pay-as-you-go* system in which payroll taxes on the wages of current workers finance the benefits being paid to current retirees. While the program is running a small surplus at present, large deficits loom in the future. Deficits are first projected to appear in 15 years; by 2080, the Social Security deficit is projected to exceed 2.3 percent of GDP.

The coming deficits in Social Security are driven by two demographic shifts that have been in progress for several decades: people are having fewer children and are living longer. The President has called for new initiatives to modernize Social Security to contain costs, expand choice, and make the program secure and financially viable for future generations of Americans.

This chapter assesses the need to strengthen Social Security in light of its long-term financial outlook. The key points in this chapter are:

- The most straightforward way to characterize the financial imbalance in entitlement programs such as Social Security is by considering their long-term annual deficits. Even after the baby-boom generation's effect is no longer felt, Social Security is projected to incur annual deficits greater than 50 percent of payroll tax revenues.
- These deficits are so large that they require a meaningful change to Social Security in future years. Reform should include moderation of the growth of benefits that are unfunded and can therefore be paid only by assessing taxes in the future. A new system of personal retirement accounts should be established to help pay future benefits. The benefits promised to those in or near retirement should be maintained in full.
- The economic rationale for undertaking this reform in an era of budget deficits is as compelling as it was in an era of budget surpluses.

The Rationale for Social Security

All developed countries and most developing countries have publicly administered programs to provide benefits for the elderly, including programs to support surviving spouses and the disabled. Government involvement in markets for goods or services is typically predicated on a failure of private markets to achieve an efficient or equitable result. There are three main problems in the market for providing support to the elderly that justify a government role in old-age entitlement programs.

First, a strictly private market to support old age would require all individuals to choose the level of consumption that they would like in retirement and to save accordingly. Some individuals may not be capable of making the relevant calculations themselves and may not be able to enlist the service of a financial professional to advise them. For these people, Social Security provides a minimal level of financial planning. Social Security requires people who otherwise would not save for retirement to participate in a system that makes them pay for insurance against old-age poverty. It also provides a mechanism for everyone to share in the burden of taking care of those who are truly in need of assistance.

Second, well-being in retirement is subject to two types of risk that are not easily insured in private markets. The first risk is low income during working years, which can lead to poverty in old age. Low income may be caused by a specific event like disability, and Social Security provides workers with disability insurance. Private disability insurance plans exist but participation is quite low. Low income may also be caused by other events beyond an individual's control. However, these events do not lend themselves to private insurance contracts because income can also be low for a variety of reasons that are under an individual's control but that are difficult for an insurer to observe (such as low work effort). Social Security partially overcomes this problem through its progressive benefit formula—retirees with lower earnings during their working years get benefits that are higher as a share of preretirement earnings.

The other risk to well-being in old age is the possibility that retirees will live an unusually long time and thereby exhaust their personal savings. To protect against this risk, a portion of the retirement wealth that a worker has accumulated must be converted to an annuity, a contract that makes scheduled payments to the individual and his or her dependents for the remainder of their lifetimes. The annuity payments should be indexed to inflation, so that their purchasing power is not eroded over time. Inflation-indexed annuities are a fairly new financial product, and even today, relatively few people participate in the indexed annuity market. A public system of Social Security, in which the government pays benefits in the form of an annuity

that keeps pace with the cost of living, can help protect retirees from outliving their means to support themselves.

Third, in other contexts, the government's fiscal policies are designed to redistribute resources from high- to low-income individuals. In most cases, such as the progressive income tax schedule, income is defined based on an annual measure. Social Security is unusual because it can redistribute income based on a lifetime average of earnings. By doing so, Social Security more accurately targets these transfers to the people who most need the assistance. Individuals with higher lifetime average earnings receive benefits from Social Security that are higher in dollar terms, but lower as a percentage of their earnings, than do those with lower lifetime average earnings.

All of these rationales are legitimate. Whether the U.S. system actually meets these goals, and whether it does so in an efficient and equitable manner, however, should be a subject of continued debate. An essential part of this debate is that none of these rationales require that Social Security be operated on a pay-as-you-go basis. Long-term solvency can be restored by advance funding of future obligations through personal retirement accounts. Personal retirement account proposals can be and often have been designed to allow for greater protection for surviving spouses and other vulnerable groups. The President has taken an important step in this debate by making the modernization and long-term solvency of Social Security a prominent feature of his Administration's domestic policy agenda.

Understanding the Financial Crisis

In a pay-as-you-go system like Social Security, the benefits paid to current beneficiaries are financed largely by the payroll taxes collected from current workers. In any given year, the system will be in balance when the *income rate* equals the *cost rate*. The income rate is the total amount of tax revenue collected (from both the payroll tax and the income taxation of Social Security benefits for moderate- and high-income beneficiaries) divided by the total amount of payroll on which taxes are levied. The cost rate is the total amount of scheduled benefits divided by the total payroll on which taxes are levied. The *annual balance* is the difference between the income rate and the cost rate in a given year.

The impending financial crisis in Social Security is due to the rapid growth in the cost rate relative to the income rate in the future. This growth is attributable to two demographic factors that have become critically important over the last half century: people are having fewer children and living longer in old age. As a result of these lower rates of both fertility and mortality, the size of the elderly cohort will expand relative to the younger cohort over time.

Chart 6-1 Demographic Change and the Cost of Social Security Through 2080
The increase in the cost rate for Social Security closely tracks the change in the dependency ratio.

Percent of taxable payroll   Beneficiaries per 100 workers

Dependency ratio
(right scale)

Social Security cost rate
(left scale)

2003   2013   2023   2033   2043   2053   2063   2073

Source: Social Security Administration.
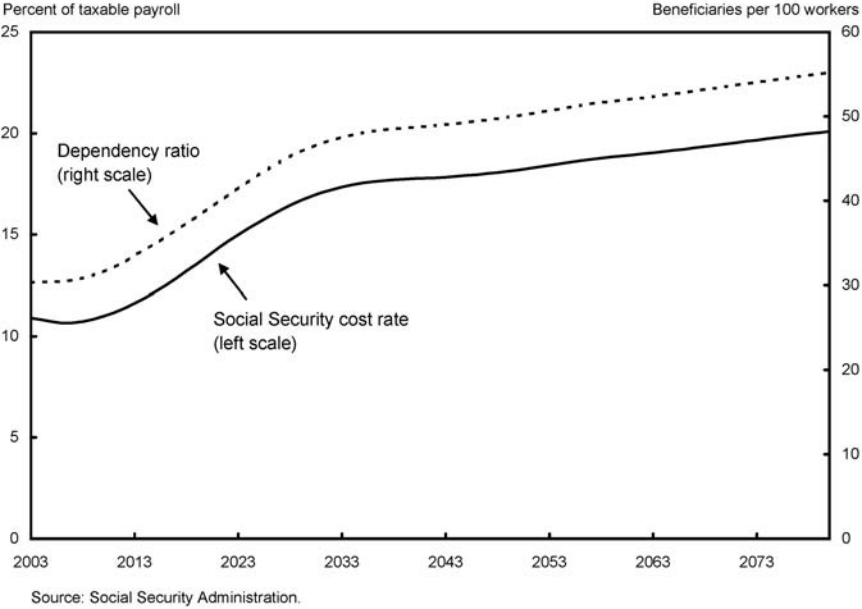
Chart 6-1 compares the Social Security cost rate with the *dependency ratio*, which is the number of beneficiaries per hundred workers. The projections are based on the intermediate assumptions made by the Social Security Trustees in their 2003 report. The dependency ratio rises from 30.4 in 2003 to 55.2 in 2080, an increase of 82 percent. Stated another way, the number of workers paying payroll taxes to support the payments to each beneficiary will fall from 3.3 workers per beneficiary in 2003 to 1.8 in 2080. With fewer workers to support each retiree, it is not surprising that the cost rate is projected to increase, in this case from 10.89 percent of payroll in 2003 to 20.09 percent in 2080. This 84 percent increase is almost identical to the rise in the dependency ratio. While changes in productivity, immigration, interest rates, and other factors also affect the long-term solvency of the program, changes in population structure are at the center of the looming crisis.

Chart 6-2 graphs the single-year projections of Social Security's income and cost rates, along with its annual balances. The solid curve that rises over the period represents the cost rate (this is the same curve as in Chart 6-1). The dashed line is the projected income rate, which reflects revenue received by the Social Security trust funds from the payroll tax of 12.40 percent plus a portion of the income tax on current benefits. Income taxation on benefits currently being paid generates an amount equal to 0.30 percent of taxable payroll. Thus,

Chart 6-2 Social Security's Annual Balances Through 2080

Social Security begins to run deficits in the next decades. Those deficits continue to widen over time.

Percent of taxable payroll
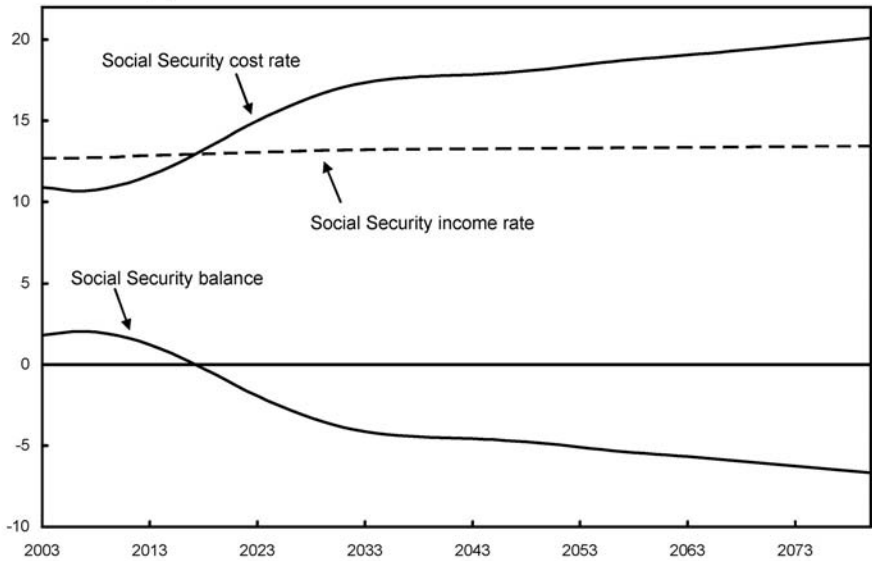
Source: Social Security Administration.

the income rate in 2003 was 12.70 percent. Because the income thresholds at which Social Security benefits become taxable are not indexed for inflation, a greater share of benefits become taxable over time as the price level rises. In 2080, income taxation of benefits is projected to generate 1.03 percent of taxable payroll, resulting in an income rate of 13.43 percent.

The annual balance, the difference between the income rate and the cost rate, is projected to deteriorate. For 2003, the annual balance is 1.81 percent of taxable payroll (12.70 – 10.89). The annual balance is graphed at the bottom of Chart 6-2 as a solid curve that declines over the period. The substantial increase in the cost rate relative to the income rate in the future causes this annual balance to change from surplus to deficit by 2018 and to widen considerably thereafter. In 2080, the annual balance will be -6.67 percent of taxable payroll (13.43 – 20.09, as reported in the Trustees Report, with the small discrepancy due to rounding).

Unless the Social Security system is reformed before that time, the payroll tax would have to rise from 12.40 percent to 19.07 percent to pay all benefits scheduled by current law, even with the assumption that benefit taxation continues under current law to provide a rising share of program revenues. Such an increase represents an expansion of the payroll taxes associated with the program of over 50 percent (Box 6-1).

The annual deficit of 6.67 percent of payroll is the most straightforward way to represent the long-term fiscal challenge confronting the Social Security program. To describe a proposed reform as having restored solvency to Social Security, the reform must greatly reduce or eliminate these annual deficits. The only desirable way to restore solvency is to do so without

---

Box 6-1: The Retirement of the Baby-Boom Generation

It is common in public discussions to associate the financial crisis in Social Security with the approaching retirement of the baby-boom generation, those born between the years 1946 and 1964. This explanation, however, is only partly correct. The problems confronting Social Security are more fundamental than the aging of an unusually large birth cohort. In 2080, for example, the youngest baby boomer will be 116 years old, and almost all benefits in that year will be paid to retirees who were born after the baby-boom generation. Even with virtually no baby boomers among the beneficiaries, Social Security in 2080 is projected to have an annual deficit equal to 6.67 percent of its payroll tax base.

The retirement of the baby-boom generation *does* have an important impact on the system's finances, as can be seen in Chart 6-1. The period of rapid increase in both the dependency ratio and the cost rate occurs during the two decades starting roughly in 2008 when the baby-boom generation becomes eligible for retirement benefits. Chart 6-2 shows that over this same period, the annual balance in Social Security will deteriorate by over 5 percentage points of payroll. If the retirement of the baby-boom generation were the only source of Social Security's financial crisis, then the cost rate would begin to decline as that generation passed away and the dependency ratio fell.

As shown in Chart 6-1, however, the cost rate continues to climb even as the baby boomers age and pass away. The dramatic increase in the cost rate associated with the retirement of the baby-boom generation is, in fact, a permanent transition to an economy in which a higher ratio of beneficiaries to workers makes pay-as-you-go entitlement programs more expensive to maintain. This transition would be more apparent already were it not for the presence of the baby-boom generation in the workforce today. The huge numbers of baby boomers in the workforce have held down the ratio of beneficiaries to workers over the past several decades. Judged from this point forward, the retirement of the baby-boom generation does not cause the financial crisis; it simply makes the long-term problem in the pay-as-you-go system appear sooner rather than later.

---

continued reliance on general revenues. While these numbers are only estimates and are revised over time, recent efforts by the actuaries at Social Security to consider the uncertainty in the projections show that there is essentially no chance that the system will be in balance in the long-term (Box 6-2).

Box 6-2: Long-Term Projections and Uncertainty

Recent experience with short-term forecasts has shown that there is considerable uncertainty about how the economy will evolve. That uncertainty is compounded over the 75-year period that the Social Security actuaries must consider. Traditionally, the Trustees Report has included projections based on three different sets of assumptions—low cost, intermediate cost, and high cost. The low-cost scenario has higher fertility rates, slower improvements in mortality, faster real wage growth, and lower unemployment. All of these changes work to reduce the projected deficits. The high-cost scenario changes the assumptions in the opposite direction and results in larger projected deficits.

Policy discussions seldom include any mention of the low- and high-cost scenarios. Part of the reason is that these alternatives are accompanied by no information on how likely they are to occur. In the 2003 Trustees Report, a new method of dealing with uncertainty was included in an appendix. The method, called *stochastic simulation*, is based on the idea that each of the main variables underlying the projection (like the interest rate or economic growth rate) will fluctuate around the value assumed in the intermediate scenario. These fluctuations are modeled by an equation that captures the relationship between current and prior years' values of the variable and introduces year-by-year random variation, as reflected in the historical period. A stochastic simulation consists of many different combinations of possible outcomes for the random variables. Each combination generates a unique path for the key financial measures, each one analogous to the single assumed path generated by the intermediate-cost scenario. Taken together, these paths represent a wide range of possible outcomes for Social Security.

Chart 6-3 shows the range of outcomes for the cost rate generated by the simulation model. These simulations are based on the assumptions and methods in the 2002 Trustees Report, when the deficits reported in the last year of the projection period (2076) were 1.11, 6.42, and 14.66 percent of taxable payroll in that year for the low-, intermediate-, and high-cost scenarios, respectively. Each curve, starting with the lowest, corresponds to a successively higher percentile of the distribution of outcomes each year. In the last year of the projection

---

Box 6-2 — *continued*

period, the median cost rate is 20.33 percent of taxable payroll, which is slightly higher than the value of 19.84 percent based on the interme- diate assumptions. Overall, 95 percent of the cost rates are between 14.53 and 28.98 percent of payroll. Thus, the low-cost estimate of 14.24 and the high-cost estimate of 28.51 correspond to very extreme outcomes in the overall distribution.

   Modeling the uncertainty underlying the demographic and economic components of the projection is a large step forward in assessing the future obligations of Social Security. The simulation model used in the Trustees Report likely understates the variation that is possible for future costs of Social Security. Nonetheless, the simula- tions show that based on random year-to-year fluctuations, it is highly improbable that the system will have a cost rate below its income rate in the long-term. Uncertainty in the underlying projections only strengthens the case for reform.

Chart 6-3 Probability Distribution of Projected Annual Cost Rates
Simulations that incorporate economic and demographic uncertainty show a wide range of possible outcomes for Social Security's long-term costs.
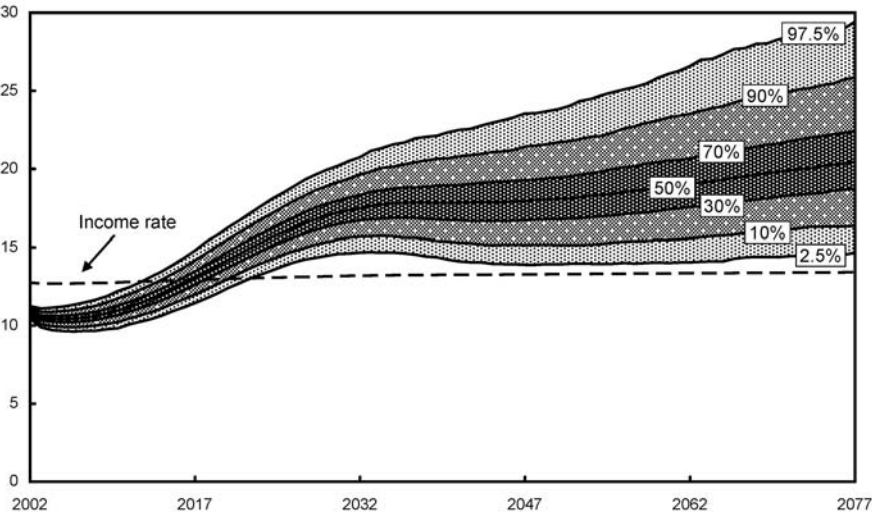
Note: Lines represent percentiles of the distribution, as labeled. Darker shaded areas are more likely outcomes.
Source: Social Security Administration.

Misunderstanding the Financial Crisis

Altough the Social Security program is operated on a largely pay-as-you-go basis, discussions of the financial condition of the program often focus on the trust funds out of which benefit payments are made. There are two trust funds—one for the old-age and survivors benefits and one for the disability benefits—that will be referred to collectively as the "Social Security trust fund." In a year when the government collects more in payroll taxes than it needs to pay out in Social Security benefits (net of the income taxes on benefits), surplus revenues are allocated to the Social Security trust fund. The trust fund is held in a portfolio that consists of special-issue Treasury bonds. The interest rate on the portfolio reflects the yields on long-term Treasury bonds. In a year when Social Security benefit payments exceed revenues, some of the bonds in the trust fund must be redeemed to cover the gap.

In the 2003 Trustees Report, the *trust fund ratio* for the Social Security program was reported as 288 percent for 2003. The trust fund ratio is the proportion of a year's benefit payments that could be paid with the funds available at the beginning of the year. Thus, a trust fund ratio of 288 percent means that in 2003, the amount of bonds held in the trust fund could have been redeemed to cover nearly three years of Social Security benefit payments. A positive trust fund ratio is the standard way of assessing the solvency of Social Security at a point in time. A trust fund ratio of 100 percent is considered to be an adequate reserve for unforeseen contingencies, such as an unexpected drop in payroll tax collections in a particular year.

When the Trustees Report is released, the reaction in the popular press almost always focuses on the date at which the trust fund is projected to go to zero as an indicator of Social Security's financial health. In the 2003 Trustees Report, this date was 2042, and this was widely reported as good news because the prior year's report had projected that date at 2041. The additional year before all of the bonds are redeemed reflects higher annual balances in Social Security through 2042 than were projected in the prior year's report.

Focusing on the date of trust fund exhaustion is inadequate as a measure of Social Security's financial health because this date by itself gives no indication of how dire the fiscal situation becomes *after* the trust fund hits zero. When the trust fund is projected to be exhausted in 2042, for example, the gap between the income and cost rates on the Social Security program is projected to be 4.54 percent of taxable payroll (or 37 percent of the revenues collected by the payroll tax). If such a gap existed in 2003, it would be nearly $200 billion. Reform proposals that are based on pushing back the date when the trust fund is exhausted by a few years will be insufficient to address Social Security's long-term financial imbalance.

As a means of providing a longer-range summary of the finances of the program, the Trustees Report also projects the 75-year *actuarial deficit* in Social Security. Long-range actuarial projections are made over 75 years because this is approximately the remaining lifetime of the youngest current Social Security participants. The 75-year actuarial deficit is equal to the percentage of taxable payroll that could be added to the income rate for each of the next 75 years, or subtracted from the cost rate for each year, to leave the trust fund ratio at 100 percent at the end of the 75-year period.

In the Trustees Report for 2003, this 75-year actuarial deficit was 1.92 percent of taxable payroll using the intermediate assumptions, up from 1.87 percent in the prior year's report. That is, in order to have one year's worth of benefits left in the trust fund in 2077 (the last year of the 75-year projection period starting in 2003), Social Security payroll taxes would have to be 14.32 percent each year for 75 years.

The 75-year actuarial deficit is a widely used measure of the system's financial condition. However, even this measure understates the long-term challenge facing Social Security's finances. Although an increase in the income rate of 1.92 percentage points in each of the next 75 years leaves the trust fund with a positive balance at the end of the 75-year period, the trust fund will rapidly decline to zero in the years after 2077. This occurs because the payroll tax increase of 1.92 percent does not cover the annual deficits of over 6.5 percent that are projected for those years.

Relying on the 75-year actuarial deficit as a guide to solvency is only marginally better than considering the date of trust fund exhaustion. A reform that purported to close the 75-year actuarial deficit would be sufficient only to push the date of trust fund exhaustion to a year just beyond the projection period.

The actuarial deficit over any finite period, even one as long as 75 years, can dramatically understate the financial imbalance in Social Security when the program's annual deficits are getting wider over that period. For example, the 2003 Trustees Report estimates that the present value of the unfunded obligations for the program over the next 75 years is $3.5 trillion. In other words, if this amount of money were available today and invested at the rate of return that is credited to trust fund assets, it would provide just enough to cover the program's deficits over the next 75 years. However, the Trustees Report also estimates that the present value of the program's unfunded obligations over the infinite horizon—the next 75 years and all years thereafter—is $10.5 trillion. The $7.0 trillion difference reflects the continued annual deficits that persist after the first 75 years. Thus, the first 75-year period represents only one-third of the present value of the total shortfall.

A projection period limited to 75 years also biases the discussion of potential reforms in favor of those that are based on pay-as-you-go, rather than

advanced, funding. Some reform proposals would allow a portion of the payroll tax to be used to establish voluntary personal retirement accounts (PRAs). People who establish their own personal retirement accounts would be able to direct some of their payroll taxes into their PRAs in exchange for accepting lower benefits from the pay-as-you-go system in retirement. The additional funding requirements to maintain benefits for current retirees while allowing some of the payroll tax to be used for personal retirement accounts for current workers necessarily appear in the first 75 years. However, much of the benefit of advanced funding—in terms of reduced obligations of the pay-as-you-go system—occurs outside of the 75-year projection period.

Recognizing that even a 75-year actuarial deficit cannot fully reflect the long-term financial shortfalls in Social Security, the Trustees have increased their focus on the annual balance in the last year of the 75-year projection period as a guide to the financial shortfalls in the program. If the trust fund ratio is to continue to play a role in discussions of solvency, then, at the very least, the standard for restoring solvency to the program should be to have not only a positive trust fund in the terminal year of the projection period, but also a trust fund that is not declining toward zero in that year.

The Nature of a Prefunded Solution

To restore solvency to Social Security on an ongoing basis, the income and cost rates cannot be moving apart over time. If the income and cost rates are moving together at the same level, then there is no need for a large trust fund, because the program's annual balance will be roughly zero in each year. As noted above, the annual deficit is currently projected to grow to 6.67 percent of taxable payroll by 2080. Only by reducing annual benefits or increasing the payroll tax (or the income tax on benefits) by a total of 6.67 percent of taxable payroll can solvency be restored in the long term on a pay-as-you-go basis.

If these benefit cuts or tax increases are not desired, then an alternative is to allow the gap between the cost and income rates to persist (provided that it is not increasing over time) and rely on the investment income from a portfolio of assets to cover the gap. Such a portfolio would have to be accumulated in the intervening years, in order to prefund the difference between the program's scheduled obligations and revenues.

In 1983, the last time a major reform of Social Security was undertaken, the program was changed to begin accumulating annual surpluses in the Social Security trust fund. In 2003, the trust fund balance was $1.5 trillion. However, the intervening two decades provide little assurance that the Social Security surpluses during that time have increased the resources available to the government as a whole to pay future benefits.

The balance in the Social Security trust fund has a clear meaning as an accounting device. At any point in time, the trust fund balance shows the cumulative amount of additional revenue—plus interest—the Social Security program has made available to the Federal government to spend on other purchases. The special-issue Treasury bonds in the trust fund are IOUs from the rest of the government to the Social Security program to cover its deficits in future years. The trust fund balance shows the extent of the legal authority for the Social Security program to redeem those IOUs in the future. Administratively, the Social Security program is not authorized to pay benefits unless the trust fund ratio is positive; that is, it can only pay benefits to the extent that it has been a net creditor to the rest of the government.

The question of what the government has done with the revenues made available by past Social Security surpluses has important implications for what the trust fund represents in economic terms and for the design of Social Security reform. There are two competing conjectures about the government's actions. The first is that the surpluses in the Social Security program have had no effect on the surpluses or deficits in the rest of the government's budget. If this is true, every dollar that the government received from past Social Security surpluses and thus allocated to the trust fund served to reduce the amount of Treasury bonds held by the public by a dollar. In the future, drawing down the trust fund when Social Security is projected to run annual deficits simply involves selling the debt back to the public so that when the trust fund is exhausted, the amount of debt held by the public in the future will be the same as it would have been had there not been any Social Security surpluses. Under this conjecture about government budget policy, the Social Security surpluses have been a source of higher national saving and the trust fund represents real resources available to pay future benefits.

The second conjecture is that the surpluses in the Social Security program have encouraged the government to run smaller surpluses or larger deficits in the rest of its budget. If this conjecture is true, the Social Security surpluses have not been used to repurchase existing Treasury bonds held by the public but instead have been used to pay for government expenditures, such as defense, health care, or education. Drawing down the trust fund in future years will involve selling Treasury debt to the public, as in the first case. However, unless future government spending is reduced, the debt held by the public will be higher than it would have been in the absence of Social Security surpluses by the time the trust fund is exhausted. Under this conjecture about government budget policy, the Social Security surpluses have not resulted in higher national saving, and the balance in the trust fund does not represent additional real resources available to pay future benefits.

Analysts have argued in favor of both of these conjectures. Determining which one is correct requires making an assumption about what the

government would have done in the counterfactual case that Social Security had not run annual surpluses. The unified budget deficit (including Social Security) has been the focus of budget discussions for almost all of the last two decades. This provides a strong *prima facie* case that government expenditures outside of Social Security were higher due to the presence of Social Security surpluses during this period.

Allocating Social Security surpluses to special-issue Treasury bonds in the trust fund provides no guarantee that future Social Security obligations are prefunded. It would therefore not be appropriate to simply accumulate government bonds in a trust fund as a way to restore solvency. One way to overcome the vagueness in trust fund accounting is to require that the prefunding occur by allocating a portion of Social Security's annual revenues to the purchase of private rather than government securities and to treat these purchases as annual expenditures of the Federal government. Doing so would break the link between Social Security surpluses and the issuance of debt by the Federal government. This would allow the Social Security program to accumulate a portfolio of financial claims on private sources to pay for future obligations.

Some simple arithmetic shows the size of the portfolio of private securities that would be required to close the entire long-term annual deficit in this manner. Suppose that investments in a portfolio of stocks and corporate bonds earn a 5.2 percent expected return, net of inflation and administrative costs. To obtain an income flow of 6.67 percent of taxable payroll (the annual deficit in 2080) would require a portfolio of assets equal to 6.67/5.2 = 128 percent of taxable payroll. In 2080, taxable payroll is projected to be 34.7 percent of GDP, so that the required stock of assets would be equal to 44.5 percent of GDP. If such a fund existed in 2003, when GDP was estimated to be $10.9 trillion, the fund would have a value of $4.9 trillion. This calculation assumes that either taxpayers or beneficiaries will absorb the financial risk associated with investments in corporate stocks and bonds. Repeating the same arithmetic using a 3 percent real interest rate—the projected return on the Treasury bonds in the Social Security trust fund—shows that the fund would have to be $8.4 trillion.

For portfolios of this magnitude, prefunding by investing in private securities would require that individuals establish their own personal retirement accounts. To put these figures in some perspective, as of November 2003, the net assets of all mutual funds in the United States were estimated to be $7.24 trillion. Thus, in order to cover the annual deficit in 2080 through prefunding, a portfolio the size of at least two-thirds and possibly more than 100 percent of all mutual funds would have to be accumulated. A portfolio of this size is simply too large to be administered centrally without political interference and without disruption to the capital markets.

In light of these issues, a Social Security reform plan should have two components. First, it should restrain the growth of future pay-as-you-go benefits for those not currently in or near retirement to bring the cost rate of the program in line with the income rate in the long term. Second, it should establish personal retirement accounts for each worker. The personal retirement accounts serve a dual purpose. First, because the accounts can be located outside of the government's budget, the accumulation of assets in these accounts would not provide any impetus for higher government spending in the non-Social Security part of the budget. Second, the personal retirement accounts would provide a way for individual workers to accumulate assets to offset the reduction in their total retirement income that otherwise would occur due to the lower benefits in the pay-as-you-go part of the system.

Can We Afford to Reform Entitlements?

While Social Security's long-term solvency has been an ongoing concern for over 25 years, the report of the 1994-1996 Advisory Council on Social Security prompted a new round of policy discussions that included serious proposals to prefund future obligations with private securities. These discussions were bolstered by the appearance of surpluses in the Federal government's budget and budget forecasts during the late 1990s. Shortly after the President took office in 2001, a bipartisan commission on Social Security was established. The commission's final report discusses three reform options that would involve the use of personal retirement accounts to prefund a portion of future benefits.

Some critics of personal retirement accounts have suggested that Social Security reform requires surpluses in the unified budget (including Social Security) or even the non-Social Security portion of the budget to begin investing in the accounts while maintaining pay-as-you-go benefits to current retirees. Since the budget surpluses forecasted a few years ago have not materialized, critics argue that adding personal retirement accounts to Social Security is impossible or impractical. In reality, the need to add resources to the Social Security system is no less pressing now that the surpluses have disappeared; indeed, it may be even more so. The change in the budget outlook makes reform neither less necessary nor less economically feasible.

As an illustration, consider the recent President's Commission's Model 2, under the assumption that all eligible workers will voluntarily choose to establish a personal retirement account (thereby maximizing the transition costs to be discussed below). This plan has two main components. First, it slows the growth of benefits from the pay-as-you-go system by indexing

future benefits to prices rather than to wages. Prices generally increase more slowly than wages. Second, the plan allows workers to receive a tax cut now, if they place the tax cut into a personal retirement account, in exchange for specific reductions in the pay-as-you-go benefits they would receive otherwise. When workers choose this option, private saving is increased. Under the conjecture that Social Security surpluses are saved rather than spent, government saving is reduced and national saving is essentially unchanged. However, the long-term solvency of the pay-as-you-go system is maintained, and government and national saving increase to the extent that having resources go into personal retirement accounts rather than the Social Security trust fund prevents the government from using Social Security revenues to pay for non-Social Security expenditures.

The economic rationale for undertaking this type of Social Security reform does not depend on the current budget situation. This is clearly true with respect to the first component of reform—restraining the growth of future pay-as-you-go benefits to a level that is commensurate with future payroll tax revenues. The value of pursuing this objective does not depend in any important way on whether, due to prior economic and budgetary events not related to the reform, future generations will be paying interest on a large or small stock of public debt. If anything, easing the payroll tax burden on future generations is *more* important if they face a greater interest burden. Relying on personal retirement accounts also remains necessary. Compared to government saving, saving in personal retirement accounts gives workers greater freedom to prepare for their own retirement. Saving in personal retirement accounts also ensures that the additional resources being accumulated for Social Security are not available to be tapped for additional government spending.

Even if both components of reform are still necessary, though, are they feasible? Chart 6-4 shows the plan's effect on the unified budget deficit and total government debt held by the public assuming that the first contributions to personal retirement accounts are made in 2004. Even under the favorable conjecture that Social Security surpluses do not facilitate higher government spending outside of Social Security, the deficit initially increases, but then falls as the reform is fully phased in. At its maximum, in 2022, the incremental deficit increase is less than 1.6 percent of GDP. The higher deficits in turn lead to a greater stock of debt in subsequent years, followed by repayment. The maximum increment to the debt is 23.6 percent of GDP, in 2036.

The hump-shaped pattern for the impact of reform on the deficit reflects the combined effects of the two parts of the reform. Personal retirement accounts widen the deficit by design—they refund payroll tax revenues to workers in the near term while lowering benefit payments from the pay-as-

you-go system in later years. After 2048, an incremental surplus emerges as the benefit reductions phased in through price indexation begin to outweigh the net effect of the personal retirement accounts on the deficit.

Is this temporary increase in government borrowing a problem? Not from an economic perspective. The increased borrowing does not shift any burden to future generations. The tax cuts given to today's workers are paid for by reductions in the share of their future benefits that must be paid from future tax dollars. Nor are current workers harmed. They save this money in their own accounts, which can give them retirement income just as surely as if the government were promising it to them.

While the government's budget situation does not affect the *economic* necessity and feasibility of Social Security reform, under some assumptions about the political constraints on the budget process, the *political* feasibility and desirability of reform may be shaped by the overall budget picture.

Reforms will lead to larger unified budget deficits in the near term but smaller deficits in the long term. The presence of a deficit in the non-Social Security part of the budget may make it more difficult to persuade lawmakers to reform Social Security, if the transition costs of the reform cause the deficit to eclipse a previous record. However, avoiding Social Security reform will not keep deficits in check. If nothing is done to reform Social Security, under current projections, the growth of Social Security,

Chart 6-4 **The Potential Impact of Commission Model 2 on Deficits and Debt**
Relative to GDP, reform initially increases then reduces the deficit and debt.
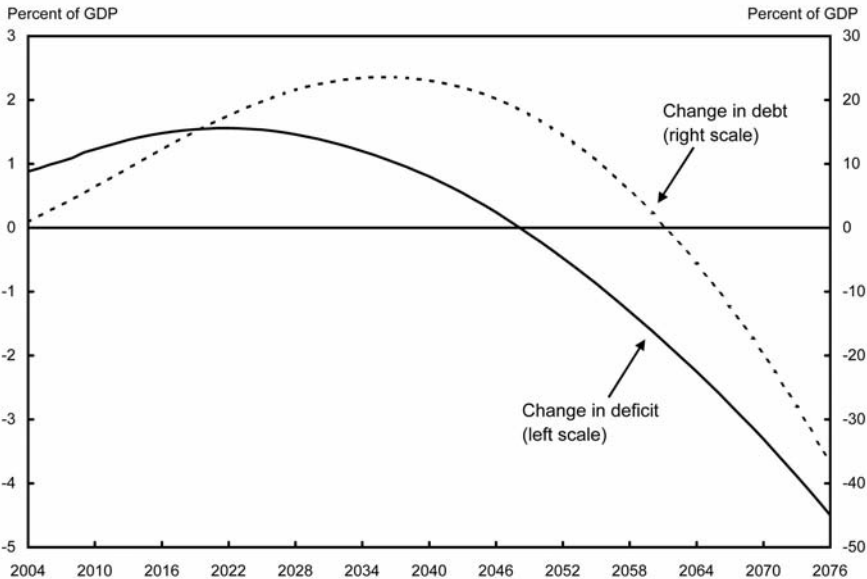
Source: Social Security Administration.

Medicare, Medicaid, and the interest on the borrowing required to finance their growth will lead to unified budget deficits that surpass previous records as a share of GDP.

Chart 6-5 shows the projected costs and revenues in the unified budget under the assumption that no reforms are made to Social Security. The projections are based on the President's policies in the fiscal year 2004 budget, modified to include relief from the alternative minimum tax. The chart assumes that all scheduled Social Security benefit payments are made, financed through additional debt after the trust fund is exhausted. The stacked areas represent total scheduled Federal spending as a share of GDP. Even with nonentitlement spending fixed at 8.1 percent of GDP and excluding interest payments, Federal spending surpasses 20 percent of GDP in 2025, 25 percent in 2050, and 30 percent in 2080. The solid line shows total revenue. The budget deficit, which is the height of the areas above the black line, grows sharply in upcoming decades.

The impact of Social Security reform on the baseline deficit is shown in Chart 6-6, which graphs the evolution of the deficit under two scenarios: the baseline from Chart 6-5 in which no reform is implemented and a reform that includes all of Model 2, with 100 percent participation. Recall from Chart 6-4 that this reform causes the budget deficit to increase temporarily before falling to a lower share of GDP as the reform is fully phased in.

Chart 6-5 The Long-Run Budget without Social Security Reform
The unified budget deficit widens considerably over the coming decades.
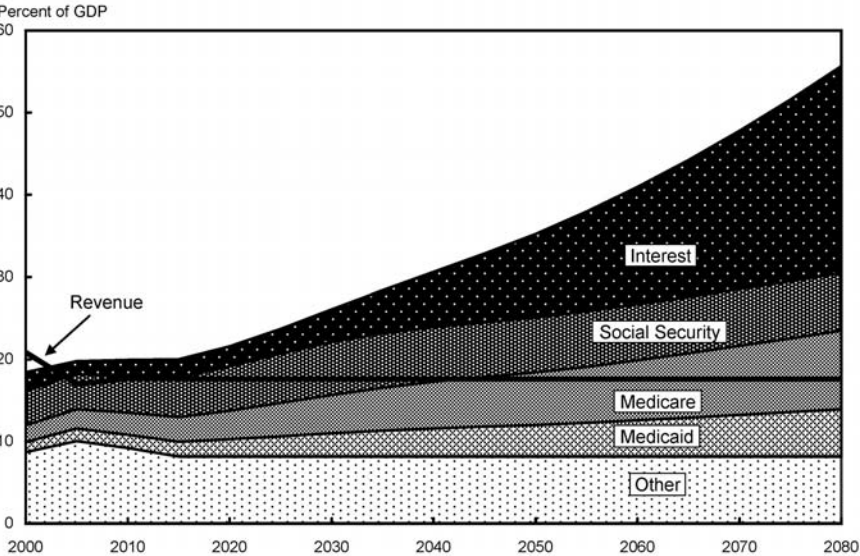
Sources: Social Security Administration, Centers for Medicare and Medicaid Services, Office of Management and Budget, and Council of Economic Advisers.

Chart 6-6 The Long-Run Budget Deficit with Social Security Reform

Enacting Social Security reform leads to lower unified deficits when fully phased in.

Percent of GDP
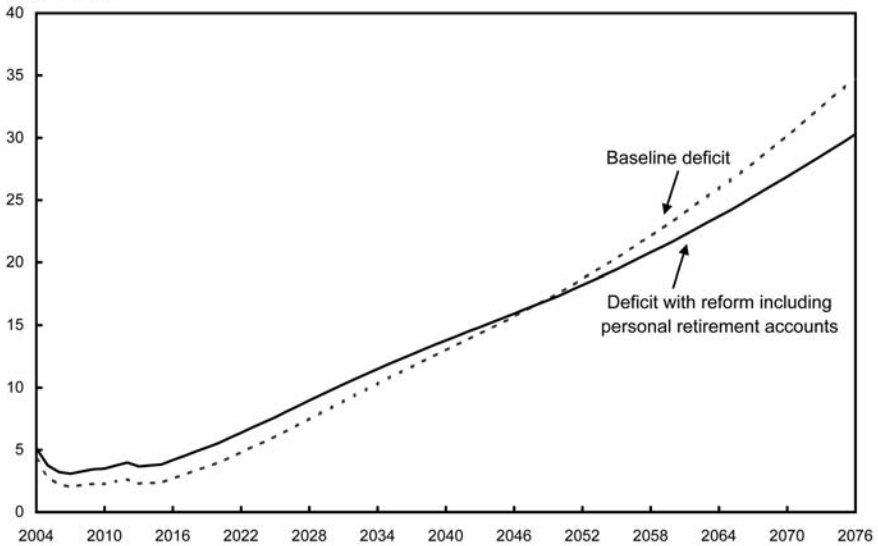
Sources: Social Security Administration, Centers for Medicare and Medicaid Services, Office of Management and Budget, and Council of Economic Advisers.

With the reform, the unified budget deficit reaches 5 percent of GDP in 2019. Without reform, this deficit is reached instead in 2023. The benefits of the reform appear over time, making a positive impact on the Federal budget after 2048.

Policy makers concerned about the unified deficit will have to decide how they will restrain Federal spending over the upcoming decades—they will have to confront this question even if nothing is done to reform Social Security. The benefit of reforming Social Security is that it alleviates, to some extent, the financial burden that unreformed entitlement programs will place on future generations.

Conclusion

The Nation must act to avert a long-foreseen future crisis in the financing of its old-age entitlement programs. The crisis results mainly from the fundamental demographic shifts to lower birthrates and longer lives rather than the impending retirement of the baby-boom generation. However, the scope for enacting meaningful reform will disappear as the baby-boom generation begins to retire and an ever greater share of the population sees its current income arrive in the form of a government check. The design of the Social Security program has failed to keep pace with emerging demographic realities. The benefits promised to those currently in or near retirement must be honored, but a new course must be set to ensure that Social Security is viable and available to Americans in the future.

To do nothing at this point to restrain the growth of entitlement programs would bequeath to future generations an increasing tax on their income to support Social Security. The only way to avoid such an outcome without reducing the living standards of future retirees is to save more today. Greater saving will increase the capital stock and increase the productive capacity of the economy so that it can support those higher payments. The combination of reducing the projected cost of taxpayer-financed benefits and shifting the revenues into personal retirement accounts provides the best mechanism for achieving that result.

Government Regulation in a Free-Market Society

An important reason for Americans' high standard of living is that they live in a free-market economy in which competition establishes prices and the government enforces property rights and contracts. Typically, free markets allocate resources to their highest-valued uses, avoid waste, prevent shortages, and foster innovation. By providing a legal foundation for transactions, the government makes the market system reliable: it gives people certainty about what they can trade and keep, and it allows people to establish terms of trade that will be honored by both sellers and buyers. The absence of any one of these elements—competition, enforceable property rights, or an ability to form mutually advantageous contracts—can result in inefficiency and lower living standards. In some cases, government intervention in a market, for example through regulation, can create gains for society by remedying any shortcomings in the market's operation. Poorly designed or unnecessary regulations, however, can actually create new problems or make society worse off by damaging the elements of the market system that do work.

The key points in this chapter are:

- Markets generally allocate resources to their most valuable uses.
- Well-designed regulations can address cases where markets fail to accomplish this goal.
- Not all regulations improve market outcomes.

How Markets Work

Free markets work through voluntary exchange. This voluntary nature ensures that only trades that benefit both parties take place: people give up their property only when someone agrees to exchange it for something that they value more highly. In most transactions, sellers receive money rather than goods in exchange for their property. Sellers then use that money to become buyers in other transactions.

What ensures that producers are providing the commodities that consumers want? Market prices play the critical role of coordinating the activities of buyers and sellers. Prices convey information about the strength of consumer demand for a good, as well as how costly it is to supply. By conveying information and providing an incentive to act on this information, prices induce

society to shift its scarce resources to the production of goods that are valued by consumers. In this way, markets usually allocate resources in a manner that creates the greatest net benefits (benefits minus costs) to society. An *efficient* allocation is one that maximizes the net benefits to society.

In general, efficiency requires that the price of a good reflects the incremental cost of producing that good, including the cost of inputs and the value of the producer's time and effort. In this way, prices induce consumers to economize on goods that are relatively expensive to produce and to increase their purchases of goods that are relatively inexpensive to produce. A key advantage of free-market competition is that it generally leads to a situation in which price equals incremental production cost. This outcome occurs because in a competitive market environment, a seller who charges a price above the cost of production will be undercut by competitors, including new entrants. In contrast, if prices are artificially high because of limited competition, consumers will buy less of the good than they would if they faced the competitive price. Furthermore, some consumers who would benefit from buying the good at a competitive price may not buy it at all.

When market conditions change, prices usually change as well and signal buyers and sellers to modify their behavior. For example, if a disruption in the gasoline supply were to occur and prices and behavior remained unchanged, there would not be enough gasoline supplied to satisfy consumer demand at predisruption prices. The result would be a gas shortage. To eliminate this shortage, some form of rationing would be required to ensure that the quantity of gasoline demanded by consumers matched the quantity of gasoline provided by suppliers.

In a market economy, rationing is done by prices. As prices of gasoline increase, two changes in behavior typically occur. First, consumers as a whole reduce their consumption of gasoline, and second, producers as a whole increase the quantity of gasoline available for sale. These aggregate changes are the result of many individual decisions. For example, some consumers may carpool, others may cancel trips, and some may be willing to spend more on gasoline to continue on as before. On the supply side, producers may ship gasoline from areas not affected by the supply disruption, refineries may increase production, and firms may lower inventories of gasoline in storage. Eventually, prices increase to the point at which the reduced quantity of gasoline demanded equals the increased quantity of gasoline supplied. In a market economy, all of this happens without any centralized control mechanism.

Market Imperfections

Sometimes markets do not allocate resources efficiently. Under such circumstances, it may make sense for the government to intervene in markets beyond providing a legal foundation for market transactions. Chapters 8 and 9, which deal with energy and the environment, discuss some regulations designed to address two such market failures—externalities and market power. These chapters look at both the benefits and potential problems that can result from imposition of regulations.

Poorly designed or unnecessary government regulations can actually reduce society's overall well-being. The possible costs of government regulation include the costs imposed on consumers and producers, impeded innovation, and unintended negative consequences such as the creation of unforeseen barriers to competition. It is essential to consider whether the costs potential regulations impose on society are greater than the benefits society receives from fixing any market failures.

Regulation and Externalities

Externalities (also known as *spillover effects*) can lead to a situation in which the price of a commodity does not reflect its full incremental cost to society. A *negative externality* exists when the voluntary market transaction between two parties imposes involuntary costs on a third party. For example, a power plant might produce and sell electricity to consumers to both their advantage, but the production process might emit air pollution that negatively affects the population. The costs that this pollution imposes on the population might not be considered when the firm decides where to locate a plant, which technologies to use, or how much electricity to produce. It could be that if these costs were taken into account in the same way as all of the other costs of producing electricity, the plant might be relocated to a place where its pollution would affect fewer people, the firm might put greater emphasis on pollution-reducing technologies, or the plant may not produce as much electricity. The existence of a negative externality can lead to an outcome that is worse for society than one that takes the externality into account.

As discussed in Chapter 9, *Protecting the Environment*, in many cases the best remedy for externalities is to define property rights and allow the affected parties to transact privately to achieve a mutually beneficial outcome. Sometimes, however, establishing property rights can be expensive. Even with clearly defined property rights, it may be costly for affected parties to collectively agree on a mutually beneficial transaction. Under such circumstances, other forms of government intervention may be appropriate, including taxes, subsidies, and direct regulation.

Addressing Externalities Through Taxes

One approach to dealing with externalities would be to levy a tax (known to economists as a *Pigouvian tax*) on market participants such that the amount of tax collected equals the incremental cost of the externality. For example, if a power plant's emissions are easy to monitor and the costs of pollution are easy to assess, the tax on each unit of pollution could be set equal to the cost of the externality. Alternatively, if the amount of pollution is not easily monitored, the tax could apply to each unit of production (each kilowatt produced by the plant, for example) rather than the pollution itself, and could be set equal to the additional external cost of pollution from each unit of production.

In general, taxes distort economic activity (see the discussion of the income tax in Chapter 4, *Tax Incidence: Who Bears the Tax Burden?*). However, proponents of Pigouvian taxation argue that it can improve the allocation of resources by forcing producers and consumers to confront the full costs of production. Indeed, some advocates of the use of such taxes go further and argue that revenues from Pigouvian taxes could be used to finance a reduction in the rates on other taxes that do distort behavior, such as the income tax. This idea is sometimes called the *double-dividend hypothesis* because it increases efficiency in the market with the externality *and* in the markets that are distorted by the income tax.

This argument must be viewed with caution. To see why, recall that Pigouvian taxes drive up the prices of the goods that are produced using technologies that involve pollution. The increase in prices reduces the buying power of households' incomes. This is effectively a decrease in the real wage rate because a given dollar amount of wages buys fewer goods and services. Put another way, Pigouvian taxes are, to some extent, also taxes on earnings. If the labor market is already distorted because of an income tax (as is the case in the United States and other industrial economies), the Pigouvian tax makes the distortion worse. In some cases, the added distortions in the labor market can actually outweigh the gains from correcting the externality. The desirability of Pigouvian taxes as a policy instrument must be determined on a case-by-case basis.

Addressing Externalities Through Limits on Quantity

Another possible problem with Pigouvian taxes is that determining their magnitude can be challenging because it may be difficult to measure the amount of pollution, as well as the value of the damage it causes. Moreover, the appropriate tax may change with market conditions. If, for example, the cost of the externality increases with output, the optimal tax would need to go up if output increases.

It is also difficult to know beforehand the tax level that will reduce emissions by the desired amount. Moreover, as the economy changes, the tax will need to be adjusted to maintain the desired amount of emissions reduction. A system in which a firm must own a government-issued permit for each unit of pollution addresses these problems because the government determines the number of permits to create. A *cap-and-trade* system, which allows firms to trade these permits, accomplishes the environmental goal at least cost.

Addressing Externalities Through Subsidies

Another option for dealing with externalities is to subsidize alternative behaviors that do not produce the negative externality. For example, concern over externalities from fossil fuels has led to government subsidies of some alternative sources of electricity, such as wind and solar power. However, such subsidies have some limitations. First, using the example of electricity, subsidies encourage overconsumption by keeping the cost of electricity below the level that market forces would set if the costs of the externality were taken into account. Second, subsidies raise some difficult administrative issues. In particular, the government needs to identify all the behaviors that should qualify for a subsidy. In the case of the power plant that emitted pollution, a fully efficient policy would be to subsidize all other ways of generating electricity and all conservation activities. Such attempts quickly become unwieldy in practice.

Addressing Externalities Through Command-and-Control Regulation

The government can also attempt to limit negative externalities with *command-and-control* regulations that mandate certain behavior. For example, the government requires automobile producers to meet overall fuel-efficiency standards. There have also been proposals to mandate that a certain percentage of electricity be generated by renewable fuels such as wind and solar power.

Command-and-control regulations can sometimes be the only way to deal with an externality. In general, however, they should be avoided because they discourage flexible and innovative responses to externalities and can result in higher costs than alternative policies. For example, mandating use of a particular technology to lower emissions could lessen firms' incentives to develop more effective techniques to reduce pollution. Furthermore, people adapt to command-and-control regulations in unintended ways that can limit their effectiveness over time. For example, one unintended consequence of the automobile fuel-efficiency standards was to increase the demand for light trucks and sport utility vehicles (SUVs), which were not as stringently regulated.

Regulation and Market Power

Market power, which arises in the presence of impediments to competition, is another potential source of inefficiency in a free-market system. Firms that have *market power* typically have the ability to charge prices above the competitive price level and maintain those high prices profitably over a considerable period. In some cases, the impediment is a law that makes it difficult for competitors to enter a market, but market power can also arise from the nature of the industry itself. For example, the high cost of wiring residential neighborhoods for electricity makes it unlikely that multiple firms would be willing to compete to distribute retail electricity. In these cases, regulation can be useful to prevent firms with market power from charging consumers prices that substantially exceed the cost of providing the good.

Policy makers need to recognize, however, that regulations themselves affect firms' and consumers' behavior and incentives. Regulations that do not take these effects into account can result in excessive consumption, misaligned incentives, stunted innovation and investment, and needless waste. Even regulations that do account for these effects may be rendered obsolete or counterproductive by changes in the industry that occur over time. For this reason, it is important to periodically reevaluate regulatory policies. Chapter 8, *Regulating Energy Markets*, discusses opportunities for reevaluation in further detail.

Regulation in the Absence of a Market Failure

Some government regulations attempt to reverse what would otherwise be efficient market outcomes due to beliefs that a particular market-based allocation of resources is undesirable. For example, regulations to prevent "price gouging" might be seen as fair, but the economic consequences of these regulations must be recognized (Box 7-1). Attempts to circumvent the market in this way must confront a basic reality—resources are scarce, so that if market prices are not used to ration commodities, some other mechanism has to be used instead. For example, resources could be allocated to consumers using ration coupons, a lottery, or first-come, first-served. Resources could also be allocated based on cronyism or other discriminatory means. These nonprice methods cannot guarantee that the scarce resources go to the consumers who value them the most. Furthermore, they reduce suppliers' incentives to increase production. For example, if prices are capped, suppliers may not work overtime to increase supplies or pay extra transportation costs to bring in supplies from distant areas. As a result, resources are not put to their best uses.

Box 7-1: Market Responses to Unexpected Shortages

When there are large, unexpected increases in demand or decreases in supply for a good, a normal market response is for prices to increase by enough to restore balance between supply and demand. Consumers might accuse sellers of "price gouging" when such price increases occur in response to a natural disaster or a failure of supply infrastructure. A number of states have laws that make price gouging illegal. Even without such laws, some businesses might choose not to increase prices during an emergency for fear of a consumer backlash. If prices do not increase, however, consumers do not receive a signal to cut their consumption and suppliers might not have the proper incentives to increase supply adequately.

By not allowing market forces to restore the balance between supply and demand after the shock, nonprice rationing must be implemented instead. For example, after a pipeline break reduced the supply of gasoline into the Phoenix, Arizona, area in August 2003, press reports indicated that some stations ran out of gasoline, consumers waited in line for hours, and some drivers started following gasoline tankers as they made their deliveries.

Changes in demand can induce shortages as well. For example, in the days leading up to the arrival of Hurricane Isabel in the Mid-Atlantic states in September 2003, press reports indicated that many retailers sold out of flashlights and D batteries. The flashlights and batteries went to the first people to show up at the store, rather than to those who valued them the most. It also meant that people who were able to buy the goods might have bought more than they would have at the higher price, leaving fewer for others. Without price increases, there was no mechanism to allocate the available goods to their highest-valued uses. For example, if prices were higher, early customers may have decided not to buy new batteries for their fifth flashlight and later customers would not have been forced to sit in the dark.

While allowing prices to increase in the face of a natural disaster or a supply disruption may seem unfair, the alternative would be to restrict the allocation of scarce supplies and to possibly keep supplies from those who need them most. Artificially low prices remove incentives for consumers to conserve and for suppliers to meet unfilled demand, potentially prolonging the shortage. Society must decide whether the perceived fairness resulting from regulations to hold down prices is more important than allowing the market to provide incentives for resolving the shortage as quickly as possible, while making sure that scarce resources are available for those who value them the most.

Conclusion

In general, market systems allocate resources toward their most highly valued uses. Importantly, no one *directs* society to this result. Rather, it is the outcome of a process in which each consumer and each producer observe prices and privately make the decisions that maximize their well-being. The coordination of economic activity is done by prices, which provide signals of the costs to society of providing various goods. However, in the presence of market power, externalities, and other types of market failure, market-generated prices may not incorporate all of the relevant information about costs. Under these conditions, there are opportunities for government to intervene and improve the allocation of resources.

The fact that the market-generated allocation of resources is imperfect does not mean that the government necessarily can do better. For example, in certain cases the costs of setting up a government agency to deal with an externality could exceed the cost of the externality itself. Therefore, proposed remedies for market failure must be evaluated on a case-by-case basis.

Energy and the environment are two areas in which government intervention may play a role in correcting market failures. Such interventions are likely to be more successful when they harness market forces to the extent possible. The next two chapters illustrate the challenges in properly designing regulations in these areas. An important implication of the analysis of both chapters is that in order to make society better off, regulatory policy must be based on a solid economic foundation.

Regulating Energy Markets

Energy is essential to the U.S. economy, both as a final good and as an input into the production of most other goods. In 2000, energy expenditures equaled $703 billion, or 7.2 percent of GDP. The markets that provide this energy function well and are generally competitive. However, parts of the energy industry have characteristics that are associated with market failures. For example, the large fixed costs required to construct distribution networks for electricity and natural gas make it unlikely that more than one firm would be willing to invest in the infrastructure needed to serve residential customers in a particular area. The distribution company, therefore, may have *market power*, the ability to charge prices significantly above the competitive price level and profitably maintain those prices for a considerable period. Another type of market failure involves *negative externalities*, costs that economic transactions impose on third parties that the parties to the transaction do not face. For example, energy producers and consumers may not fully take into account the fact that burning fossil fuels may cause acid rain or smog.

This chapter discusses economic issues relevant to several different energy markets, including natural gas, gasoline, electricity, and crude oil. The use of these different types of energy involves different market structures and different potential market failures. An important focus of the chapter is on the design of regulations to address market failures in energy markets while minimizing disruptions to the market. The key points in this chapter are:

- Markets generally work well for energy products, which in most ways are like other products in the U.S. economy. While some aspects of energy markets may require regulation, most segments of these markets function well without regulation.
- Federal, state, and local regulations can have conflicting goals. If the conflicting goals are not balanced, competing regulations could lead to worse problems than the market failures the regulations attempt to address.
- Regulations need to be updated as markets evolve over time to ensure that the original goals still apply and that these regulations are still the lowest-cost means of meeting those goals.
- The United States benefits from international trade in energy products.

Market Forces and Regulation
in the Market for Natural Gas

Some energy markets require regulation. For example, because of the high cost of natural gas distribution services, the market generally supports only one local distribution company. Thus, the delivery infrastructure, including pipelines and gas meters connected to individual residences, is regulated. However, certain segments of the natural gas industry are amenable to competition. They do not require regulation even though the distribution segment does. Indeed, in many areas of the country parts of the natural gas market have been deregulated. For example, producers of natural gas are no longer subject to price regulation. Furthermore, although prices for transporting natural gas to homeowners are regulated, in some states multiple firms now compete for the right to sell the gas to homeowners. This type of partial deregulation has also been applied to electricity markets; in many areas, local distribution lines are still regulated while generation and retail marketing are deregulated.

The last year has demonstrated how market forces have worked to allocate scarce resources in the natural gas market. Demand for natural gas is highly seasonal, with the greatest consumption by far during the winter heating season. During the summer, a portion of natural gas production is stored for use in the following winter. Natural gas inventories in spring 2003 were unusually low after a colder than normal winter in 2002-2003. This led to large increases in natural gas prices in the spot and futures markets. In turn, these high prices encouraged consumers to switch to other fuels or reduce consumption over the summer, encouraged producers to increase production, and encouraged importers to bring in additional natural gas from outside North America. In combination, these actions resulted in a near-record increase in natural gas inventories in time for the winter heating season. As a result, the United States entered the winter of 2003-2004 with slightly above-average natural gas inventories. High prices have also given firms an added incentive to invest in new projects, such as liquefied natural gas (LNG) facilities, to bring additional supplies of natural gas to the market in the future.

Market Forces and Regulation
in Gasoline Markets

Recent and past events in the gasoline market have shown how unexpected shortages affect market prices and how government regulation can make the situation worse. Wage and price controls imposed in the early 1970s to combat inflation included government regulations that kept gasoline prices below the market level. As a result, when oil supplies were disrupted in 1973 and 1979 by geopolitical events in the Middle East, consumers wanted to buy more gasoline than suppliers were willing to supply at the artificially low prices.

Regulations that prevented suppliers from increasing prices meant that consumers had to wait in lines or face limits on the amount of gasoline they could purchase. As a result, some gasoline likely went to consumers who valued it less than other consumers because those who would have cut consumption as prices rose continued to buy gasoline at the artificially low price. Keeping gasoline prices artificially low also reduced the incentive for oil companies to refine new sources of crude oil into gasoline—a supply response that would have lessened the shortfall.

Gasoline markets also demonstrate how markets react to unexpected changes in supply when prices are not regulated. For example, several refinery problems on the West Coast in recent years have, on occasion, temporarily reduced the supply of California Air Resource Board (CARB) gasoline that meets strict California specifications for reducing air pollution. After these disruptions, prices typically increased quickly, and usually stayed high for only a matter of weeks. These increased prices led consumers to reduce their gasoline consumption.

During supply disruptions that were expected to last a relatively long time, the high prices also led distant refineries to produce and ship CARB gasoline to California. These refiners had to shift their operations to make CARB gasoline instead of their normal product, find an available tanker, and then ship the gasoline to California—a process that takes three weeks or more. High prices rewarded the refiners that were able to get CARB gasoline to California quickly, while refiners whose shipments arrived too late (that is, as prices started to come down again) would lose money. The price spike provided an incentive for distant refiners to risk making and shipping CARB gasoline to California, thus helping to alleviate California's gasoline shortage.

Local and Federal Regulations May Conflict

As illustrated in the example above, not all gasoline sold in the United States is the same. Differences in local specifications are often the result of how local and state governments have responded to the Clean Air Act of 1990. Chart 8-1 shows which areas of the United States have adopted different fuel specifications. Flexibility in how localities address air pollution abatement allows them to implement an approach that best meets their needs. However, different local or regional gasoline specifications add complexity to the national gasoline production and distribution infrastructure, reducing the reliability and availability of gasoline supplies.

The proliferation of fuel varieties produced for various locations (called *boutique fuels*) reduces the number of potential suppliers of each particular fuel and slows the industry response when there are local or regional disruptions to the gasoline supply. Boutique gasoline specifications likely contributed to the price spike in the Midwest in 2000, which occurred after several refineries experienced production problems around the same time that two major pipelines supplying the region went out of service. Chicago and Milwaukee were particularly hard-hit in part because of their local

Chart 8-1 Required Specifications for Gasoline
Different local environmental regulations have lead to a patchwork of gasoline specifications.

RFG - North	RFG w/Ethanol
RFG - South	NV CBG
Oxygenated Fuels	7.2 RVP
CA CBG	7.0 RVP
RFG/CA CBG	7.8 RVP, MTBE-No Increase
AZ CBG	7.8 RVP
Oxy Fuels/7.8 RVP	7.0 RVP, 30 ppm S
Oxy Fuels/7.0 RVP	300 ppm S
Conventional	800 ppm S
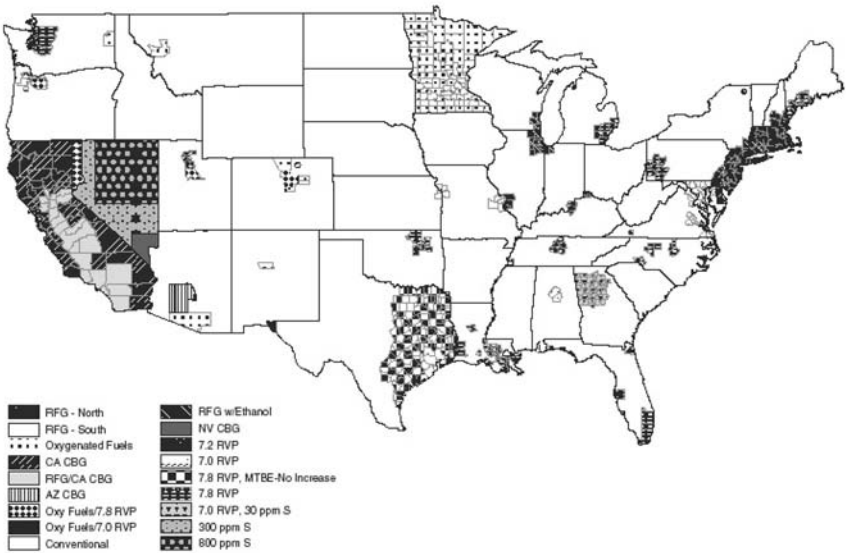
Note: RFG refers to reformulated gasoline, CBG to cleaner burning gasoline, RVP to Reid vapor pressure, MTBE to methyl tertiary butyl ether, and ppm S to parts per million sulfur.
Source: ExxonMobil, November 2003.

requirements for reformulated gasoline using ethanol. Nearby cities using reformulated gasoline had different specifications, so that existing reformulated gasoline stocks could not be shipped to the area.

The impact of boutique fuel regulations demonstrates that there may be benefits from standardizing regulations across geographic areas for goods that are sold regionally or nationally. Gasoline markets in the eastern half of the United States are interconnected by pipelines, barges, and tankers. Reducing the number of gasoline specifications could allow for increased flexibility of the gasoline supply system. For example, production lost because of a refinery problem in Chicago could be offset by shipments of gasoline from refiners in other areas. The President's National Energy Plan asked the Environmental Protection Agency (EPA) to study ways to increase the flexibility of the Nation's fuel supply.

While there may be benefits from standardizing regulations across geographic areas, standardization may require some areas to use gasoline that is more expensive than necessary to meet local air-quality standards. The benefits of standardization must be weighed against any increased costs.

Local and State Regulations Lead to Different Market Outcomes

State regulations can also increase the cost of marketing and distributing gasoline to consumers. For example, several states and the District of Columbia have divorcement laws that restrict refiners' ability to own and operate retail stations. These regulations have been found to increase prices at the pump; prices in states with divorcement laws are almost 3 percent higher than they would be without such laws. Similarly, regulations in Oregon and New Jersey ban self-service gasoline sales because of putative safety and environmental concerns. Economists have estimated that gasoline prices in these states are between 2 and 6 cents per gallon higher than they would be without the self-service ban (gasoline prices in New Jersey are lower than in surrounding states because of New Jersey's low gasoline taxes, but prices would be even lower if self-service were allowed).

Market Forces and Regulation in Electricity Markets

While a mix of market forces and well-designed regulation can lead a market with market failures to perform more effectively and efficiently, improper regulation can lead to worse outcomes than even an imperfect market without regulation. The market for electricity is a case in point.

Some existing regulations in the United States have the unintended effect of making the Nation's electricity supply less reliable and more expensive. The same attribute that makes competition in electricity difficult to achieve—provision of electricity over a single network on which the amount of electricity supplied must equal the amount of electricity consumed at every moment—makes the consequences of poorly designed regulation particularly costly. For example, California's rolling blackouts in January 2001 appear to have stemmed in part from regulations that fixed retail electric rates. As a result, there was an insufficient supply of electricity during the daily peak periods of demand. Fixed retail electric rates provided little incentive for consumers to reduce their consumption of electricity during these high-usage periods.

The Evolution of the Electric Industry from Local to Interstate Markets

As the electric industry has evolved from local, largely self-contained systems to a more national, integrated system, the appropriate combination of state and Federal regulations has changed as well. For many years, electricity was provided by integrated utilities—local monopolies that generated power and distributed it to residents and companies in a specific area—that were regulated by state public utility commissions.

Over time, a high-voltage transmission network linking the local monopolies developed. The network was originally designed to boost reliability, but it has also had the effect of reshaping the economics of the electricity market. The existence of this network (called the *transmission grid*) gave rise to a market for wholesale electricity through which utilities could buy electricity generated elsewhere for use by their own customers.

Regulatory changes complemented the technological and structural changes to make the electricity business more competitive. In 1978, new Federal regulations mandated by the Public Utilities Regulatory Policies Act (PURPA) required state-regulated utilities to buy power generated using renewable energy sources and cogeneration plants (plants that produce electricity while producing other products such as steam heating). These regulations led to an expansion of wholesale markets in which regulated utilities bought electricity generated by other firms and demonstrated that independent electricity generators could coexist with existing state-regulated utilities. In the late 1980s, Federal regulators began revising regulations to encourage the development of independent producers more generally. In 1996, Federal regulators began requiring the public utilities that owned transmission lines to make them available to independent electricity generators. Today, more than half of all the electricity generated is exchanged on the wholesale market before it is sold to consumers.

Electricity Regulation in an Evolving Market

Wholesale electricity generation will become more efficient over time as unregulated generating companies add new capacity based on competitive market signals. Market signals will influence both the timing of when new generation capacity is built and the type of fuel these plants will use. For fully regulated electric utilities, these decisions are made with the approval of local or state regulators. Without the discipline of competitive markets, regulated utilities are able to pass increased costs on to consumers regardless of whether the utilities have made the most efficient choices.

Effects of Regulation on Transmission Capacity

Regulations in the electricity market continue to impose barriers to competition and greater efficiency. Today's regulatory structure may not encourage regulators in one jurisdiction to take into account the full effects of their actions on the rest of the transmission grid because the regulatory system is based on an industry structure that no longer exists. For example, the transmission grid crosses state boundaries, so what happens in one state affects the residents of other states. However, state regulators might not consider the costs and benefits of their actions on citizens of other states. As a result, regulation of the transmission grid has not kept up with changes in the market.

Extensive blackouts in the Northeast and Midwest in August 2003 and in the West in August 1996 demonstrated the potential costs of not updating and coordinating Federal, state, and local regulations. Despite the growing demand for electricity and the growing demand for transmission capacity to satisfy the wholesale market, construction of new transmission facilities has declined by about 30 percent since 1990. The current mix of regulations has facilitated increased use of transmission capacity, but has not done enough to encourage companies to invest in building new capacity. For example, some state and local regulations have discouraged the construction of new local facilities, thus encouraging increased transmission from more distant locations.

State deregulation may also give local utilities the incentive to import lower cost electricity from generators in other states. The growth of interstate transmission of electricity has increased the need for Federal, state, and local governments to coordinate their regulations that affect the interstate transmission grid.

Another problem with existing regulation is that state and Federal regulators approve transmission rates to provide the owners of transmission lines a fixed rate of return, but the chosen rate may not be high enough to encourage firms to invest in sufficient new transmission capacity. One factor that is not fully considered in rate-of-return calculations is the lengthy and uncertain permitting process that requires companies to deal with multiple regulators. Because these costs are not fully accounted for, the effective rate

of return often is too low to attract investment. Such regulatory uncertainties are just one of many factors that make investing in new transmission capacity risky. Higher rates of return may be needed to spur investment.

Insufficient investment in new transmission capacity is not the only problem stemming from improper regulation of rates of return. Such regulation may also prevent investment from being channeled to areas that most need new transmission capacity. Higher prices for use of the most congested parts of the grid would reduce transmission over these parts of the grid and send a signal to potential investors to expand capacity in those areas. Grid operators in some parts of the country now use locational marginal pricing to set prices in different locations based on both the cost of generation and the cost of congestion. Areas that are served by congested transmission lines pay higher prices reflecting the cost of such congestion.

Congestion in the transmission grid leads to both lower reliability and less competition. The lack of competition results from the low-cost generators' inability to send power to high-cost areas, forcing the high-cost areas to use less efficient, locally-produced electricity. Adding new transmission capacity between low-cost and high-cost areas could increase prices in low-cost areas in the short run. However, these price increases would likely lead to new generating capacity being built in low-cost areas, reducing prices back toward existing levels.

Regulations That Require Updating

As electricity markets have become more competitive, Federal regulations designed to prevent utilities from abusing their government-granted monopoly power may have ceased to serve the public interest. For example, the Public Utilities Holding Company Act (PUHCA) was originally passed in the 1930s to limit the size and type of operations in which a public utility may engage, including the types of companies that can own utilities. Today, these limits may actually increase prices to consumers by preventing utilities from engaging in activities that could make their businesses more efficient. These limits also may prevent public utilities from expanding their operations in ways that would increase competition in other parts of the country.

The evolution of the electric power industry from a natural monopoly to an increasingly competitive market calls for regulations that facilitate rather than hinder efficiency and innovation. The Federal Energy Regulatory Commission (FERC) is working on new regulations for wholesale electricity markets with the goal of having market forces encourage the lowest-cost generators to provide electricity.

Demand Response to Electricity Production Costs

Many residential electric rates today are fixed throughout the day at a level based on the average cost of generating and delivering electricity to the residential customer. The cost of producing electricity, however, is not fixed throughout the day. Instead, electricity generators constantly adjust production to meet demand hour by hour or even minute by minute. As a result, the marginal cost of electricity production—the cost to produce one extra unit of electricity—varies widely over the course of a day. Wholesale prices reflect this, with lower prices in the middle of the night (a period of low demand) and higher prices in late afternoon (a period of peak demand). Under the current regulatory structure, however, many consumers are charged the same rate regardless of the wholesale cost of electricity so that utilities cannot raise prices to reflect the true cost of generation. As a result, local regulated utilities must have access to enough generating capacity to meet peak demand, as well as enough transmission and distribution capacity to get the electricity to all customers. Chart 8-2 illustrates the fluctuations in electricity consumption and wholesale prices over a week in August 1999. During the week illustrated, the regulated utilities were at times forced to sell electricity at a loss because wholesale prices rose above the fixed retail rate.

It is not cost-effective to store large quantities of electricity. Therefore, the requirement that electric utilities meet all demand at fixed retail prices

Chart 8-2 Hourly Electricity Consumption, Wholesale Prices, and Retail Prices in California
The volume and wholesale price of electricity consumed fluctuate throughout the day, while retail prices are fixed, so that residential consumers have no economic incentive to change their consumption as costs change.

Note: Residential retail price is the monthly average for August 1999.
Sources: California Energy Commission, California Power Exchange, and Department of Energy (Energy Information Admistration).

means that they must build enough capacity to meet the highest peak demand during the year. They also need to maintain reserve capacity to offset any supply lost due to generation or transmission problems. Some of this capacity is only required during the relatively few hours of the year when demand peaks, for example, on the hottest days in August.

Without the ability to increase retail prices during peak demand to encourage consumers to cut their energy consumption, insufficient generation capacity would lead to a rationing of supply, for example, rolling blackouts. While some electric utilities offer time-of-day pricing, a system in which retail rates are higher during periods of peak demand, these prices do not vary with the actual cost of generating electricity on a particular day. These programs reduce the average peak demand but do not provide the needed incentives to cut power usage on days with extreme peak demand.

Some consumers also receive lower rates in exchange for allowing the electric company to interrupt their service if wholesale costs increase above a certain level. There are some programs that allow the utility to cut off all of a consumer's power, while others simply allow the utility to turn off the consumer's air-conditioning. There are also typically limits on how long or how many times the utility can cut off power. These programs to reduce peak electricity usage thus represent only a partial implementation of variable pricing.

A reduction in peak demand achieved through variable pricing would allow regulated utilities to build less generation, transmission, and distribution infrastructure. Because they cannot increase retail prices, these utilities use other means to reduce peak demand, such as rebates to consumers who purchase energy-efficient appliances or incentives to improve weatherization of homes. While these programs reduce peak demand by increasing energy efficiency, they do not use the market to determine which ways of cutting electricity demand would have the lowest cost. Furthermore, as electricity markets evolve, there may no longer be one firm that can capture all of the benefits from reducing peak demand. As a result, these programs may not be able to continue because individual companies have less incentive to implement them.

Current programs that attempt to reduce peak demand still leave customers unaffected by changes in the cost of production until shortages and interruptions in service result. If retail prices were allowed to increase, consumers could decide to cut their consumption (possibly to zero). This approach could improve overall welfare by reducing the number of peaking plants needed; that is, it may be less costly to curtail demand than to add to supply by building expensive generation capacity that is rarely used. However, for variable pricing to be completely implemented, new meters and smart appliances may be needed so that consumers can acquire the information and technology needed to adjust their usage as electricity prices change.

Energy and Trade

The United States benefits greatly from global trade in energy markets. Gasoline and diesel fuels refined from crude oil are currently the most widely used transportation fuels. By importing petroleum, U.S. firms are able to continue to supply gasoline and diesel at real prices comparable to historical averages, even as environmental regulations have increased the costs of refining. Adjusted for inflation, gasoline prices are much lower than at their peak in 1981. However, this beneficial trade requires reliance on imports that could be subject to supply disruptions.

Because crude oil is traded throughout the world, its price is affected by global changes in supply and demand. Disruptions to the supply of oil from areas that do not supply the United States would affect domestic prices of oil even if U.S. imports are not directly affected. Indeed, domestic prices of oil would be affected even if the United States produced all of its oil domestically (unless petroleum exports were prohibited). The outcome is the same because the price of oil is set in global markets.

Meeting all U.S. energy needs from domestic sources would require significant changes to the U.S. economy, including changes in the types of transportation fuels used by Americans. The costs of these changes would probably exceed the costs resulting from periodic unexpected increases in the global price of oil. This is suggested by the fact that prior oil market disruptions did not lead to such structural changes in the U.S. economy. Moreover, oil markets have undergone tremendous changes since the 1970s that likely reduce the risks to the U.S. economy from a disruption in crude oil production and imports.

U.S. Energy Sources

Most energy consumed in the United States is produced in North America. In 2002, the main energy sources were petroleum (39 percent), natural gas (24 percent), coal (23 percent), and nuclear power (8 percent). In 2002, roughly 80 percent of U.S. energy needs were met by North American sources, including 59 percent of crude oil, 99 percent of natural gas, 100 percent of coal, and roughly 45 percent of uranium for nuclear power generation. Petroleum is the main energy source that the United States imports in significant amounts from outside North America. Hence, discussions of energy security focus on imports of crude oil. In the future, analysts expect the United States to import more natural gas, but there are many potential suppliers.

The United States also imports a large share of uranium from outside North America, but there are sufficient North American reserves of uranium that

could be used if less-expensive foreign sources were not available. Furthermore, uranium fuel represents a relatively small portion of the cost of nuclear electricity generation. Also, most uranium is produced in stable parts of the world, with Canada and Australia producing about half of the world's total.

Changes in the Oil Market

A disruption in crude oil production in an area that does not supply the United States would still affect the United States by raising oil prices in the worldwide market. However, the power of the Organization of the Petroleum Exporting Countries (OPEC), or of any one country, to affect world oil prices is less today than it was in the past. OPEC's influence on the market has fallen with the decline of its market share from 55 percent in 1973 to 39 percent in 2002. Other evidence of the diversification of sources of crude oil is that in 1973, the top eight producing countries produced 75 percent of the world's oil, while in 2002 the top eight producing countries produced only 54 percent. Access to a greater number of sources of oil reduces the impact of a disruption in any one region on the world oil price. In addition, the increased sophistication of financial markets for oil and related products has made it easier to hedge oil price risks. With financial instruments such as futures contracts, firms are better able to avoid having potential disruptions in the crude oil market lead to substantial immediate cost increases from their energy inputs.

Another significant difference between today and the 1970s is that the United States no longer has price controls on gasoline and oil. During the oil shocks of the 1970s, Federal government mandates kept consumer prices artificially low and dampened the amount of gasoline conservation that otherwise would have occurred in response to increased prices. As a result, people wanted to consume more gasoline than suppliers were willing to supply at the artificially low price leading to shortages in the United States.

When prices are not regulated, large swings in oil prices do not disrupt the economy nearly as much. For example, between June 15, 1998, and November 27, 2000, the price of West Texas Intermediate (WTI) crude oil more than tripled from $11.69 to $36.24 per barrel without throwing the economy into disarray. These price increases did not cause major economic disruptions for two main reasons. First, energy consumption per 1996 dollar of real GDP has dropped 43 percent, from 18,360 British thermal units (BTU—a measure of the energy content in fuels) per dollar in 1973 to 10,450 BTU per dollar in 2001. Second, market signals have worked to increase the flexibility of U.S. energy markets, allowing them to adjust and adapt to market changes. This is why market forces work better to allocate goods than command-and-control measures such as price controls.

Another change from the 1970s has been the expansion of the strategic reserve of crude oil that can be used during severe disruptions to the oil market. Created in 1975, the Strategic Petroleum Reserve held 634.7 million barrels as of December 2003—enough oil to replace U.S. crude oil imports from the Persian Gulf for approximately 287 days. While maintaining the Strategic Petroleum Reserve entails storage and inventory costs, holding reserves to increase energy security is less likely to distort the market than other measures, such as attempting to replace U.S. oil imports with more expensive sources of energy.

Trade in Oil and Price Stability

In considering whether it is worth taking steps to decrease U.S. reliance on petroleum imports from outside North America, it is useful to compare the movement of oil prices with the prices of other commodities in which the United States is self-sufficient. It turns out that having a supply of a commodity in the United States or North America is not an assurance of stable prices. Numerous factors affect both the supply and the demand of goods so that commodities such as natural gas, wholesale electricity, and many agricultural goods also exhibit price volatility even when supplied wholly from North American sources.

Relying on imported oil reduces the United States' overall expenditures on energy. Without crude oil imports, the cost of gasoline and other petroleum products (or alternative transportation fuels) would be higher. Therefore, the United States would have to devote a greater portion of its resources to paying for the costs of energy, especially for transportation, than is the case today. Without petroleum imports, it would be necessary to use significantly less gasoline and more transportation fuels made from corn, soybeans, or other agricultural products, or liquid fuels from coal, natural gas, oil sands, or oil shale. Under current technologies, these substitutes all cost substantially more to produce than gasoline from crude oil.

The Evolution of Energy Markets

Energy sources have changed as society's needs have evolved over time. Wood was replaced by coal, which was replaced by petroleum. Eventually, the energy market may evolve to include substantial energy production from new sources, such as renewable energy, hydrogen, or nuclear fusion. Government policy can help move this evolutionary process forward by encouraging research in new energy technologies. However, forcing the transition to new technologies before the market signals that old technologies should begin to be phased out could involve tremendous costs to society.

Market signals have already altered U.S. energy consumption. In response to higher crude oil prices, U.S. crude consumption fell by 21 percent between 1978 and 1983 even as real GDP grew by 7.8 percent. Demand shifted towards coal, which experienced the smallest price increase of any major fuel, and away from oil and natural gas, which experienced the greatest increases. Even with the increased consumption of coal, total U.S. energy consumption declined 1.8 percent annually between 1978 and 1983. This decrease occurred despite the longer-term upward trend of energy consumption, which averaged 1.1 percent annually between 1971 and 2001. Energy conservation programs and other nonmarket forces may have been responsible for some of the reduced demand for energy. However, at least 80 percent (and probably more) of the demand reduction can be attributed to higher prices and overall changes in the economy.

Market signals have also triggered a great deal of innovation to lower the cost of finding and extracting oil. For example, three-dimensional seismic technologies have lowered the cost of finding oil, and directional drilling has lowered the cost of extracting oil so that reserves that were not viable in the past can be extracted profitably today. Similarly, technological advances have lowered the cost of extracting oil from oil sands so that production from oil sands is competitive at today's oil prices. As a result, at least one industry publication has classified a portion of Canada's large oil sand deposits as proved oil reserves; estimates of Canada's proved oil reserves are now second only to those of Saudi Arabia.

The technology exists to convert large North American reserves of oil sands, oil shale, natural gas, coal, wood, and agricultural products into liquid fuels such as gasoline, diesel, methanol, and ethanol. Some of these processes are now prohibitively expensive, but these fuels could compete with fuels produced from crude oil if oil prices increased or if research and development lowered their production costs. Chart 8-3 illustrates the range of estimated costs of producing synthetic fuels that could compete with oil in the market for liquid fuels. For example, at a price for oil of $20 a barrel, liquid fuels from oil sands and natural gas may be able to cover production costs, while oil shale, coal, ethanol, and biodiesel would not be viable sources. Higher prices could eventually make these alternatives commercially viable. Note that the extraction process for some of these fuels may have adverse environmental consequences that could limit their use and that some of these processes yield low-sulfur fuels that may burn more cleanly than fuels produced from crude oil. The chart does not consider either the costs of the externalities or the benefits of the cleaner fuels.

There is a role for government in subsidizing research and development into new energy sources. For example, hydrogen shows strong potential as a possible future fuel, though many technological hurdles must be overcome

Chart 8-3 Production Costs and Reserves of Alternative Transportation Fuel Sources
Synthetic fuels have the potential to become commercially viable if the price of oil increases sufficiently.
Energy reserves not traditionally used as a source of liquid fuels would become available at this point.

Trillion barrels of oil equivalent

Range of estimated costs	Estimated reserves

Oil sand/heavy oil

Gas to liquids

Oil shale

Coal to liquids

Ethanol

Biodiesel

Dollars per barrel of oil equivalent (2002 dollars): 0 20 40 60 80

Reserves: Crude oil, Oil sand/heavy oil, Gas to liquids, Oil shale, Coal to liquids
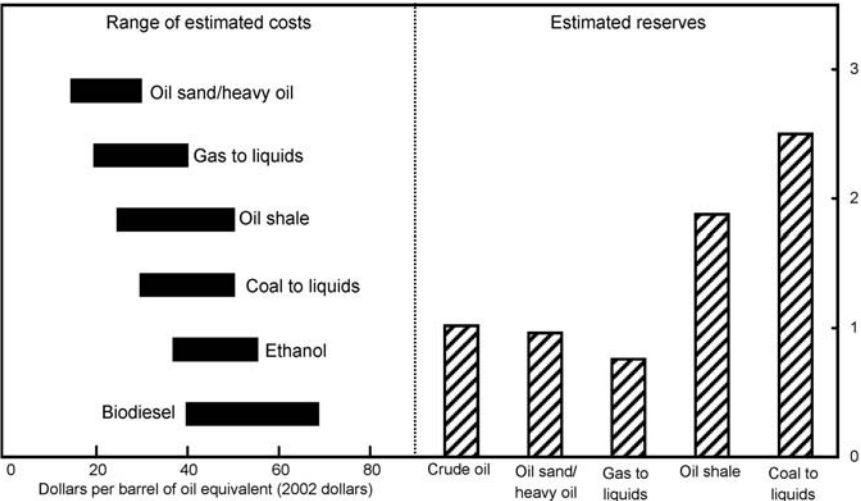
Note: Ethanol and biodiesel reserves are assumed to be infinite. Crude oil reserves are included for comparison.
Sources: Department of Agriculture, Department of Energy, and Council of Economic Advisers.

before it becomes practical for everyday use. Even if hydrogen became a feasible energy source, there would be still more problems to be resolved before the technology became economically competitive. Government subsidies for research and development may aid the private market in developing technology to produce, transport, and use hydrogen economically as a fuel. However, market forces should decide when commercial adoption of hydrogen as an energy source will be competitive.

Policy makers should avoid forcing commercialization of new energy sources before market signals indicate that a shift is required. One potential problem with forcing this process is that technological breakthroughs may lead to alternatives that are not seriously considered today. Premature adoption of new technologies would raise energy costs before the need arises, causing society as a whole to spend more on energy than needed, a misallocation of resources that would hurt the U.S. economy. For example, forcing adoption of energy sources other than oil to gain complete energy independence would be prohibitively expensive; it would require tremendous reductions in the use of energy derived from crude oil through the use of alternative energy sources that are far from competitive.

Conclusion

Regulations can improve the performance of energy markets by addressing market failures such as externalities and market power. However, it is essential to design regulations to address these potential market failures without reducing the benefits from markets. An added complication occurs when the goals of local and Federal regulators conflict. Regulators should adjust the rules as markets evolve and ensure that the regulations' goals are achieved. Finally, regulators should be careful not to adopt regulations that cause more harm than the potential market failure.

Protecting the Environment

Economic growth and environmental improvements go hand-in-hand. Economic growth can lead to increased demand for environmental improvements and can provide the resources that make it possible to address environmental problems. Some policies aimed at promoting environmental improvements can entail substantial economic costs. Misguided policies might actually achieve less environmental progress than alternative policies for the same economic cost. It is therefore important to weigh the direct benefits of environmental regulations against their economic costs.

While the free-market system typically promotes efficiency and thus enhances economic growth, the absence of property rights for environmental "goods" such as clean air and water can lead to negative externalities that reduce societal well-being. This can be addressed by establishing and enforcing property rights that will lead the affected parties to negotiate mutually-beneficial outcomes in a market setting. If such negotiations are expensive, however, the government can design regulations that consider both the benefits of reducing the environmental externality as well as the costs the regulations impose on society. Regulations should be designed to achieve environmental goals at the lowest cost possible, thus helping to achieve environmental protection and continued economic growth.

The key points in this chapter are:
- Establishing and enforcing property rights for the environment can address environmentally-related market failures. Any needed regulations should consider both the benefits and the costs.
- Environmental risks should be evaluated using sound scientific methods to avoid possible distortions of regulatory priorities.
- Market-based regulations, such as the cap-and-trade programs promoted by the Administration to reduce common air pollutants, can achieve environmental goals at lower cost than inflexible command-and-control regulations.

The Free Market and the Environment

In a free-market system, only trades that benefit both parties will take place. Market prices coordinate the activities of buyers and sellers and convey information about the strength of consumer demand for a good, as well as how costly it is to supply. In the context of the environment, a market failure may

occur if a voluntary transaction between parties imposes involuntary costs on a third party. These involuntary third-party costs are known as *negative externalities* (or *spillovers*), and their existence in a free market can lead to inefficient outcomes; that is, outcomes that fail to maximize the net benefits to society. For example, a plant might produce and sell a good to a consumer to both their advantage, but the production process may result in emissions of air pollutants that negatively affect others not involved in the transaction. The root of the market failure is that there are no clear property rights for the surrounding air. The interests of the third party—the people affected by the plant's emissions—are not represented in the market transaction.

If those affected by the plant's emissions had a right to demand compensation for the costs imposed on them by the pollution, then the firm would take these costs into account when making its production decisions. The plant would produce only up to the point where the benefit of another unit of production equals the additional cost of producing the good plus the cost to the people negatively affected by the pollution. Any additional emissions due to producing more goods would require compensation that is greater than the monetary gain the plant gets from selling the additional goods. Likewise, if the property right belonged to the plant, the people negatively affected by the emissions could compensate the plant for reduced emissions. Either way, all three parties (consumers, the firm, and those affected by the emissions) would transact voluntarily to everyone's benefit, resulting in an efficient outcome. If the government were to assign and enforce the property right, and if it were costless for parties to collectively agree on compensation, then an efficient use of resources would result from private bargaining, regardless of which party was assigned the property right. This insight is known as the *Coase theorem*.

The Role of Government in Regulating the Environment

The existence of property rights does not always guarantee an efficient outcome. If there are many sources of pollution or there are many parties affected by the emissions, then it might be difficult for the parties collectively to agree on the compensation, and an efficient outcome might therefore not be achieved. This presents an economic justification for government involvement and regulation. Government regulation might also be justified in order to address distributional concerns associated with environmental problems.

Regulations that address negative externalities can therefore improve societal welfare. To improve the environment while still promoting economic growth, sound policies must consider both the benefits and the costs of regulations. Economic growth itself can contribute to environmental improvements

(Box 9-1). As the economy grows, the demand for environmental improvements increases and the greater wealth provides more resources to better address environmental concerns. It is therefore important to weigh the direct environmental benefits of regulations against their economic costs.

Box 9-1: Economic Growth Can Improve the Environment

Much research has shown that economic growth contributes to environmental gains. In the early stages of economic development, environmental degradation may occur because nations place higher priority on basic needs such as food and shelter. As wealth increases, however, so does demand for a cleaner environment, and greater wealth provides more resources to better address these environmental concerns. After a certain level of national income is attained, the balance shifts and environmental degradation is arrested and then reversed. For several decades in the United States, many environmental indicators have been improving as the economy has also grown.

From 1975 to 2002, concentrations of five of the six common air pollutants (the pollutants for which there are reliable data) decreased by an average of 60 percent (Chart 9-1), as real gross domestic product (GDP) increased by about 130 percent, energy consumption increased by 35 percent, and the population increased by 34 percent. While the Nation's air quality has improved substantially since passage of the Clean Air Act of 1970, air quality was improving prior to 1970, perhaps due to market-induced technological advancements (such as improvements in energy efficiency) that accompany economic growth. The limited air-quality monitoring data available before 1970 indicate that average annual concentrations of particulate matter in urban air dropped 16 percent from 1957 to 1970 and these total suspended particulates (liquid or solid particles in the air) across the country fell by about six percent from 1958 to 1970 (Chart 9-2).

As the Nation's productive output has increased and environmental quality has improved, so too has the health and well-being of Americans. In the last century, life expectancy at birth increased from 48 to 80 years for women and from 46 to 74 years for men. Infant mortality dropped to the lowest level ever recorded in the United States. The death rates for heart disease, cancer, and stroke are also decreasing. This well-documented correlation between wealth and health extends across time and nations. More-developed countries have higher life expectancy, and globally, life expectancy has increased as per capita wealth has increased.

Chart 9-1 National Concentrations of Air Pollutants

Concentrations of five major air pollutants have been declining since 1975.

Index, 1975 = 100
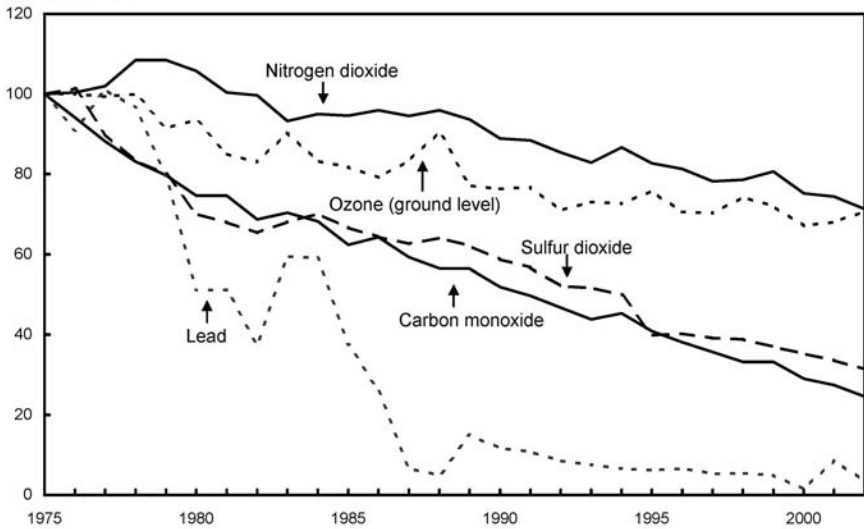
Note: The number of air monitors varies across pollutants and time.
Source: Environmental Protection Agency.

Chart 9-2 Particulate Matter Concentrations

While the data for monitoring airborne particulate matter have not been uniform, limited data indicate that particulate matter began declining steadily prior to passage of the Clean Air Act in 1970.

Micrograms per cubic meter
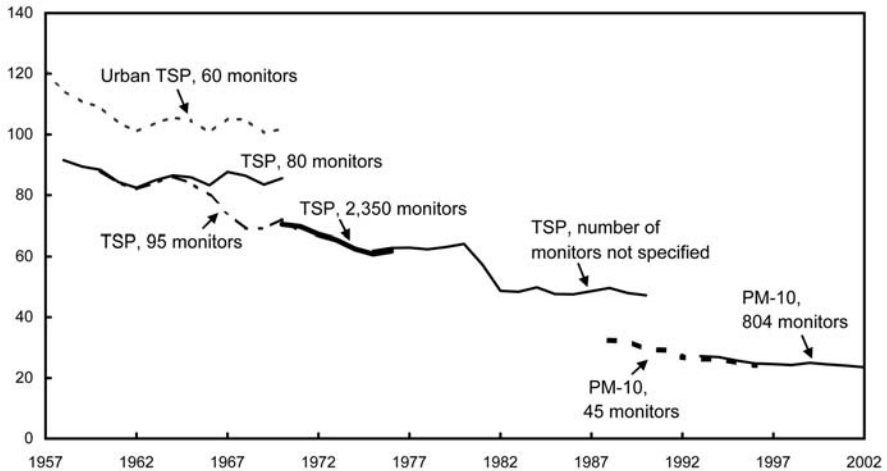
Note: TSP refers to total suspended particulates and PM-10 refers to particulates with diameter of 10 micrometers or less. The figures for monitors are the number that collected data for a particular series.
Sources: Environmental Protection Agency and Council on Environmental Quality.

Misplaced Reasons for Government Intervention

In making environmental policies, it must be recognized that government measures themselves might create further inefficiencies. When it is difficult to determine the extent of an environmental externality, an attempt to rectify it might end up making matters worse by imposing unintended costs on third parties without achieving an efficient outcome.

This inefficiency can arise even from well-intentioned environmental regulations. Two fallacious arguments are frequently used to justify inefficient regulations. One such misplaced rationale is that regulations improve the economy and spur job growth. The reasoning goes as follows: environmental regulations lead firms to install pollution-control technologies, which they must purchase from other firms. These technologies are built, delivered, installed, and operated by workers who otherwise would not be doing this work. Similarly, the regulations may promote environmentally-friendly industries that hire people who would not be hired otherwise. For these reasons, the regulations are said to "spur" the economy and job growth. By this reasoning, throwing a rock through a window also improves the economy, because it necessitates the hiring of someone to repair the window. What this ignores is that the resources spent to comply with an unnecessary or inefficient regulation are diverted from other uses. The money and people involved could have been used instead to produce more goods for consumers or to build new factories or machinery. The jobs associated with complying with environmental regulations are a cost of regulation, not a benefit.

Another misplaced view of environmental regulation is that the goal of regulations should be to eliminate or substantially reduce risks without considering costs. This approach is embodied in some well-intentioned laws. The 1970 Clean Air Act, for example, directs the Environmental Protection Agency (EPA) Administrator to set national ambient air quality standards (NAAQS) that achieve "an adequate margin of safety," and the Supreme Court has ruled that "the Clean Air Act…unambiguously bars cost considerations in the NAAQS setting process." Similarly, the stated goal of the Occupational Safety and Health Act of 1970 is "to assure so far as possible every working man and woman in the Nation safe and healthful working conditions," without considering the costs of doing so. While the goals of these laws are noble, they do not recognize the inevitable trade-offs involved. Not all environmental laws preclude cost considerations. For example, the Safe Drinking Water Act Amendments of 1996 explicitly acknowledge the importance of benefit-cost analysis when considering the appropriate level of regulation for contaminants in drinking water.

Regulations Impose Benefits and Costs

The failure to consider costs inhibits the goal of making regulations that maximize the difference between benefits and costs. Furthermore, the failure to consider costs can lead to a misallocation of resources, because a regulation that is made without considering costs might receive more resources than other regulations that warrant greater attention. While the benefits of many regulations include both health and non-health related benefits, many regulations primarily address fatality risks, and there is a wide range of cost per expected life saved across such regulations. For example, one survey of cost per life saved across regulations found that the regulation for childproof cigarette lighters costs approximately $100,000 per life saved (in 2003 dollars) whereas the formaldehyde regulation costs approximately $80 billion per life saved (in 2003 dollars). Shifting resources from regulations where the cost per expected life saved is high (for example, formaldehyde regulation) to regulations where it is low (for example, childproof lighter regulation) would result in more lives saved for the same cost to society. Many of the differences in cost per life saved occur because legislative mandates only sometimes allow agencies to consider costs when crafting regulations.

Stringent regulations may appear to be good for society because they save lives. However, because the Nation's ability to bear costs is limited, the wide range of costs per life saved across regulations implies that more lives could be saved at the same cost by shifting resources to the regulations with lower costs per expected life saved. One study found that society could save twice as many lives with the same budget if it designed regulations in a way that maximized lives saved. Some of the more costly health-based regulations might actually lead to a net increase in fatality risk because their high costs diminish the resources available for improving other health and environmental outcomes.

Using Science to Help Set Regulatory Priorities

Sound regulatory policy must be based on scientific assessments of environmental and health risks. Scientific assessments involve a careful examination of the risks involved and of the expected health outcomes for the people exposed to the risk at hand. This allows for an unbiased evaluation of environmental and health threats in which to target regulatory actions. Unfortunately, regulatory risk assessments at times overestimate some threats, or overemphasize risks to "hypothetical" (rather than real) people. These practices can lead to a distortion of regulatory priorities.

Overestimating the Risks: The Problem with "Cascading Conservatism"

In a well-intentioned attempt to be prudent, regulatory agencies sometimes rely on scientific assessments of environmental and health risks based on assumptions that overstate actual risks. When estimating chemical toxicity, for example, risk assessors have at times relied on high-end default assumptions that are likely to overestimate the actual risk of a chemical. Toxicity testing is evolving to use information that permits assessors to move away from assumptions that lead to overstated risks. When more data are available, regulatory risk assessors do not need to rely on high-end default assumptions and can instead attempt to estimate more accurately the expected level of risk. Because the EPA's primary goal is public health protection, however, it still relies on high-end default assumptions when there is uncertainty about scientific data.

Similarly, regulatory agencies sometimes use high-end estimates of the likelihood of people being exposed to a certain risk. These exposure estimates are then combined with toxicity estimates that are themselves likely to overstate risk. The multiplicative impact of combining several high-end component estimates is known as *cascading conservatism*. This practice can lead to risk estimates that greatly overstate the threat of environmental problems and thus overstate the benefits of regulating those risks. One study found that in a sample of hazardous waste sites, over 40 percent of the sites requiring cleanup under the Superfund program would shift into the discretionary cleanup range if not for the overestimation of risks resulting from cascading conservatism.

Such high-end risk estimates can lead to several types of problems. First, the practice overstates the risk of all environmental health problems relative to other types of hazards. This overstatement can cause too many resources to be allocated to addressing low-priority concerns. An example of such a distortion is the commonly-held view that synthetic chemical pollutants such as insecticides are a leading contributor to cancer. In reality, the evidence suggests that such chemicals account for a low percentage of human cancers. The main contributors to human cancer appear to be smoking and poor diet—each of which accounts for about one-third of cancers. The result is that regulatory efforts are directed at addressing the risks of synthetic chemicals that may well pose lower risks of causing cancer than many common natural chemicals.

A second problem with the high-end risk estimates caused by cascading conservatism is that they can distort the allocation of resources among different environmental health concerns. If each uncertain component that goes into a risk assessment overstates the risk, then the multiplicative impact of cascading conservatism will result in higher risk estimates for threats that have more uncertain components. For example, if there are two equally

effective pesticides, with one posing a higher threat to the population than the other, the safer pesticide might be assessed as more of a threat if there are more uncertain components involved in its risk assessment. This assessment could result in the safer pesticide receiving stronger regulatory emphasis by the government. It is better to target regulatory dollars to the risks expected to be higher in a reasonable scenario or range of scenarios than to the risks that might be higher in a worst-case scenario.

Population-Weighted Risk Assessments

Regulatory efforts can also be distorted when risk assessments ignore the number of real people potentially exposed to an environmental risk. For example, an environmental hazard at one location might pose a greater risk to any person exposed to the hazard than an environmental hazard at a second location. However, if no one lives near the first location and many people live near the second location, the expected risk to society is higher at the second location.

The case of *United States* v. *Ottati & Goss* offers one example of such misplaced regulatory priorities. In this case, a company litigated for relief of an EPA-required cleanup that would have cost the company $9.3 million to remove small amounts of contaminants from a site that was already mostly decontaminated. The company had already spent $2.6 million to clean the site so that small children playing on the site could eat small amounts of dirt daily for 70 days each year for three and a half years without significant harm. The additional $9.3 million would be used to burn the soil, which would allow children to eat a small amount of dirt each day for 245 days per year without significant harm. However, there was little chance that children would ever be exposed to this site because it was located in a swamp. The courts ruled in favor of the private party and refused to enforce the proposed remediation goal.

Objective Versus Perceived Risk

Regulatory decisions should be based on scientific assessments of risks rather than perceived risks. This approach would help properly order priorities for regulatory decisions. Perceived risks often differ from expert assessments of risk because laypeople have difficulty assessing the frequency of low-probability events. Chart 9-3 compares survey respondents' perceived risks of dying from various hazards to the objectively measured risks of dying. In this chart, the dashed line represents where the perceived risk equals the actual risk; if all the points on the chart fell on this line, it would indicate that survey participants precisely estimated the risk of dying from various hazards. All points to the left of the dashed line represent hazards for which the

Chart 9-3 Relationship Between Actual and Perceived Risk of Dying
People tend to overestimate the risk of dying from low-fatality events and underestimate the risk of dying from high-fatality events.

Perceived frequency of death per 100 million people, log scale
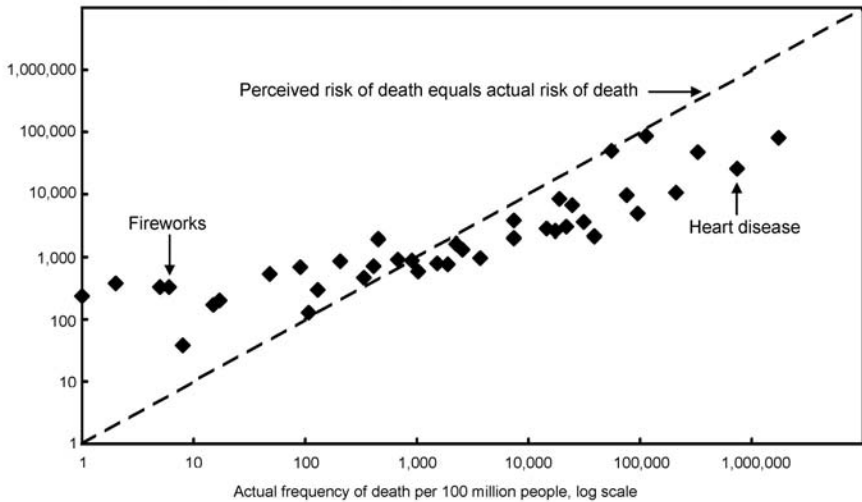
Source: Sarah Lichtenstein, Paul Slovic, Baruch Fischhoff, Mark Layman, and Barbara Combs, "Judged Frequency of Lethal Events," *Journal of Experimental Psychology: Human Learning and Memory*, November 1978.

perceived risk of dying is higher than the actual risk, and all points to the right of the line indicate hazards for which people thought the risk of dying is lower than it actually is. The chart suggests that it is common to overestimate fatalities associated with low-probability events and to underestimate fatalities associated with high-probability events. These systematic misperceptions may lead to misplaced pressures to overregulate small environmental risks at the expense of addressing larger ones.

Achieving Goals Through Cost-Effective Regulations

As discussed in Chapter 7, *Government Regulation in a Free-Market Society*, when the assignment of property rights is insufficient to achieve an efficient outcome, government intervention may help achieve efficiency. Chapter 7 discusses government actions that can, in principle, achieve an efficient outcome by incorporating the costs of externalities into the market's price mechanism. It is important that any regulatory mechanism that addresses externalities do so in the least costly (that is, the most cost-effective) way so that society's scarce resources are not wasted. This section focuses on how to achieve air-quality goals cost effectively, but many of the

lessons can be applied toward achieving other environmental goals, such as clean water protection and energy-efficiency standards.

Command-and-Control Regulations

Air-quality command-and-control regulations prescribe specific technologies that individual firms must use to control emissions, or they set specific emission rates for individual firms. The United States currently has many such environmental regulations. These regulations are inherently inflexible and are ill-suited to achieving emissions reductions in the least costly manner. While some command-and-control air-quality regulations may be just slightly more costly than cost-effective regulations, studies show that others are up to 22 times more expensive than the most cost-effective set of controls.

The reason command-and-control regulations are more expensive is straightforward: suppose the regulatory goal is to halve the emissions emanating from two firms. A command-and-control regulation might require each firm to cut its emissions by half. However, if it is less costly for one firm to reduce emissions, then—so long as the health effects of the emissions depend only on the total from the two emission sources—shifting the burden to the firm with lower abatement costs would result in the same environmental improvement at a lower cost. In general, the greater the differences across firms in their emissions before the regulation, and the greater the differences across firms in the rate at which each firm's costs rise with additional reductions, then the more costly a command-and-control approach is compared to more flexible approaches. Cost-effective emissions reduction is achieved when the cost of reducing an additional unit of emissions (the *marginal abatement cost*) is equal across all firms.

An example of an inflexible command-and-control regulation is the mechanism by which the Clean Air Act Amendments of 1990 address hazardous air pollutants (HAPs). The Act specifies that the emissions reduction standards for categories of existing HAP polluters must be set at "the average emission limitation achieved by the best performing 12 percent of the existing sources." While some flexibility is allowed in establishing the emission limitations, the command-and-control standard for regulating HAPs has frequently been interpreted in a way that ignores the differential costs of reducing emissions across existing sources within a category. This likely results in higher costs than would a more flexible regulation.

Command-and-control regulations also fail to provide market incentives for firms to explore less expensive means of reducing emissions. More flexible, incentive-based regulations would provide signals to the market of the increased demand for emissions reductions. With proper incentives in place, markets can respond to such an increase in demand with technological innovation and efficient reallocation of their scarce resources to achieve the goal.

Command-and-control regulations can also unintentionally lead to outcomes that are contrary to their environmental goals. An example of this is the New Source Review component of the 1977 Clean Air Act Amendments. This legislation required a strict control technology for most new industrial facilities and for facilities that undertook significant modifications, but it exempted existing facilities that did not make major modifications from the same standards. It was thought at the time to be more efficient to add new pollution control technology when plants were upgrading or when building new plants. This situation is known as *new source bias* because it provides an incentive for existing sources of emissions to continue their business operations for longer than would have been the case under normal market conditions without the regulation. It also provides an incentive for existing plants to forgo modifications.

New pollution-causing production sources tend to be cleaner than old ones even in the absence of regulations, so extending the business operations of older plants without making modifications could result in higher emissions. Applying different regulations for "routine" versus "major" modifications also leads to ambiguity, litigation delays, and uncertainty in business planning, all of which can harm the economy and may impede environmental improvements. The Administration recently addressed this problem by establishing clear rules that remove disincentives for facilities to modify and undertake routine maintenance, repair, and replacement activities that could improve the safety, reliability, and efficiency of the plants.

Market-Based Price Regulations: Emission Fees

Environmental regulations that provide firms with market-based incentives for emissions reduction avoid the complications of command-and-control regulations and achieve the same goals at lower costs. In particular, emission fees and cap-and-trade programs are usually less expensive than command-and-control approaches at achieving regulatory goals. An emission fee involves a charge to polluting sources for each unit of pollution emitted. Because each successive unit of emissions reduction typically involves increased costs, each source will reduce emissions until it would cost more to reduce the next unit of emissions than it would to pay the emissions fee. This results in equal marginal abatement costs across all affected firms.

With an emission fee, the total level of emissions reduction will depend on the per unit fee: a higher rate will achieve more emissions reduction. The emission fee also provides incentives to reduce emissions, because the better a firm is at reducing emissions, the lower the total fee the firm must pay. This sends a market signal that pollution has a price (equal to the emission fee), and any innovative means of reducing emissions will save firms from paying the fee. This market signal is likely more adept than the government at

spurring technological innovation, adapting to changes in the economy, and shifting resources to reflect the increased demand for emissions reduction.

Market-Based Quantity Regulations: Cap-and-Trade

The main problem with an emission fee is that it is difficult to know beforehand what fee level will achieve the desired amount of pollution reduction. A cap-and-trade regulation addresses this issue and provides market incentives to reduce emissions in a cost-effective way. Such regulations "cap" the amount of allowable emissions and require that a firm own a permit for each unit of pollution emitted in a given period (for example, a year). This permit effectively establishes a legal property right for the air affected by the pollution, so that any emissions must be paid for by the firm. The government allocates the pollution permits to the emission sources and then allows the sources to buy and sell permits from each other.

Under a cap-and-trade system, a source with a high cost of reducing an additional unit of emissions would be willing to purchase a permit from a source with a lower marginal abatement cost. With a well-functioning market for the permits, sources will trade permits until the price for the permits equals the marginal abatement cost. As with the emission fee, the marginal abatement costs will be equal across sources, leading to a cost-effective result. The cap-and-trade system also provides an incentive to reduce emissions because each unit of emissions reduction saves the source the price of another permit. This regulation sends a market signal that there is a price for emissions and any innovative means of reducing emissions will save firms from paying the price. The cap-and-trade system therefore achieves the target level of pollution reduction at the lowest cost.

One consideration for a cap-and-trade system is how to allocate the permits initially. A cap-and-trade system that allocates the permits based on historic emissions or other firm characteristics, known as *grandfathering*, in essence gives away a valuable asset—the permits. A grandfathering system could establish a barrier to entry for new firms because any new entrant would have to purchase permits from existing firms.

One way to avoid these problems is to auction the permits at some regular interval to the highest bidders. Firms with higher marginal abatement costs would bid more for permits than those that can achieve less-costly emissions reductions. While auctioning the permits would result in lower profits for the regulated firms (compared to giving away the permits), it would not affect the firms' output decisions. Grandfathering versus auctioning the permits is primarily a question of distribution, not efficiency—it is a question of whether a public asset should be given to firms for free or sold as a means of generating public revenues.

A notable example of a cap-and-trade system is the sulfur dioxide (SO_2) trading program created under Title IV of the Clean Air Act Amendments of 1990. The program set a goal of reducing emissions by 10 million tons from the 1980 level by 2010. This was to be accomplished in two phases. The first phase, which began in 1995, initially capped the SO_2 emissions at 263 individual units which were owned by 110 electric utility power plants in 21 eastern and midwestern states. These plants, which were primarily coal-fired, emitted the greatest amounts of pollution among power plants in these regions. From 1995 to 2000, an additional 182 units were allowed into the program. The second phase, which began in 2000, further decreased the annual emissions of SO_2 and required all large fossil fuel-fired power plants in the contiguous 48 states and the District of Columbia to hold permits to cover their emissions.

In both phases, power plants could purchase permits from other power plants in order to meet their emissions coverage. The program also allowed plants to carry over (or *bank*) unused permits to use in later years, which gives firms even greater flexibility in achieving long-term pollution reduction. In contrast to a command-and-control system, this cap-and-trade system allows plants that find it costly to reduce their SO_2 emissions to purchase credits from plants that can reduce SO_2 at lower cost.

Evidence indicates that such cost-saving trades did indeed take place as firms took advantage of the system's inherent flexibility (Chart 9-4). Each bar in the following chart represents the emissions rate each plant achieved after trading permits in 1997. The superimposed line in the figure shows the level of emissions each plant would have had to achieve in the absence of trading. Bars below the line indicate plants that reduced their emissions by more than the required amount and sold their excess permits or banked them. Bars above the line indicate plants that purchased permits or used previously banked permits to avoid costly abatement. The figure shows that almost every plant took advantage of the flexibility of the system, suggesting that plant-level costs of reducing SO_2 emissions vary greatly.

The trading program has achieved its pollution-reduction goals at great cost savings. By the end of the first phase, emission reductions were almost 30 percent below the required level. The flexibility of this approach has been estimated to provide cost savings of approximately $0.9 billion to $1.8 billion a year compared to costs under a command-and-control regulatory alternative; other tradable-permit markets have had significant cost savings as well (Table 9-1).

Chart 9-4 Unit-Level Sulfur Dioxide Emissions Trading in 1997

Variation in actual plant-level emissions for units in the Acid Rain Program indicates that firms took advantage of the flexibility and cost-savings inherent in the cap-and-trade system.

Emissions rate (pounds of sulfur dioxide per mmbtu)
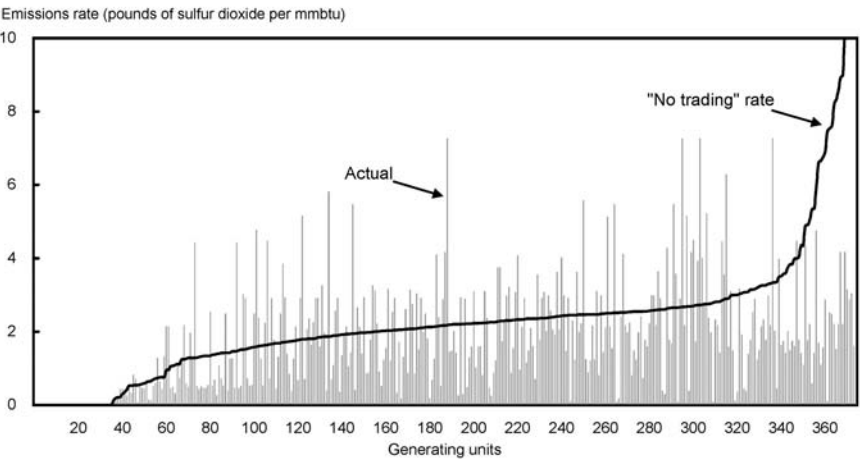
Note: Corresponding graphs for other years of the program show similar patterns.
Sources: Richard Schmalensee, Paul L. Joskow, A. Denny Ellerman, Juan Pablo Montero, and Elizabeth M. Bailey, "An Interim Evaluation of Sulfur Dioxide Emissions Trading," *Journal of Economic Perspectives*, Summer 1998. Update from personal communication between A. Denny Ellerman, Massachusetts Institute of Technology, and the Council of Economic Advisers.

TABLE 9-1.— *Cost Savings of Tradable-Permit Systems*

Program	Traded commodity	Years of operation	Cost savings (2003 dollars[1])
Emissions trading program	Criteria air pollutants	1974–present	Total, $1–$12 billion
Lead phasedown	Rights for lead in gasoline	1985–1987	Total, $400 million
Acid rain reduction	SO_2 emission reduction credits	1995–present	Annual, $0.9–$1.8 billion

[1] Base year for values for emissions trading program not specified.

Sources: Robert W. Hahn, "Economic Prescriptions for Environmental Problems: How the Patient Followed the Doctor's Orders," *Journal of Economic Perspectives*, Spring 2000; Curtis Carlson, Dallas Burtraw, Maureen Cropper, and Karen L. Palmer, "Sulfur Dioxide Control by Electric Utilities: What Are the Gains from Trade?" *Journal of Political Economy*, December 2000; and Environmental Protection Agency.

Emission Fees Versus Cap-and-Trade

As mentioned previously, one problem with emission fees is that it is difficult to know beforehand at what level to set the fee to achieve the desired pollution reduction. This might require periodic adjustments of the fee level, and such adjustments would introduce uncertainty that could interfere with firms' planning decisions. The emissions fee does, however, allow the government to set with certainty the marginal cost of emissions reduction. For each emission fee there is a corresponding allocation of permits that

would achieve the same results; however, it is difficult to know beforehand what the market price for permits will be once trading actually takes place.

One way to reconcile these issues is to offer a cap-and-trade system with a *safety valve*. The safety valve sets a maximum price for a permit, which guarantees that the price of reducing emissions does not exceed the expected benefits. The regulatory agency issues and sells extra permits on request from any firm at this fixed safety valve price, thus guaranteeing that the market permit price does not exceed this level. A cap-and-trade program with a safety valve achieves the target level of emission reductions in a cost-effective manner, while protecting the regulated firms against unexpected short-term price increases in emissions reduction.

The President's Cap-and-Trade Program

An example of a well-designed incentive-based regulatory approach is the President's Clear Skies proposal for reducing emissions of sulfur dioxide, nitrogen oxides, and mercury from electric utility generators by approximately 70 percent by 2018. Clear Skies would cost-effectively reduce emissions by establishing a cap-and-trade system for each of the three pollutants. The EPA has estimated the benefits of the Clear Skies Act at $113 billion annually by 2020, compared with $6 billion in projected annual costs. These include $110 billion in annual health benefits (including the prevention of 14,100 premature deaths and 30,000 hospitalizations and emergency room visits) and $3 billion in annual benefits from increased visibility at national parks. Under the existing Clean Air Act, the EPA issues national air-quality standards for certain pollutants, including particulate matter and ozone. The EPA projects that compared with existing programs, the Clear Skies Act would lead 35 additional eastern U.S. counties to meet the particulate matter standard by 2020, leaving only eight counties not meeting the standard. The EPA expects that the remaining counties not meeting the standards would move closer to achieving them due to the Clear Skies Act.

To mitigate the effects of market shocks that potentially affect the costs of emissions reduction, Clear Skies would establish a safety valve price for permits of each pollutant. It would also provide regulatory certainty by achieving the reductions of all three pollutants in two phases. Firms would therefore plan their reductions of the three pollutants together and over the long term. Indeed, because the Clear Skies plan allows the banking of permits for future use, it provides an incentive for firms to achieve reductions quickly. Additionally, Clear Skies would provide revenue for the government because it phases in an auction system for the permits.

Clear Skies demonstrates the lessons learned from past regulatory experiences: instead of imposing an inflexible, command-and-control regulation to achieve

emissions reduction, it offers a market-based, cost-effective, cap-and-trade program to achieve large reductions in emissions from electric utility generators.

Conclusion

Economic growth and environmental improvements are at times incorrectly seen as competing aims. Increased economic production can indeed lead to greater environmental degradation. However, an increase in economic resources provides more options (most notably, technological advancements) for addressing environmental problems. Moreover, a growing economy can also lead to increased demand for environmental improvements. It is therefore important to weigh the direct environmental benefits of a regulation against its economic costs. The goal should be to maximize the net benefits to society, while also giving due consideration to distributional issues. Maximizing net benefits is best achieved in a free-market setting unless there are spillover costs to third parties.

Spillover costs are best addressed by establishing property rights that will lead the affected parties to negotiate a mutually-beneficial outcome. If the costs of such negotiations are prohibitive, however, government should respond carefully and always keep in mind the possible government spillover costs. To make effective regulations, the government must first assess the environmental problems using sound, unbiased estimates of the hazards and then craft incentive-based regulations to address them. Such regulations can address the spillover costs of environmental problems at lower costs to society than the traditional command-and-control regulatory methods. These principles, and the lessons learned from our past regulatory experiences, as described throughout this chapter, should guide our future regulatory endeavors to achieve environmental improvements coupled with economic growth and efficiency.

Health Care and Insurance

The breadth and pace of innovation and change in the provision of health care in the United States over the past few decades have been no less than astounding. Technological progress in the form of new medical knowledge, medicines, treatments, and medical devices has allowed Americans and people worldwide to live longer, healthier lives.

As new treatment options become available, it is not surprising that the United States and other major industrialized countries continue to shift more resources to health care. Research suggests that between 50 and 75 percent of the growth rate in health expenditures in the United States is attributable to technological progress in health care goods and services. However, the increase in resources devoted to health care has led to concern about its affordability, both for families worried about tight budgets and for the Nation as a whole. A strong reliance on market mechanisms will ensure that incentives for innovation are maintained while providing high-quality care in the most cost-efficient manner. Americans should have more choices, more information, and more control over their health care decisions.

Health insurance plays a central role in the workings of the U.S. health care market. An understanding of the strengths and weaknesses of health insurance as a payment mechanism for health care is essential to the design of reforms that retain incentives for innovation while reining in unnecessary expenditures.

This chapter discusses the roles of innovation, insurance, and reform in the health care market. The key points in this chapter are:

- U.S. markets provide incentives to develop innovative health care products and services that benefit both Americans and the global community.
- Over reliance on health insurance as a payment mechanism leads to an inefficient use of resources in providing and utilizing health care.
- Reforms should provide consumers and health care providers with more flexibility and information.

The U.S. Health Care System as an Engine of Innovation

Innovation and new technology have changed the practice of medicine over the past few decades. Diagnostic tools such as magnetic resonance imaging and computed tomography scanning have made it possible for doctors to see otherwise invisible problems. Innovations such as balloon angioplasty treat conditions that previously required extensive surgery. Minimally invasive surgical techniques such as arthroscopy provide treatment options that lead to shorter hospital stays and faster recoveries. Restorative surgeries such as hip and knee replacements are now commonplace and provide patients with improved mobility and thus improved quality of life. New pharmaceuticals treat conditions that were previously intractable or help to avoid more costly surgeries and lengthy hospital stays. The list of advances is long and impressive.

The Value of Health Care Innovation

Innovation in health care goods and services, including advances in scientific knowledge that have changed many people's day-to-day behavior, has markedly improved the lives of Americans. Life expectancy at birth in the United States increased from 68.2 years in 1950 to 77.2 years in 2001. Medical advances have also increased the quality of life through innovations that improve mobility, sight, and hearing.

Some might argue that these advances are not unique to the United States and that Americans spend too much for health care relative to other countries. The United States expends a higher fraction of GDP on health care than does any other industrialized country. According to an international comparison released in 2003, the United States spent 13.9 percent of GDP on health care in 2001, while the average among industrialized countries was 8.4 percent of GDP. Measures of health outcomes such as longevity and infant mortality, however, are not markedly different in the United States than in other advanced economies that spend substantially less on health care.

The argument that the U.S. health care system is overly costly relative to other countries implicitly assumes that if two countries spend different amounts for health care and get the same health outcomes, then the higher-spending country must be inefficient and wasteful. This argument is not correct in the case of health care for two reasons that are related to the leading role of the United States as a source of research and innovation.

First, in general terms, while all countries can benefit from research and development expenditures made by a single country, only the health expenditures in the innovating country will include the costs of research and development. Health expenditures in non-innovating countries will exclude the research and development costs.

Second, free markets incorporate incentives for innovation that generate products, services, and knowledge that potentially benefit all countries. Markets naturally encourage and reward innovation. Unfettered by government price controls or access restrictions, innovative products, talented health care practitioners, and skilled health care professionals are rewarded in the marketplace. This leads to technological advances by encouraging talented people to participate in the health care industry and by increasing investment in new products and research. The financial rewards for innovation will be reflected in U.S. health expenditures through a combination of higher prices and wages, and higher usage than in other countries. Once a product or service is developed through the combination of talent and capital, however, it becomes available for use outside the United States. Countries in which government regulation has supplanted market forces will still have the opportunity to take advantage of U.S. innovation without having to pay as much for it.

As an illustration of how U.S. health expenditures reflect the incentives for innovation, consider products such as medical devices and pharmaceuticals. The patent system exists to encourage innovation for these types of products. The innovator's incentive in a patent-based system is the opportunity to hold a monopoly on a product for a limited period of time. Therefore, the innovator can temporarily charge a higher price and earn more profits than he would without patent protection. The higher consumer expenditures that can result from monopoly pricing will be reflected in health care expenditures.

Once the patent system has led to the development of a product, it is available for use throughout the world, not just in the United States. This leads to an opportunity for other countries with centralized health agencies to negotiate a price close to production costs, thereby paying lower prices than they would in a free market that fully respected patent rights. What this implies is that other countries can reap the benefits of U.S. innovations in health care goods and services but pay only a fraction of the costs. It follows that if the United States attempted to reduce health expenditures by adopting cost-control policies found in other countries, innovation would slow and both Americans and citizens of other countries would be affected.

U.S. Leadership in Health Care Technology

Several pieces of evidence point toward the preeminence of the United States in providing health care technology. First, since 1975, the Nobel Prize in medicine or physiology has been awarded to more Americans than to researchers in all other countries combined. Second, according to data collected through 1993, 15 of the 19 marketed "biotech" drugs used for nondiagnostic purposes were the product of U.S. companies alone. U.S. companies shared credit with companies from other countries for two more of the 19 drugs. As of 2002, eight of the world's ten top-selling drugs were produced by companies headquartered in the United States.

A third example of U.S. leadership is that many important medical innovations in the past 30 years arguably originated in the United States. This evidence is based on a survey designed to determine the relative importance of a variety of medical innovations developed over approximately the last 30 years. Starting with a review of the medical literature, researchers compiled a list of 30 major medical innovations and then surveyed over 300 leading general internists in the United States concerning the relative importance to their patients of the innovations. Based on the survey, researchers ranked the innovations in order of importance. The first and second columns of Table 10-1 reflect the results for the top ten innovations.

The table also includes countries of origin, a category that was not included in the original research. Assignment of country was based on the

TABLE 10-1.—*Important Medical Innovations and Associated Country of Origin*

Rank	Technology	Description	Country of Origin
1	Magnetic resonance imaging (MRI); Computed tomography (CT)	Noninvasive methods to view internal workings of the body	United States, United Kingdom; United States, United Kingdom
2	Angiotensin converting enzyme (ACE) inhibitors	Drugs to treat hypertension and heart failure	United States
3	Balloon angioplasty	Minimally invasive surgery to treat blocked arteries	Switzerland
4	Statins	Cholesterol-reducing drugs	United States, Japan
5	Mammography	Diagnostic tool to detect breast cancer	Indeterminate
6	Coronary artery bypass graft (CABG) surgery	Surgery for heart failure	United States
7	Proton pump inhibitors (PPIs); H2-receptor antagonists	Antiulcer drugs	Sweden; United States
8	Selective serotonin re-uptake inhibitors (SSRIs)	Antidepressant drugs	United States
9	Cataract extraction and lens implants	Eye surgery	United States
10	Hip replacement; Knee replacement	Joint replacement with mechanical prosthesis	United Kingdom; Japan, United Kingdom, United States

Sources: Victor R. Fuchs and Harold C. Sox Jr., "Physicians' Views of the Relative Importance of Thirty Medical Innovations," *Health Affairs*, September/October 2001. Descriptions and countries of origin from various sources.

location where the first clinically viable form of the innovation was developed or produced, or where research important to its creation occurred. The United States dominates this chart as the innovating country for these important medical developments. Of the ten, eight include the United States as a key country. The United Kingdom and Japan, the next closest sources, are associated with just two of the innovations each.

Table 10-1 should not be misinterpreted. Scientific advances by their nature are evolutionary, with recent advances building upon prior discoveries. The process of identifying a single person or team for progress that relies upon previous work is necessarily subjective. Nevertheless, such judgments are regularly made in selecting awards such as the Nobel Prize. But even taking into account the unavoidable limitations of such a list, it does suggest a dominant role for the United States in the development of new and useful medical technologies.

Box 10-1: Price Regulation and the Introduction of New Drugs

A recent study suggests that pharmaceutical firms tend to avoid or delay introducing new drugs in countries with price controls. In the study, which includes data from 25 countries on 85 new chemical entities introduced in the United States or the United Kingdom between 1994 and 1998, the three countries that did not require price approval before launch (the United States, Germany, and the United Kingdom) introduced the most new drugs. Analysis controlling for per capita income and other country and firm characteristics shows that countries with lower expected prices or smaller expected market size have fewer launches and longer launch delays. In the European Union, where drugs can be approved through a centralized procedure for use in the entire region, countries with price controls still experience significant launch delays.

According to the study, the connection between price controls and delayed access to drugs lies in the tendency for price controls to "spill over" from one country to another. Firms have an incentive to avoid or delay launching drugs in markets with price controls if they fear that the low prices will "spill over" to other markets. There are two main mechanisms by which price controls in one country can affect pharmaceutical profits in another: parallel trade and external referencing. With *parallel trade*, one country can take advantage of regulated low prices in another country through trade. With *external referencing*, countries can incorporate external price controls into domestic prices through price-setting formulas that depend on prices in other countries. Overall, the study suggests that there is a tradeoff between low prices and rapid access to new drugs.

Insurance Reform as a Means of Providing Health Care More Efficiently

While the U.S. health care market provides excellent incentives for innovation, there are legitimate concerns about cost. Rising health expenditures for families and firms can lead to difficult decisions over how best to allocate limited budgets. Pressure on government budgets continues to increase due to major health care programs such as Medicare (health insurance primarily for the elderly) and Medicaid (health insurance primarily for the poor). Physicians and hospitals struggle with government regulations, rising liability costs, and growing administrative burdens. To craft adequate responses to such challenges, it is important to understand the economic forces at work.

Technological progress in health care has been very beneficial, but it has led to growth in health care expenditures as the new technology has been applied to increase the length and to improve the quality of life. Research suggests that between 50 and 75 percent of the growth rate in health expenditures in the United States is attributable to technological progress in health care goods and services. Potential sources of the remaining 25 to 50 percent of the growth rate include: higher demand for health care due to increasing incomes and the aging of the U.S. population; the increased practice of "defensive medicine" (that is, medical procedures with limited therapeutic value that are performed by physicians to avoid lawsuits); and increased use of health insurance plans as a payment mechanism for health care.

There are various ways to reduce health care costs. Reducing the incentive to practice defensive medicine has the potential to lower the level of health care costs and is therefore an important objective. Modifying the health insurance system offers an especially attractive target for cost-saving reform because it would affect both the level and the growth rate of health expenditures. Reforms could be targeted to reduce administrative costs and the incentive to overuse health insurance as a payment mechanism. Understanding the strengths and the weaknesses of the health insurance system is central to developing policies that will lead to more cost-effective health care and to greater access to health care for those underserved by the current market.

The Appropriate Use of Insurance

Insurance is an indispensable tool in modern economies. Individuals insure automobiles against the possibility of an accident and homes against the possibility of a fire. Life insurance provides financial security to loved ones in case of an untimely demise. In each of these examples, the basic principle is the same: for a fee—the insurance premium—the insurer promises that some financial benefit will be forthcoming if a well-defined event takes place such as a car accident, a house fire, or a death.

Insurance is a valuable economic commodity. By giving up some income in the form of a premium, a consumer can avoid the large decline in wealth associated with an unfortunate event. Even if the event does not occur, a consumer benefits from the reduced uncertainty provided by insurance.

Insurance is generally not needed when there is little uncertainty or when financial risks are small. For example, insurance policies usually do not pay for items such as groceries, clothing, or gasoline, although it would certainly be possible to create such policies. Suppose, for example, that an individual could purchase a clothing insurance policy with a "coinsurance" rate of 20 percent, meaning that after paying the insurance premium, the holder of the insurance policy would have to pay only 20 cents on the dollar for all clothing purchases. An individual with such a policy would be expected to spend substantially more on clothes—due to larger quantity and higher quality purchases—with the 80 percent discount than he would at the full price. However, the insurance company would need to charge a high premium to cover expenses. The premium would need to cover the 80 percent discount on the clothing that the individual would have bought had he or she been paying full price. Additionally, the premium would need to cover the insurer's expense for clothes purchased because the individual buys clothes as if they cost only 20 cents on the dollar. Few individuals would find such an expensive policy cost-effective.

Moral Hazard

The clothing insurance example suggests an inherent inefficiency in the use of insurance to pay for things that have little intrinsic risk or uncertainty. It also illustrates the broader problem in insurance markets known as moral hazard. *Moral hazard* refers to the idea that policy holders will make different choices when they are covered by an insurance policy than when they are not, but the insurer cannot fully monitor or restrict their actions. In the clothing example, moral hazard results in insured individuals spending more on clothing than they would without insurance.

Optimal insurance contracts must balance the value that consumers place on reducing their exposure to risk against the inefficiency arising from moral hazard. In the absence of uncertainty, insurance is wasteful because moral hazard will lead to excessive use and there is no benefit to the consumer from risk-reduction. Inefficient use of insurance will be reflected in an unnecessarily high cost for insurance. Standard features of insurance contracts such as coinsurance rates, copayments, and deductibles are attempts to mitigate the moral hazard problem. Even so, inefficiencies of this sort are pervasive in the U.S. health care system.

Adverse Selection

Another issue that arises in discussions of insurance markets is adverse selection. *Adverse selection* occurs when an insurance policy attracts certain types of people, and the insurer cannot identify these people before they enroll. If the premium is based on the average individual, but the policy disproportionately attracts those who spend more than the average person (in the clothing example, individuals with particularly expensive tastes in clothes), the policy will lose money for the insurer. The policy will then either increase in price or not last in the marketplace.

Adverse selection illustrates a problem that exists when the consumer knows more about his or her characteristics than the insurer. As a result there is a market inefficiency where, in the extreme, some consumers do not purchase insurance because the only policy available to them is priced for the most expensive consumers. If insurers could distinguish among different types of consumers, policies could be tailored to specific types and priced accordingly. With better information, an efficiently functioning insurance market would be able to provide insurance in a way that would maximize individual consumer welfare.

Health Insurance in the United States

Health insurance in the United States has several unique features. First, the employer portion of premiums for employer-provided health insurance is generally exempt from income and payroll taxes. The employee portion of premiums is similarly tax-exempt for the roughly one-half of workers covered by tax-advantaged health plans. This leads to the second, and unsurprising, feature, which is that most health insurance is provided through employers. Over 60 percent of all individuals in the United States have employer-provided health insurance. The central role of employer provision makes health insurance very different from other types of insurance, such as fire and car insurance.

Third, health insurance policies in the United States also tend to cover many events that have little uncertainty, such as routine dental care, annual medical exams, and vaccinations. For these types of predictable expenses, health insurance is more like prepaid preventative care than true insurance. If automobile insurance were structured like the typical health policy, it would cover annual maintenance, tire replacement, and possibly even car washes.

Fourth, health insurance tends to cover relatively low-expense items, such as an office visit to the doctor for a sore throat. Although often unforeseeable, this expense would not have a major financial impact on most people. To continue the analogy, it would be similar to car insurance covering relatively small expenses such as replacing worn brakes.

Box 10-2: Who are the Uninsured?

The U.S. Census Bureau estimates that in 2002, 242.4 million people in the United States had health insurance for the entire year, while the remaining 43.6 million people were uninsured. Uninsurance persists in the face of public programs such as Medicare, Medicaid, and the State Children's Health Insurance Program. In general, these programs provide health insurance to the elderly, the very poor, and the children of the moderately poor, respectively.

The uninsured are a diverse and perpetually changing group. The Congressional Budget Office claims that due to sampling techniques, the U.S. Census Bureau estimate of 43.6 million (15.2 percent of the population) more closely represents the number of people who are uninsured at a point in time than the number of people who are uninsured for an entire year. Just under half of all new spells of uninsurance end within four months. The number of people who were uninsured for all of 1998 (the most recent year for which comparative survey data are available) is estimated to have been 21 million to 31 million (7.6 to 11.2 percent of the population).

Some individuals included in survey-based counts of the uninsured may in fact have access to public coverage. For instance, the number of people who report having Medicaid is smaller than the number determined to be enrolled based on the program's administrative data. The reasons for this discrepancy are not well understood. People might fail to report this coverage because of a possible stigma associated with being on Medicaid or because the survey questions are confusing. In addition, some individuals who are uninsured are eligible for Medicaid but have not enrolled. These people are counted as uninsured in surveys, but they are effectively insured because they can enroll in Medicaid should they require medical treatment.

Others who lack insurance coverage possess economic or demographic characteristics that suggest many of them may remain uninsured as a matter of choice. For example, some have levels of household income that are above the median for the population. Over 32 percent of uninsured individuals report a household income of $50,000 or more. Others have access to employer-provided coverage but do not opt to participate. Researchers believe that as many as one-quarter of those without health insurance had coverage available through an employer but declined the coverage. Still others may remain uninsured because they are young and healthy and do not see the need for insurance. In fact, more than two-fifths of uninsured individuals are between the ages of 18 and 34.

A Brief History of Health Insurance in the United States

The historical background of health insurance coverage in the United States helps explain why health insurance is different from other types of insurance. In the early twentieth century, health insurance tended to cover wage loss rather than payment for medical services. This insurance is comparable to present-day disability insurance or workers' compensation. Limited health care coverage reflected the small number of options available to the medical profession for improving health—there were few costly treatments to insure against.

The first modern health insurance policy appears to have been started in 1929 when a group of teachers contracted with Baylor University Hospital. For an annual premium of $6, the policy guaranteed up to three weeks of hospital coverage. Providing insurance through employers, rather than to individuals, lowered administrative costs for insurers. It also mitigated the problems from adverse selection because the insured group was formed without regard to health status.

Employer-based coverage was encouraged by legal provisions during World War II that allowed employers to compete for employees by offering health benefits during a period of wage and price controls. Separately, a 1943 administrative tax court ruled that some employers' payments for group medical coverage on behalf of employees were not taxable as employee income.

A consequence of exempting premiums paid on employer-provided insurance is that tax receipts to the Federal government are lower than they otherwise would be. It has been estimated that Federal tax receipts in 2001 were about $120 billion lower as a result of the tax exemption. Research suggests that the tax preference for insurance induces people to buy more expansive health insurance—for example, people buy policies that cover a broad array of health services—and policies that have low deductibles and low coinsurance rates, which lead to the associated inefficiencies from moral hazard.

To summarize, health insurance markets can be improved in at least three ways. The first is to encourage contracts that focus on large expenditures that are truly the result of unforeseen circumstances. The second is to strengthen health insurance markets outside the traditional employer-based group markets. The third is to provide a more standardized tax treatment of all health care expenditures.

Proposals for Modernizing the Health Care Market

Health insurance reforms have the potential to increase the cost-effectiveness of health care markets without sacrificing the incentives that are essential to continued innovation. Reforms that lead to more direct interaction between consumers and health care providers, relying less on third-party payers such as insurance companies, have the potential to increase the efficiency and therefore the cost-effectiveness of health care markets. Coupled with changes that provide consumers with more flexibility and more information, such reforms would continue to provide the market signals important for developing new and useful health care innovations. The President has proposed several reforms that promise to move the Nation in the direction of achieving these goals. Taken together, these reforms will help preserve the innovative strengths that have proven so valuable to Americans and will improve the efficiency of the U.S. health care system.

Medicare Prescription Drug, Improvement, and Modernization Act of 2003

The Medicare Prescription Drug, Improvement, and Modernization Act of 2003, enacted in December, adds a prescription drug benefit to the Medicare program. The new drug benefit will give more Medicare beneficiaries access to prescription drug coverage and will provide benefits for individuals with limited means and low incomes. A prescription drug discount card will be available for beneficiaries until the full drug benefit is available nationwide.

The Act also establishes another key element of the President's health care agenda, Health Savings Accounts (HSAs). With an HSA, individuals and their employers may contribute pretax dollars to fund an account that can then be used to pay for medical expenses. Once established, this money belongs to the individual and can accumulate over time. The account remains with the individual if he or she changes employers. With such accounts, there is an increased incentive to purchase insurance that only

covers events that are truly random and large, and to pay for other expenses using an HSA. Indeed, the law requires that such accounts be coupled with a high-deductible insurance plan.

With less reliance on insurance for routine health expenses, consumers would place a greater value on information about health care options and providers. More prudent use of insurance would also reduce "middle-man" costs of involving an insurance company in what could otherwise be a simple transaction between the patient and the caregiver.

Next Steps in Improving Health Care Markets

The passage of the Medicare bill was a major accomplishment, but much remains to be done. A number of proposals on the President's agenda for health care reform would lead to improvements in the health care market.

Association Health Plans (AHPs)

The AHP proposal enables small businesses and associations to purchase health insurance for employees and their families. These plans offer small businesses and self-employed individuals the potential for lower health insurance premiums resulting from decreased administrative costs and increased bargaining power with insurers and medical providers.

New Tax Deduction for Health Insurance Premiums

The President has proposed a new tax deduction for health insurance premiums. Individuals who purchase a high-deductible insurance policy coupled with an HSA would be able to deduct the value of the insurance premium from their income taxes even if they do not itemize their deductions. This would encourage the use of high-deductible insurance by providing a tax benefit similar to that given to employer-provided insurance.

Refundable Health Credit

Many workers do not have the option to obtain insurance through their employment. The President has proposed a refundable health credit that could be used to purchase insurance. This credit will help expand health care access for low- and middle-income workers who do not have good employer-based coverage options.

Reducing the Cost of Medical Care Through Liability Reform

Malpractice premiums are a significant cost for physicians and hospitals. The President has proposed the national adoption of standards to make the medical liability system more fair, predictable, and timely. Adoption of these proposals would lower the cost of providing health care (see the discussion of this subject in Chapter 11, *The Tort System*). Similarly, fear of litigation keeps health care providers from sharing vital information on quality problems and medical errors. The President has called for legislation to allay these fears and make it possible for health professionals to share information to reduce errors and complications.

Improving Efficiency Through the Use of Health Information Technology

The use of information technology in health care holds the promise of reducing medical errors, facilitating communication between care providers and patients, and reducing administrative costs. Computerized physician order entry, a type of technology that allows physicians to write medication orders electronically, has been shown to reduce significantly the rate of serious medication errors. Intensive care telemedicine, a type of technology that allows remote specialists to monitor patients continuously with video-conferencing and computer-based data transmission tools, has been found to decrease intensive care costs substantially in certain settings. The President is proposing to double the funding (for a total of $100 million) for the Department of Health and Human Services to increase the use of these new technologies through demonstration projects.

Conclusion

The U.S. health care system has provided tremendous benefits for both American citizens and the global community. New knowledge, innovative products, and life-saving medical procedures are the results of the U.S. market for health care. The proposed policies will help preserve the strengths of the U.S. market and will improve the efficiency and affordability of health care.

The Tort System

*T*ort is the civil law through which injured individuals seek compensation from another party alleged to have caused or contributed to their injury. The tort system in the United States is intended to compensate accident victims and to deter potential defendants from putting others at risk. Expenditures in the U.S. tort system were $233.4 billion in 2002, equal to 2.2 percent of gross domestic product (GDP), more than twice the amount spent on new automobiles in 2002. The expansive tort system has a considerable impact on the U.S. economy. Tort liability leads to lower spending on research and development, higher health care costs, and job losses. This chapter examines the growth of the tort system, the benefits the United States receives from it, and how alternative injury-compensation systems compare with the present tort system in terms of costs.

The key points of the chapter are:

- The evidence is mixed on whether the tort system serves to deter negligent behavior.
- The tort system is a costly method of providing insurance against injuries, and has a number of adverse effects on the economy.
- Possible ways of reducing the burden of the tort system include limiting noneconomic damages, reforming class action procedures, setting up trust funds for payments to victims, and allowing parties to avoid the tort system contractually.

The Changing Role of Tort Law

Until the 1960s, tort law covered injuries involving strangers, such as those caused by automobile accidents. Injuries resulting from the interaction between individuals with a prior relationship, such as physicians and patients, were covered by contract law instead of torts, which enabled individuals to define the terms the court would use to resolve any injury disputes in advance. This division between the tort system and contracts limited the courts' role to hearing cases involving injuries in which one person had harmed another with no predetermined specification of damages by the parties—either because no contract existed or because the existing contract did not cover a particular set of circumstances. In essence, the courts' job was to decide if the defendant was

liable (at fault) and to determine compensation for the plaintiff (the victim). An important feature of the legal environment was that courts assigned liability for an injury by applying the *negligence standard*, under which the court assessed whether the injury had occurred because the defendant had failed to exercise the caution of a reasonable person under the circumstances of the accident. Changes to tort law since the 1960s have altered the standard of care courts apply in considering claims for compensation. Although some tort cases, such as those alleging medical liability, still use the negligence standard, others, such as product liability, are now generally decided using *strict liability*. Under this standard, defendants are held responsible for any product-related injuries even if they were not negligent. More injuries have become eligible for compensation as a result of this change, thus increasing the number of injuries litigated in the tort system.

Another change since the 1960s is that the tort system now serves to provide insurance against harms relating to any goods or services consumers or businesses purchase. This function is in addition to the original purpose of punishing negligence in order to deter future injuries. The right to sue for damages means that the tort system today effectively obligates suppliers of goods and services to provide this insurance along with their products. As recently as the late 1950s, ladder manufacturers would not have been liable for falls from ladders, doctors would not have been liable for birth defects, and diving-board manufacturers would not have been liable for injuries resulting from diving; in today's tort system, they are. Courts used to presume that falls from ladders were caused by deviations from normal use and not, as is currently the case, that ladder manufacturers were potentially liable for not warning consumers about the dangers of their product.

The Expansion of Tort Costs

Expenditures associated with the tort system have risen along with its increased role in society. One estimate based on insurance industry data finds that aggregate expenditures in the tort system were $233.4 billion in 2002. This estimate includes the legal costs of defending policyholders, benefits paid to parties injured by policyholders, insurance companies' administrative costs, and estimates of medical liability and self-insurance costs. Tort costs as a percentage of GDP increased after 1974 and peaked in 1987 (Chart 11-1).

The number of injuries handled in the tort system has increased along with expenditures. The number of filings per capita started to rise in the early 1980s and peaked in the mid-1980s, at least in the 16 states for which data on lawsuit filings are available between 1975 and 2000 (Chart 11-2).

Chart 11-1 Tort Costs as a Percent of GDP

Tort costs as a percent of GDP have been rising since 1974 but, until recently, have fallen from their peak in 1987.

Percent of GDP

Source: Tillinghast-Towers Perrin, "U.S. Tort Costs: 2003 Update, Trends and Findings on the Costs of the U.S. Tort System," 2003.

Much of the decline in filings since 1985 appears to have occurred in California, where medical liability reforms included a $250,000 limit for noneconomic damages that was found constitutional in 1985. Although there has been a decline in cases per capita since the 1980s, some types of tort awards have increased. For example, between 1990 and 2001, the median award in medical liability cases increased from about $100,000 to more than $300,000.

Expenditures in the tort system vary by the type of dispute (Table 11-1). In auto cases plaintiffs received a median award of $18,000 (in 57.5 percent of the cases). The most expensive cases tended to be those in which plaintiffs and defendants had preexisting relationships, such as product liability and medical liability. Plaintiffs won 23.4 percent of the time in medical liability cases and received a median award of $286,000. The median award in asbestos cases tried in state courts was $309,000, with 56 percent of plaintiffs receiving compensation. Large awards are relatively rare. In the 75 largest counties in the United States in 1992, 73 percent of the 377,421 tort cases disposed in state courts concerned auto accidents, which tend to result in relatively small awards at trial.

Chart 11-2 **Tort Filings in 16 States**
Tort filings rose through the 1980s but have returned to 1975 levels.

Number per 1,000 people

Note: The 16 States comprise: Alaska, California, Colorado, Florida, Hawaii, Idaho, Kansas, Maine, Maryland,
Michigan, North Dakota, Ohio, Tennessee, Texas, Utah, and Washington.
Source: National Center for the Study of State Courts.
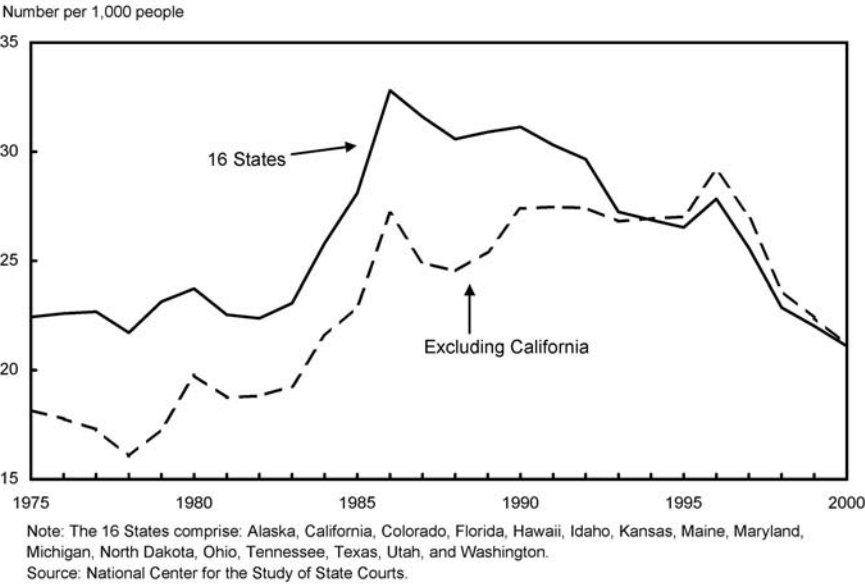
TABLE 11-1.— *Characteristics of State and Federal Tort Cases Decided by Trial, 1996*

Tort cases by type	Cases won by plaintiff (percent)	Median award	Percent of awards $250,000 or more	
			Total	$1 million or more
All tort cases				
State	48.2	$31,000	16.9	5.8
Federal	45.8	139,000	38.1	14.6
Automobile cases				
State	57.5	18,000	8.7	3.4
Federal	59.7	100,000	37.4	11.6
Medical liability				
State	23.4	286,000	51.0	20.2
Federal	39.8	252,000	54.3	22.9
Asbestos				
State	55.6	309,000	50.6	12.1
Federal	40.0	465,000	50.0	0.0
Product liability other than asbestos				
State	37.1	177,000	41.2	16.3
Federal	26.6	368,500	62.0	24.0

Source: Congressional Budget Office.

The Economic Effects of the Tort System

The economic effects of the tort system go beyond their direct impact in terms of expenditures. Resources that could be directed toward productive uses are diverted instead to the tort system or dissipated as firms and individuals take actions not needed for actual safety concerns but rather to avoid exposure to tort liability. Studies suggest that the gains to society from tort compensation and deterrence do not make up for these losses. A study of the impact of tort reform on productivity finds that limitations on the size of tort claims (for example, caps on punitive damages) enacted by states from 1972 to 1990 increased productivity by 1 to 2 percent a year, an amount equal to $955 per worker per year in 2002 dollars. Limitations on tort awards moved some injury payments out of the tort system so that the $955 figure represents an estimate of the cost of the tort system over alternative systems.

The gains from limits on the tort system come about because torts cause firms and individuals such as medical professionals to change the way they do business. Firms choose not to sell certain products so that they can avoid potential liability or they take costly extra precautions in the delivery of their products and services—precautions beyond the level that would reasonably balance costs and benefits to society. For example, torts cause doctors to practice defensive medicine, such as ordering extra tests that are a waste of time and resources. Some expenditures in the tort system, such as compensation for damages, are transfers of money from defendants to plaintiffs and do not consume resources. Other expenditures involve true economic costs in that the resources involved are not available for more productive uses; attorney's fees are an example. Additional costs include the profits and consumer benefits forgone by society when a potential defendant removes a product or service from the market or does not produce it in the first place in order to avoid frivolous lawsuits.

Torts as Injury Compensation

The tort system is not the only way in which society can deter injuries and compensate victims. There is an extensive system of regulations to improve the safety of products, medicines, and many other goods and services. Consumers have access to numerous publications and Internet Web sites that offer reviews and facilitate discussions of products. The availability of this information on product safety provides producers with a powerful financial incentive to make their products safer.

The question then becomes whether another system could provide the same benefits in terms of compensation and deterrence as the tort system but at lower cost. There is not enough evidence to determine the answer to

this broad question. Nevertheless, some evidence indicates that in certain areas, such as product liability and medical liability, the tort system does not deliver enough deterrence benefits to justify the associated administrative costs (such as legal fees, overhead to process insurance claims, and the cost of running the tort system itself).

The Principal Injury-Compensation Methods

Injury-compensation systems can be broadly classified by the type of act that leads to the compensation being provided. A *fault-based* system compensates the injured party on the basis of negligent action, intentional harm, or strict liability. In contrast, a *cause-based* system is one in which the specific cause of the injury entitles an individual to compensation. The most widespread cause-based program in the United States is workers' compensation, which pays for many workplace injuries regardless of whether the employer was negligent with regard to the worker's injury. Finally, *loss-based* systems pay compensation based only on injury or illness. Loss-based systems include private systems like health insurance and public systems like Medicare.

The tort system is not the principal means by which injuries are compensated. Private health insurance, Medicaid, and Medicare are all substantially larger providers of compensation than the tort system (Table 11-2). The portion of tort expenditures that covers only economic damages such as current and future lost wages (that is, not including noneconomic damages such as for pain and suffering) is comparable in size to either the workers' compensation system or payments for life insurance.

Administrative Costs

The tort system is one of the most expensive compensation systems to run, with administrative costs equal to 54 percent of benefits. Sixty-one percent of these administrative costs (about a third of every dollar spent in the tort system) are the legal fees generated by attorneys for plaintiffs and defendants. In 2001, administrative costs of the health insurance industry were around 14 percent of benefits paid. The overhead for the Social Security disability system was around 3 percent of benefits in 2003; a study from the mid-1980s found that workers' compensation had overhead costs of around 20 percent of benefits. Some of the high cost of the tort system may arise because it deals with accidents that are more difficult to evaluate than those of other injury-compensation mechanisms.

TABLE 11-2.— *Compensation for Injury, Illness, and Fatality in the United States, Selected Methods*

Type of injury or illness compensation system	Compensation (billions of 2002 dollars)
Fault-based	
Tort economic payment[1]	51.3
Tort noneconomic payment[1]	55.9
Cause-based	
Workers' compensation[2]	48.0
Veterans' benefits[2]	26.0
Loss-based	
Health insurance (private first-party)[2]	408.2
Life insurance (private first-party)[2]	46.1
Social/public insurance	
Health:	
Medicaid[3] and Medicare[2]	362.1
Medicaid prescription drug[4]	13.1
Disability:	
Social Security Disability[1] and Supplemental Security Income[2]	94.1

[1]Data are for 2002.
[2]Data are for 2000.
[3]Data are for 1999.
[4]Data are for 1998

Sources: Department of Commerce (Bureau of Economic Analysis); Social Security Administration; Centers for Medicare and Medicaid Services; American Council of Life Insurers, "Life Insurers Fact Book," annual; and Tillinghast-Towers Perrin, "U.S. Tort Costs: 2000, Trends and Findings on the Costs of the U.S. Tort System," February 2002.

Compensation of Noneconomic Losses

Another way in which the tort system differs from other compensation methods is that it forces consumers to accept not only coverage for economic losses such as current and future lost wages and medical costs, but also nonpecuniary losses such as pain and suffering. Of the 46 cents of each dollar spent in the tort system that goes to plaintiffs, on average, 22 cents compensates them for economic losses and 24 cents compensates them for noneconomic damages.

Damages paid through the tort system are costs to firms—and higher costs ultimately translate into higher prices for goods and services. Tort awards can thus be seen as a form of insurance: consumers pay "premiums" in the form of higher prices for goods and services and receive compensation if injured. Torts cover only a limited set of possible injuries, however, so a consumer seeking comprehensive insurance against all possible economic and noneconomic losses would still have to purchase additional insurance. In reality, few people buy insurance against noneconomic losses such as pain and suffering; people do buy insurance against economic losses such as lost wages, medical expenses, or costs to rebuild a damaged house. This suggests that insurance policies against noneconomic losses are not worth their cost to potential buyers.

Extent of Coverage

Despite the expansion of the tort system, torts still provide compensation for a relatively limited number of injuries compared to other systems such as health insurance. For example, injuries that are the sole fault of the victim do not give rise to a legal claim for compensation and hence do not fall under the purview of the tort system. Many injuries are too small in economic terms to justify litigation. The long delays inherent before the tort system delivers monetary compensation likely also dissuade many potential lawsuits from being filed. In tort cases resolved in the 75 largest counties in the United States in 1992, the median time from filing to disposition was just over two years, with nearly one out of six cases taking more than four years. For medical liability, the median time to resolution was nearly three years with almost three out of ten cases taking longer than four years.

There is evidence that the eventual compensation does not match the injury well. In medical liability cases, the tort system appears to overcompensate minor injuries relative to the compensation that would have been provided by private insurance, while more serious injuries are undercompensated. This discrepancy may exist because factors other than the medical specifics of the injury could affect the compensation received by the plaintiff. For example, the location of the trial and the composition of the jury pool appear to affect the verdicts of some tort lawsuits and the size of the compensation. In addition, compensation may be tied more to the ability of the defendant to pay than to the actual injury suffered by the plaintiff. This is particularly a concern for punitive damages (Box 11-1).

Moreover, the tort system does a poor job of identifying which injuries are entitled to compensation and which are not. Many injuries that would meet the legal definition of negligence are never pursued, and the majority of those that are pursued appear not to merit compensation. A 1984 study of the outcomes of hospitalizations in New York City found that 3 to 4 percent of hospitalizations gave rise to adverse events such as drug reactions, with just over one-quarter of these due to negligent actions. However, more than half of the medical liability claims actually filed in the tort system arose from circumstances in which neither negligence nor any identifiable injury was present. One-third arose from instances in which the patient was injured but the doctor was not negligent (for example, for injuries resulting from a previously unknown drug allergy). Only one-sixth of the cases identified instances of true negligence and injury. Moreover, in this study, these claims represented a small fraction of injuries that actually arose due to negligence. Consequently, the majority of the compensation went to people who were not injured or were not injured by the doctor accused of malpractice, while the majority of those actually injured by doctor error were not compensated at all. Only in a minority of cases did those legally entitled to compensation receive it through the legal system.

Box 11-1: Punitive Damages

Compensatory damages are intended to "make the plaintiff whole" by offsetting an injured victim's losses. Punitive damages, on the other hand, are intended to punish the party whose negligent action caused the injury. Defendants may be liable for punitive damages if a jury finds that their actions were malicious, oppressive, gross, willful and wanton, or fraudulent. The Department of Justice studied civil trial cases in the country's 75 largest counties and found that punitive damages were awarded in 4.5 percent of cases that plaintiffs won (or 2.3 percent of all cases), but represented 21 percent of all damages awarded to plaintiffs. The median punitive award was $40,000 in those cases in which the plaintiff received an award. The threat posed by large punitive damages is that they may encourage more frequent and larger settlements.

Some are concerned that punitive damages are awarded against companies because they have deep pockets rather than because they have behaved egregiously. Indeed, the Supreme Court has expressed unease over the fact that the size of certain punitive awards has seemed out of proportion to the wrongfulness of the defendant's actions. This capriciousness also has implications for the deterrence effect of punitive damages, because a deterrence effect can be realized only if firms are able to take specific actions to avoid liability. If firms cannot tell which actions will likely incur liability, they cannot avoid them. Anecdotal evidence suggests that punitive-damage awards can indeed be unpredictable. Two identical allegations of fraud against BMW were heard in the same Alabama court and before the same judge. One purchaser was awarded $4 million in punitive damages; the second purchaser received no punitive damages.

Torts As Deterrence

The threat of a lawsuit can create and enforce appropriate standards of behavior. If the tort system made products and services in the United States safer, fewer accidents would occur and the higher administrative cost of torts would provide benefits to society in terms of reduced injury rates and associated health care costs. For example, the move by a number of states to no-fault automobile insurance in the 1970s appears to have led to as much as a 15 percent increase in the highway fatality rate. Such no-fault auto insurance laws eliminate or restrict liability for auto accidents so that each driver's own insurer typically pays for his or her own accident costs regardless of how

the accident happened. Drivers who know that they will not be financially liable for other drivers' injuries in the event of an accident might be expected to take fewer safety precautions than if they were responsible for the financial consequences of their actions. In other areas of tort law such as medical liability and product liability, there is not consistent evidence that deterrence effects are large enough to justify the considerable administrative costs of the tort system. This suggests that alternatives to the tort system provide deterrence. For example, the possibility of losing a medical license could provide an adequate incentive for doctors to take steps to avoid negligence beyond the steps doctors take in the interests of their patients.

General Aviation and Deterrence

The experience of the general aviation industry over the past several decades provides an example of the role of tort liability in affecting product safety, firm profits, and the availability of goods to consumers. General aviation is the segment of the aviation industry composed of all civil aircraft not flown by commercial airlines or the military. General aviation manufacturers were the targets of a large volume of litigation in the 1970s and 1980s.

The general aviation accident rate has been declining for 50 years (Chart 11-3). In 1963, court rulings made lawsuits alleging manufacturing defects in the design of private and commercial aircraft subject to strict liability. In the most extreme cases, this meant that firms were responsible for accidents even if the accidents were caused by product defects that were not known or knowable at the time of manufacture. By the mid-1970s, this change in the law had led to a sharp rise in the number of product-liability cases and increased liability costs for the general aviation industry, with liability awards increasing nearly ninefold from 1977 to 1985.

The merits of these product-liability claims against airplane manufacturers were subject to question. A study of a sample of general aviation lawsuits filed between 1983 and 1986 showed that none of the accidents that led to lawsuits was caused by a design or manufacturing defect, as each suit had claimed. Thus, these lawsuits did not give manufacturers any additional incentives to produce safer aircraft, since the allegations of design defects appear to have been specious in the first place.

Indeed, the rise in tort claims had no discernible effect on the accident rate. An examination of the trends in the accident rate calculated over various periods shows that the steepest decline in general aviation accidents occurred between 1950 and 1969—before the dramatic rise in tort costs in the 1970s and 1980s (Chart 11-4). If liability exposure were driving the general aviation industry to build safer products, accident rates would have declined more rapidly as the increased likelihood of tort litigation pushed aircraft manufacturers to add safety features to their aircraft.

Chart 11-3 General Aviation Liability Payouts and Accident Rates

The increase in liability payouts between 1977 and 1985 did not cause a change in trends of the number of accidents or fatal accidents per 100,000 flight hours.

Percent of base year
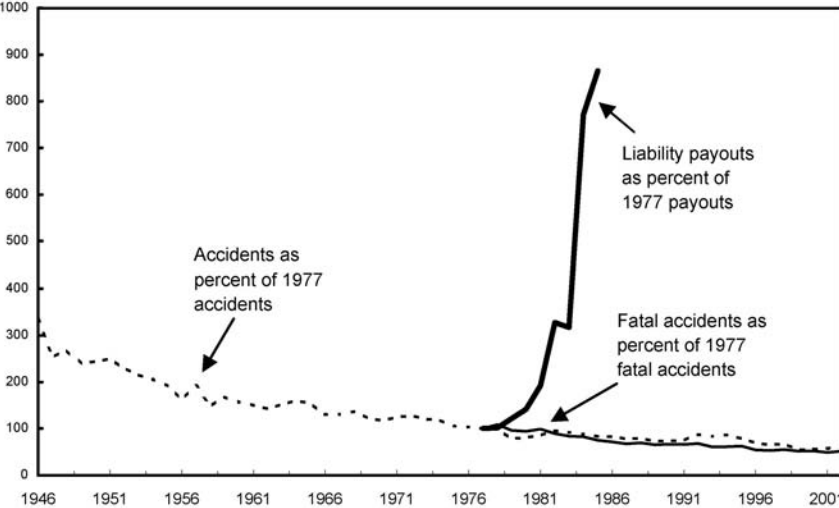
Sources: George L. Priest, "The Modern Expansion of Tort Liability: Its Sources, Its Effects, and Its Reform," *Journal of Economic Perspectives*, Summer 1991, and General Aviation Manufacturers Association.

Chart 11-4 Accident Rate for Small Aircraft

The rise in tort claims has had no discernible effect on the declining accident rate in general aviation. The steepest decline was between 1950 and 1969, which predated the rise in tort costs during the 1970s and 1980s.

Accidents per 100,000 flight hours

Years
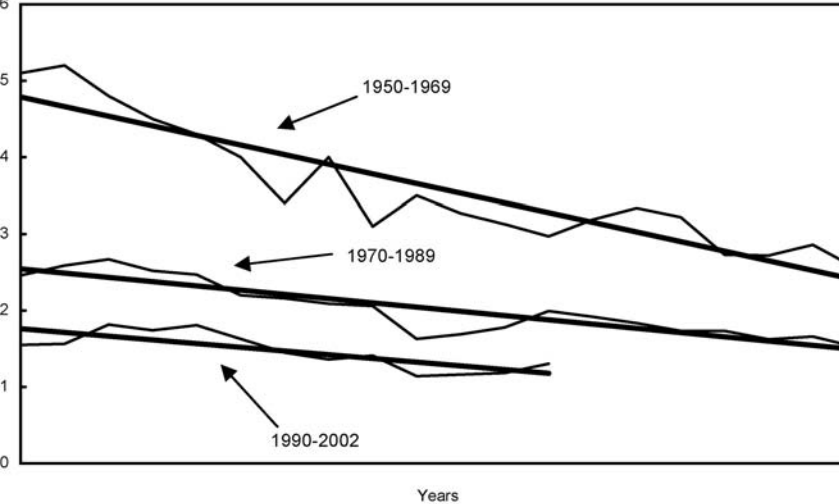
Source: General Aviation Manufacturers Association.

The rise in liability expenses did, however, cause great harm to the general aviation industry. During the period of expanding liability costs from 1977 to 1985, the financial health of the general aviation industry deteriorated markedly, with a number of firms shutting down production lines and one going bankrupt. As a result, small-aircraft production fell precipitously (Chart 11-5). By discouraging the production of new planes, tort law has created a situation in which the mix of planes in use actually presents a higher risk than would have been the case had older planes been retired and replaced by new ones. The General Aviation Revitalization Act of 1994, which exempted some general aviation aircraft older than 18 years from product-liability claims, appears to have led to a small resurgence in the industry.

Chart 11-5 **Small-Aircraft Production**
The increase in tort liability beginning in the 1970s caused a decrease in small-aircraft production. This effect was attenuated by the General Aviation Revitalization Act of 1994.

Percent of 1970 production
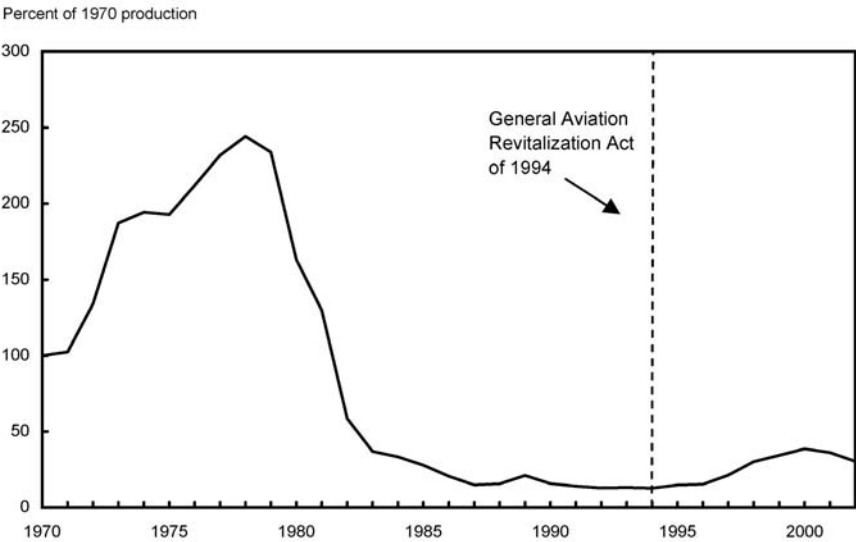
Source: General Aviation Manufacturers Association.

Other Evidence on Deterrence

It is difficult to find deterrence effects in other contexts. For example, studies examining injury rates for consumers and workers as well as death rates from workplace injuries show that such injuries did not decline more rapidly following a steep increase in litigation. Other research has examined the deterrence effect of medical liability by estimating the impact on treatment outcomes of state-imposed limits on damage awards at trial (such as California's $250,000 limit on noneconomic damages). Studies have found no appreciable impact on treatment outcomes—the lower threat of torts did not lead to more medical injuries. These findings suggest that there is at best limited deterrence from such cases.

The Limits of Tort Deterrence

Why does the tort system appear to be ineffective in improving product safety? One major reason is that market incentives already provide an important form of deterrence against unsafe products. Firms whose products cause injuries lose customers and suffer economic losses. In addition, many products and services face government regulation. The producers of such items are required to undertake investments in safety, and the tort system may have no incremental effect on safety. Similarly, medical services also face market incentives and regulation by governmental and professional bodies.

The current tort system makes it hard to predict which actions will be deemed negligent during litigation. Thus, the system does not provide much deterrence because people do not know what steps to take to avoid a lawsuit or an adverse judgment.

Potential Tort Reforms

One way to consider the effects of changes in the U.S. tort system is to compare the U.S. system with those in other advanced economies, such as Canada, Japan and the United Kingdom. Like the United States, many of these countries use a negligence standard for medical liability and strict liability for product-related injuries, yet they expend fewer resources in their tort systems than the United States (Chart 11-6). Possible explanations for this divergence are discussed in the following sections.

Percent of GDP
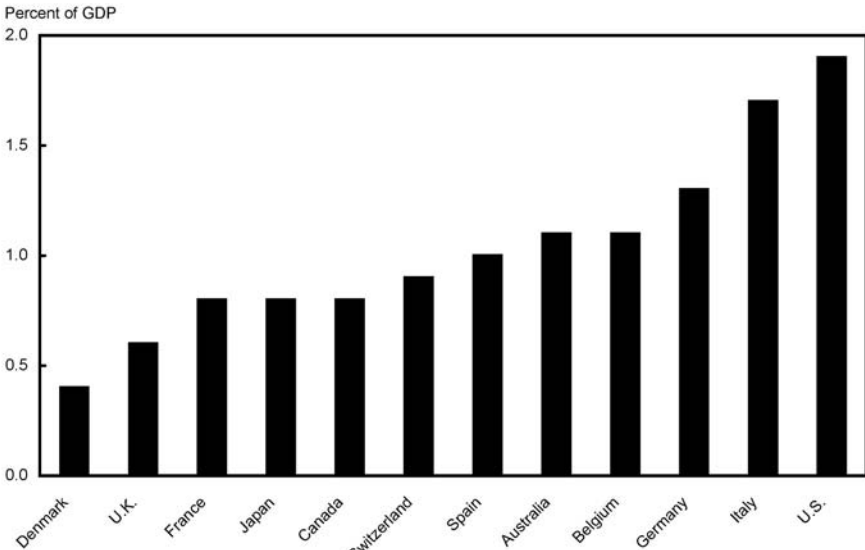
Source: Tillinghast-Towers Perrin, "U.S. Tort Costs: 2000, Trends and Findings on the Costs of the U.S. Tort System," February 2002.

Limiting Noneconomic Damages and Other Potential Reforms

One important reason for the divergence in tort costs between the United States and other countries is that awards for noneconomic damages, such as pain and suffering, appear to be much higher in the United States. Noneconomic damages account for half of all compensation awarded in the United States, but in other countries are either capped (as in Canada) or otherwise restricted (as in Germany). Reforms aimed at reducing or eliminating pain and suffering awards, such as the President's proposed $250,000 limitation on noneconomic damages in health-related cases, have the potential to reduce the cost of the U.S. tort system.

Several other differences appear to be less important in explaining the divergence than compensation for noneconomic damages. One difference is that in other countries, judges decide the vast majority of tort claims, while juries typically decide cases in the United States. Empirical evidence suggests that U.S. judges and juries decide cases in approximately the same way, suggesting this is not a major factor in explaining the divergence. Another difference is that in the United States each side pays its own legal costs, whereas in many other nations the losing side pays both sides' legal costs. A study of Florida's temporary use of a "loser-pays" method in medical liability

cases found that when the losing side paid legal expenses, plaintiffs were more likely to receive compensation either at trial or in a settlement. Furthermore, the compensation was higher. This finding suggests that apportioning legal costs to the losers discourages plaintiffs from pursuing low-quality (nuisance) cases because they would have to pay all legal costs if the case went against them.

Procedural Reforms

Some of the costs of the tort system arise because there are incentives that encourage state judges and juries to extract financial compensation from out-of-town defendants. The vast majority of tort cases are litigated in state courts. Tort cases tried before elected state judges have been found to result in higher awards when the defendant is a corporation headquartered outside of the state than when the defendant is local. By removing national class action suits from state courts, the Federal government could reduce the ability of entrepreneurial lawyers to *forum shop*, that is, to file cases in a sympathetic state court. Some evidence on asbestos tort litigation suggests that forum shopping is indeed a problem. Research also suggests that certain small counties tend to be magnets for national class actions in the sense that they attract many more cases than would be expected on the basis of their populations.

The Class Action Fairness Act of 2003 would allow removal of some class actions to Federal court if any plaintiff is from a different state than any defendant (Box 11-2). Under current law, a plaintiff's attorney who does not like a particular judge's limitations in a class action can seek a less restrictive judge in a different jurisdiction. The proposed Act would make this more difficult by reducing the ability of plaintiffs' attorneys to file national class actions in state court.

Limiting the Scope of Tort Compensation

An alternative approach to the current system would be to resolve disputes and compensate victims outside the tort system. An example of this approach is the case of compensation for individuals exposed to asbestos. The proposed Fairness in Asbestos Injury Resolution Act of 2003 would create a trust fund to compensate those injured by asbestos exposure. Disbursements from the fund would be restricted to those who are actually suffering from asbestos-related illnesses. The use of asbestos has been all but abandoned in the United States, so the focus in resolving claims is now appropriately placed on compensating injured workers rather than deterring new instances of future liability (Box 11-3).

Box 11-2: The Role of Class Actions in the Tort System

A *class action* is a legal procedure in which individuals are joined together to litigate a single case (the class refers to the group of such individuals). Class actions are used in a variety of contexts, including cases involving securities fraud, consumer protection, employment, civil rights, and exposure to toxic chemicals or other pollutants. Class actions are intended to secure compensation in cases that involve substantial aggregate losses but relatively small individual losses. In practice, private attorneys often initiate these cases, each one in effect becoming a "Private Attorney General." In this role, lawyers identify both the legal violations and a number of individuals harmed by the violations and bring an action on these individuals' behalf. To induce attorneys to take on this role, they are compensated out of the settlement fund. In many cases, this compensation is based on a *contingent fee*, a percentage of the settlement or award.

An important concern about class action suits is that many of them are filed more for the benefit of the plaintiffs' attorneys than for the plaintiffs. In individual litigation, plaintiffs enter a contract with an attorney and have an incentive to monitor the attorney's effort to ensure a favorable outcome. In class action suits, most individual plaintiffs have only a small stake in the case's outcome and thus have little incentive to monitor the activities of their lawyers. In principle, judges are expected to monitor payments to plaintiffs' attorneys and the nature of settlements. With growing caseloads, however, many judges face pressure to clear their dockets as rapidly as possible. Accepting a settlement and associated attorneys' fees is one way to accomplish this.

Without the active scrutiny of clients or judges, plaintiffs' lawyers have an incentive to collude with defendants to set higher attorney's fees in exchange for lower overall payouts from defendants to plaintiffs. One study of a small number of class action cases found that in a substantial fraction of them, class counsel received more in fees and expenses than all of the plaintiffs combined.

Box 11-3: Asbestos and the Tort System

The tort system's treatment of asbestos cases demonstrates how the system can fall short of its purported objectives of deterring harmful behavior and funding compensation. Beginning in the 1970s, increased public awareness and concern about the health effects of asbestos led to regulations limiting exposure to asbestos. By 1989, all new uses were banned, and strict regulations have limited remaining asbestos use. Between 1973 and 2001, asbestos use in the United States fell by 98 percent. With extensive regulations in place and minimal use, the tort system's role in deterring harmful behavior has been substantially reduced simply because there is little activity to deter.

Yet even as the use of asbestos declined, the number of claims rose substantially. The total number of claimants is estimated to have grown from 21,000 in 1982 to over 600,000 by the end of 2000. To be sure, some additional claims are warranted because cancers caused by asbestos can take years to develop. An estimated 90 percent of the new claims, however, are by people who have no cancers and may never develop cancer. Claims by individuals without a diagnosed asbestos-related cancer account for almost all of the growth in asbestos case loads during the 1990s and most of the compensation received by claimants goes to those without malignant cancers. Only 43 percent of the money spent on asbestos litigation is recovered by claimants—the rest goes to lawyers and administrative costs. In short, the current system neither achieves deterrence in the use of this dangerous substance nor directs appropriate compensation to its victims.

Instead, asbestos litigation has imposed costs on workers, share-holders, and those who in the future will become ill from their previous exposure to asbestos. Estimates suggest that roughly 60 companies entangled in asbestos litigation have gone bankrupt primarily because of asbestos liabilities, with most of the bankruptcies occurring since 1990. One study estimated that between 52,000 and 60,000 workers were displaced because of these bankruptcies. Moreover, bankruptcy results in a shrinking pool of money to be divided up among future claimants. The growing number of bankruptcies raises concerns that those who become ill in the future will receive little or no compensation.

For other injuries, a possible approach to compensating accident victims would be a system akin to workers' compensation, in which compensation would be provided by an insurance system. New Zealand has replaced the personal injury and medical liability aspects of its tort system with a government-run compensation system. Such a system, however, can increase the prevalence of accidents because fully-insured individuals may not take sufficient care against a loss. This is not a concern in cases where accidents have already occurred, such as asbestos exposure. In other cases, such as product liability or medical liability, the effect of changes in the system on the behavior of potential victims is an important consideration. Moreover, like the tort system, workers' compensation systems tend to be costly to administer and may encourage frivolous claims. Replacing the tort system with a more general workers' compensation system could well mean replacing one costly and inefficient system with another.

Avoiding the Tort System

Recontractualization is an alternative approach to reform that has been the subject of considerable academic discussion. According to this idea, individuals and firms would be allowed to specify by contract the types of damages for which injurers would be liable. For example, consumers or their insurers could determine individual caps on damages in exchange for lower prices for goods and services. In principle, potential defendants would enter into such contracts if they reduced the expected costs of dealing with injuries. Such a system would be voluntary, so that individuals could refuse to participate if offered a contract by a potential defendant that was inferior to the insurance associated with the tort system.

A possible drawback to this approach is that the courts currently view contracts limiting damages or defining negligence with suspicion. Courts have held that warranties that limit liability are not enforceable because they are *contracts of adhesion*—agreements that the purchaser of a product or service has no choice but to accept. Hence, it is likely that any steps toward recontractualization would require substantial institutional and legal changes. This could explain why this approach has not received much attention from policy makers.

Conclusion

The tort system has expanded in the last 30 years. By expanding the number of accidents for which accident victims receive compensation, the current tort system in effect requires the suppliers of goods and services to provide insurance to their customers. This tort-based insurance against accidents appears to be more expensive than other methods of compensating victims. At least in the cases of product liability and medical liability, the expansion of the tort system does not appear to have had an appreciable effect in deterring negligent behavior.

The President has proposed several initiatives to reduce the burden of torts on the economy. These include placing limits on noneconomic damages, reforming class action procedures, and finding alternative methods to compensate injuries such as those that have been proposed for people suffering from asbestos-related ailments. These steps would focus the tort system on those cases it can deal with most effectively and lessen the costs to society of frivolous lawsuits and awards.

International Trade and Cooperation

Since the end of the Second World War, international trade has grown steadily relative to overall economic activity. Countries that have been more open to international flows of goods, services, and capital grew faster than countries that were less open to the global economy. The United States has been a driving force in constructing an open global trading system. A series of international trade agreements has reduced barriers to trade in goods and services and has been an important element in U.S. and global growth.

During this period, new types of trade emerged and delivered new benefits to consumers and firms in trading countries. Growing international demand for goods such as movies, pharmaceuticals, and recordings offered new opportunities for U.S. exporters. A burgeoning trade in services provided an important outlet for U.S. expertise in sectors such as banking, engineering, and higher education. The ability to buy goods and services from new places has made household budgets go farther, while the ability of firms to distribute their production around the globe has cut costs and thus prices to consumers.

The key points in this chapter are:

- Trade has grown significantly since World War II. The benefits from new forms of trade, such as trade in services, are no different from the benefits of traditional trade in goods.
- The benefits of integration are substantial.
- International cooperation is an essential part of realizing the potential gains from international trade. A system through which countries can resolve disputes can play an important role in realizing these gains.

Increased Trade Flows: Facts and Trends

One way to measure the relative importance of international trade is to compare the value of trade flows to overall economic activity. In the latter half of the twentieth century and into the twenty-first, growth in world trade has outpaced growth in world output (Chart 12-1). As recently as 1950, the sum of merchandise exports of all countries equaled only 8 percent of world GDP. In 2002, the most recent year for which data are available, exports had increased to 19 percent of world output. For the United States, the sum of merchandise exports and imports rose from 7 percent to 18 percent of GDP over the same period.

Chart 12-1 **World Trade and GDP**
World merchandise exports have grown faster than world GDP from 1950 to 2002.

Index, 1950=100

Source: World Trade Organization.
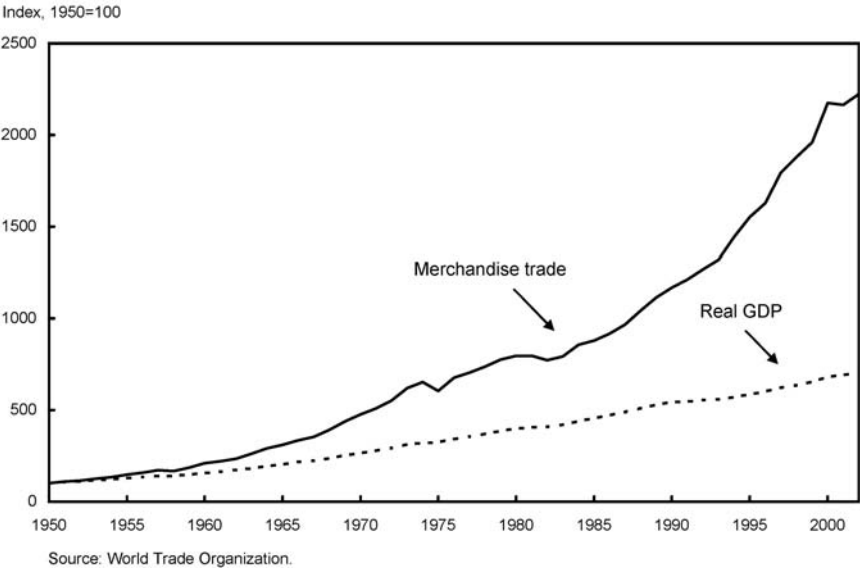
Increased trade has been accompanied by increased output growth, which can be attributed, at least in part, to the opening of markets and to the benefits derived from the international trading system. The growth in world trade also reflects lower transportation costs, which facilitate trade; new production processes, which allow companies to produce and assemble goods in different countries; and information technology, which facilitates communication between buyers and sellers. These allow trading countries to take advantage of variations in resource endowments and sectoral productivities across countries.

The United States is the largest importer and exporter of goods and services in the world, although its prowess at exporting is sometimes less apparent to the casual observer than the country's demand for imports. This is because many of the products that U.S. firms export are capital goods used in production and are not sold at the retail level to consumers (Table 12-1). The composition of U.S. exports reflects its abundance of skilled labor and high-technology expertise relative to other countries. This relative abundance explains why the United States is more likely to export aircraft and semiconductors and import footwear and clothing.

U.S. export levels depend partly on the vitality of export markets. When U.S. trading partners such as Europe and Japan experience slow growth, as has occurred in recent years, they import fewer goods from the United States. The developed countries of North America, Europe, and Japan still account for roughly two-thirds of world imports and exports of goods, and over 70 percent of world trade in services.

TABLE 12-1.— *Leading U.S. Net Exports of Goods, 2002*

Rank	Category	Millions of dollars
1	Aircraft	15,675
2	Semiconductors and related devices	15,233
3	Aircraft parts and auxiliary equipment	9,482
4	Plastics materials and resins	6,819
5	Soybeans	5,597
6	Oil and gas field machinery and equipment	5,296
7	Corn	4,988
8	Aircraft engines and engine parts	4,624
9	Motor vehicle parts	3,976

Source: Department of Commerce (International Trade Administration).

The Benefits of Free Trade

The benefits of free trade are often misunderstood. Discussions of the gains from trade often focus on the jobs created in industries that export goods and services rather than the benefits to consumers and producers from importing. The jobs created by exports are important—indeed, some research suggests that workers in export industries tend to have higher wages than those in other industries. The benefits of trade, however, are much greater. In fact, the claim that free trade is good mainly because it allows us to export misses much of the story. Free trade is good not just because it allows us to export, but also because it allows us to import. Providing goods and services to people in other countries is worthwhile because it allows Americans to consume the goods and services made in other countries. This is analogous to why most people work at their jobs—to earn the income with which to buy goods and services. That is, people "export" the product of their efforts and in return receive income with which to buy goods and services made by other people.

The benefits of exports are similar. The advantage of selling goods and services abroad is that U.S. exporters receive funds that can be spent on imports for Americans to consume. Imports allow Americans to purchase more varieties of goods and services at lower cost than if the same items were obtained from domestic producers. These cost savings free up resources to be used to produce other products. In this way, imports raise the standard of living in the United States.

Comparative Advantage

Free trade does not require that one country gains at another country's expense. Free trade is win-win. Just as the United States benefits from goods produced more cheaply abroad, other countries benefit from goods built more efficiently here. Each country gains from these exchanges because each has different capabilities. Free trade encourages countries to specialize in what they do best. Such a division of tasks raises economic well-being around the world, just as the specialization of individual workers into different jobs makes a company more productive.

Free trade also pushes American businesses to become as efficient as possible by exposing them to competition from foreign firms. For example, foreign competition over the past several decades has spurred improvements on the part of U.S. automakers. American firms and workers responded to the challenge of international competition by improving American cars and making them less expensive. American consumers are better off as a result of increased choice and better value.

Barriers to trade, in contrast, tend to help a relatively small number of firms and their workers at the expense of harming a much larger number of consumers who pay more for their goods as a result of protection. Each consumer might pay only modestly more while the beneficiaries of the protection gain substantially. The total financial costs of protection borne by consumers, however, are typically larger than the benefits that accrue to producers and workers.

The effects of trade policy on economic growth and the mechanisms by which trade affects growth have been controversial. In part, this is because it is difficult to disentangle the effects of trade liberalization on economic growth from the effects of the multitude of other policies that countries adopt. As late as the 1950s and 1960s, the idea that open markets spur economic growth was somewhat unconventional. The more common belief was that developing countries should close their borders to imports in order to support and encourage the growth of their own firms. This approach became known as "import substitution" because countries sought to develop home industries in place of imports. This was believed to be particularly important for the manufacturing sector. Advocates of this view pointed to positive past experiences with protection among currently developed countries. Developing countries that followed this strategy and tried to substitute domestic production for imports often found initial success, but subsequently encountered serious economic difficulties.

Broad comparisons of countries' experiences support the assessment that openness to trade is significantly correlated with economic growth. One study examined the experience of 133 countries from 1950 to 1998. Countries' annual real incomes per capita grew about ½ percentage point

faster after liberalizing trade policies than under their closed regimes. Further, the income gains from opening up to free trade have become increasingly significant; countries that removed trade barriers in the 1990s raised their growth rates 2½ percentage points, an additional 2 percentage points per year. While the results of these cross-country studies are not irrefutable, their findings are bolstered by studies of individual countries' problems with trade protection and successes with liberalization.

Assisting People and Communities Affected by Free Trade

Although openness to trade provides substantial benefits to the Nation as a whole, foreign competition can require adjustment on the part of some individuals, businesses, and industries. To help workers affected by trade develop the skills needed for new jobs, the Administration has built upon and developed programs to assist workers and communities that are negatively affected by trade. The Administration has reformed existing programs to make them more responsive and flexible. For example, the long-standing Trade Adjustment Assistance program offered training and income support to workers directly hurt by greater imports. This program was significantly enhanced by new legislation signed by the President in 2002 to extend eligibility to workers indirectly affected, such as upstream suppliers of the firms hurt by imports. The new legislation also expanded the benefits to include a health insurance tax credit and a wage supplement for older workers who found new jobs that did not pay as well as the jobs they had lost. This assistance, which will total $12 billion over 10 years, helps ease the adjustment for displaced workers and helps them move into jobs where they are most needed. In addition, the President has proposed a pilot program for Personal Reemployment Accounts, which would offer an innovative approach to worker adjustment. These accounts would provide unemployed individuals funds they can use for training, for job-search assistance, or as a cash reemployment bonus if they find new work quickly.

The creation and destruction of jobs is part of the way in which people and materials move from less-productive to more-productive functions in a free-market economy. Businesses fail and jobs are lost for many reasons; for example, changes in technology or new domestic competition can shake up industries and communities. In the 1980s, 70 percent of the changes in employment in U.S. manufacturing resulted from less demand for relatively low-skilled workers and greater demand for high-skilled workers within the same industry. This indicates that the job losses in the 1980s were not primarily due to foreign trade pushing workers out of a sector, but to the changing nature of manufacturing. Import competition, however, often receives a

disproportionate share of the blame. This may be because there is less that can be done to prevent the dislocations associated with technological change.

New Facets of Trade

The nature of U.S. trade has changed dramatically over the last several decades. Whereas the United States once would have exported your father's Oldsmobile in exchange for foreign-made food or clothing, the United States is now as likely to export financial or educational services, Hollywood blockbusters, or life-saving medicines. The United States still imports food and apparel, but it also imports components that go into sophisticated products (such as computer hard drives). This section explores several ways in which modern trade has evolved from the classic exchange of manufactured and agricultural goods.

Intellectual Property

The kinds of goods that have been traded for centuries, such as wine or clothing, have two important attributes: the value of the good is linked to the physical object, and it costs roughly the same to produce the second unit of the good as the first. Many of the goods in which the United States now excels—movies, books, music, software, and pharmaceuticals—are dramatically different from traditional goods. The value of a book, movie, or computer software program lies in the ideas contained within, more than in the paper and binding or disk. The cost of producing the first book includes not just the paper and ink, but the intellectual contribution of the author. To produce the second copy of the book, however, only the raw materials are required, which makes it significantly less expensive. As discussed later in the chapter, trade in goods with valuable intellectual property raises different policy questions than does more-traditional trade.

Services

Services trade is growing in importance in the world economy (Chart 12-2). The services sector, for trade purposes, includes travel and transportation-related services, royalties and license fees, and other private services, such as finance, insurance and telecommunications. The service-providing sector is the largest component of the private economy in the United States, providing more than 86 million jobs in 2003 and accounting for over half of total GDP. In 2002, the United States exported services worth almost $300 billion, about 30 percent of total exports of goods and services.

Worldwide services trade totaled $1.5 trillion in 2002, compared to goods trade of over $6 trillion, but services trade has been growing faster. Unlike

Chart 12-2 World Trade in Goods and Services
Imports and exports of services are a growing part of total trade.

Trillions of U.S. dollars
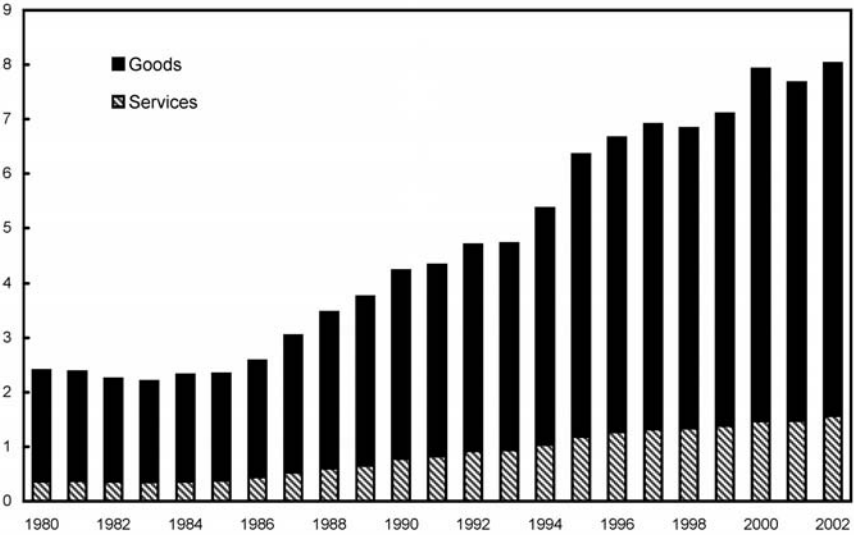
Source: World Trade Organization.

goods trade, in which a product can be loaded on a ship at one port and off-loaded anywhere in the world with little need for the exporter and importer to interact, services trade generally requires extensive interaction. Some services can be provided at a distance, such as software services. For others, such as tourism, the customer must come to the location of the service provider. For others, such as some consulting work, the service provider must come to the customer. The liberalization of services trade involves the movement of individuals as well as the regulation of investment and other business activity. For American banks to sell many of their services abroad, they must open branches in their target markets (Box 12-1). As a result, negotiations to liberalize trade in services have moved beyond border measures such as *tariffs* (taxes on imports) to deal with subjects that have traditionally been the domain of domestic regulation.

One facet of increased services trade is the increased use of *offshore outsourcing* in which a company relocates labor-intensive service industry functions to another country. For example, a U.S. firm might use a call-center in India to handle customer service-related questions. The principal novelty of outsourcing services is the means by which foreign purchases are delivered. Whereas imported goods might arrive by ship, outsourced services are often delivered using telephone lines or the Internet. The basic economic forces behind the transactions are the same, however. When a good or service is produced more cheaply abroad, it makes more sense to import it than to make or provide it domestically.

Intra-industry Trade and Intermediate Products

In classical descriptions of trade, a country with abundant land sends corn to a country with abundant capital in exchange for automobiles. In modern trade, it is common for two countries to send machine tools back and forth to each other. While the items traveling in both directions might all be machine tools, they are distinct products that draw on similar production capabilities. Even when a country is technologically capable of producing all varieties of a product, it is cost-effective to specialize in producing particular varieties and then trade with partner countries to obtain other types of the product. This type of trade is referred to as *intra-industry trade*.

Modern trade also differs from classical trade because the production of any given product may be spread across several countries. Final assembly might occur in the United States, for example, using parts (*intermediate inputs*) that were built in Canada and Brazil. In fact, a good deal of U.S. trade involves flows of intermediate inputs used for domestic production.

This kind of trade can have very different economic effects. In the conventional trade model, an increase in imports would drive out domestic production and jobs in the import-competing sectors. Evidence indicates, however, that increases in imports are correlated with increases in domestic employment in the same product category. One explanation for this result is that when production is integrated across countries, an increase in demand can stimulate both domestic and foreign production.

International Cooperation and Disputes

Countries can benefit from cooperation that increases trade. This has always been true for the shipment of goods across borders, but it is even more essential for the new types of trade described above. Trade in services and goods with high intellectual-property content often requires a deeper involvement on the part of the exporter in the importing country, as in the case of U.S. bank branches overseas.

Why Is There a Need for Cooperation?

Even if a nation understands and accepts the benefits of importing, there may still be an incentive to intervene in trade through policies such as tariffs. Countries that are large enough to affect world prices can potentially benefit by limiting their demand for imports and moving the *terms of trade* (the relative price of exports to imports) in their favor. If two large countries try to do this to each other, however, they can make their situations worse than under free trade.

One reminder of this lesson was the aftermath of the Smoot-Hawley Tariff Act of 1930. Though the United States had a trade surplus before 1930, the pressures of the nascent Great Depression led Congress to raise tariffs in an ill-conceived attempt to protect American jobs. Trading partners around the world responded by raising their own trade barriers. This was an important factor in the ensuing breakdown of international commerce, contributing to lower employment worldwide.

Many of the post-World War II international economic institutions established under U.S. leadership, such as the World Bank and the International Monetary Fund, were responses to perceived failures in international economic policy in the prewar period. The plan under which these two institutions were created also included a proposal for an organization, the International Trade Organization (ITO), to oversee cooperation in international trade. The ITO was never established due to political disputes.

For more than four decades the trading system was governed instead by a series of agreements known as the General Agreement on Tariffs and Trade

(GATT). GATT only became part of a formal organization, the World Trade Organization (WTO), in 1995. Despite the absence of a standing international body, substantial progress was made in global trade liberalization.

In the first GATT negotiations in the late 1940s, a relatively small group of countries, including the United States, looked for opportunities in which they could all benefit from reciprocally lowering barriers. This gathering to seek mutual gains from cooperation was known as a "round." The current multilateral trade talks were launched by well over 100 countries in Doha, Qatar, in 2001.

Early trade talks were primarily devoted to cutting tariffs. This era of import liberalization coincided with and contributed to an era of rapid worldwide economic growth. While tariff cuts could be painful for industries that faced new competition from imports, the United States gained better market access for exports, while consumers and firms benefited from lower prices of imports. At a practical level, tariff cutting was relatively easy. If the United States and France each had 40 percent tariffs in sensitive sectors, they could agree to cut those tariffs to 20 percent. Because of this simplicity, as well as the limited number of participating countries, the early GATT trade rounds were brief. Over time, however, GATT negotiations became more comprehensive and more complex. The negotiations were held less frequently and lasted much longer. Nonetheless, a good deal of progress was made in liberalizing world trade. Among developed countries, successive tariff cuts on manufactured goods lowered average tariff levels to below 5 percent. Barriers remained higher in developing countries.

Nontariff barriers to trade remain, but they are often more difficult to address. For example, countries' policies on protecting intellectual property can constitute a nontariff barrier with important trade consequences (Box 12-2). Other types of regulations could, if misused, also constitute a barrier to trade. For example, "sanitary and phytosanitary regulations" are rules designed to protect the health of people, plants, and animals. A foreign government seeking to block competition in a sensitive agricultural sector could seek to ban imports on the basis of a product-safety claim that is without a sound basis in science. The standard that was agreed upon in the Uruguay Round of trade talks in 1994 was that such claims must be based on sound scientific evidence. What constitutes such evidence has been the subject of dispute. This circumstance poses a challenge: trade restrictions based on sound science must be allowed and claims not founded on sound science must be avoided or dismissed, but determining the difference is frequently not an easy process.

Box 12-2: International Cooperation on Intellectual Property Rights

The protection of intellectual property is an important new trade issue. The United States has worked to ensure that copyrights, trademarks, and patents given to authors, companies, programmers, or other inventors are protected in other countries.

One implication of the high development and low production costs of goods with high intellectual property content is that they are relatively easy to steal. While it may cost $80 million to create a feature film, the blank videotape or DVD used to copy that film may cost just a few dollars. It is a fairly straightforward matter for the United States to prevent other countries from taking U.S. wheat without paying. It is more difficult to prevent an exported copy of a movie, recording, or drug from being reproduced, though the loss to the United States in forgone exports would be just as significant. These losses can occur not only through unauthorized duplication, but also through foreign government policies such as controls on drug prices. These price controls reduce the return that U.S. producers can earn from abroad and shift the burden of paying for development costs to the American consumer.

As trade in goods embodying valuable intellectual property has grown, the protection of intellectual property has emerged as an important policy concern. In the Uruguay Round of trade talks, which concluded in 1994, participating countries agreed to adopt high standards of intellectual property protection in the accord on Trade-Related Aspects of Intellectual Property Rights (TRIPS). Some have misconstrued it as preventing developing nations from addressing health emergencies such as the spread of AIDS in Africa. At Doha in 2001, WTO members agreed that the TRIPS Agreement does not and should not prevent members from taking measures to protect public health. Furthermore, in 2003, the United States and other WTO members agreed that developing countries that lack domestic manufacturing capacities in the pharmaceutical sectors should be able to override patent rights to import needed medicines from abroad in order to deal with domestic health problems.

The Administration has actively pursued measures in trade agreements to ensure the security of U.S. intellectual property rights. The inclusion of these measures in trade agreements illustrates a new way in which international cooperation benefits the United States. If countries are found to be in violation of their obligations under a trade agreement, the United States could retaliate against those countries across the entire range of transactions covered by the agreement.

The Benefits of Dispute Settlement

Another issue that arises with international cooperation in trade is the need for some way of solving disagreements among trading partners. Disputes might occur when one country disregards a commitment it made in negotiations, or when there is a disagreement over the interpretation of an agreement.

If one WTO member has a complaint about the behavior of another member, there is an established process for addressing the concern. First, the two countries are required to consult and determine whether the dispute can be resolved amicably. If this is not possible, a dispute settlement panel is established at the WTO. This panel consists of experts, generally selected from countries not involved in the dispute, who hear evidence from the complaining and responding countries and then issue findings. The panel determines whether a country has failed to follow through on commitments previously made in its trade agreements. Panel findings can be appealed to a standing body, which issues its own report and findings on the issues on appeal. Panel and Appellate Body reports are then submitted to the Dispute Settlement Body (DSB), also a standing body, for adoption. Once adopted, these findings become DSB recommendations and rulings.

After the conclusion of the dispute process, the difference between the WTO system and the domestic legal system becomes apparent. All WTO members have agreed that when a country loses a dispute settlement case at the WTO, the first preference is to bring the domestic law into compliance with the DSB recommendations and rulings. However, if the losing country chooses to maintain its initial policies, it must either negotiate compensation to the complaining country or else the complaining country can get authorization from the WTO to retaliate by withdrawing concessions of comparable value. If the latter happens, the net effect is the unwinding of the reciprocal liberalization that the countries had undertaken.

The virtue of an orderly dispute-settlement system that has the confidence of all participants is that the unraveling of cooperation is limited. Parties not involved in the dispute handle the facts and interpretation of the dispute, reducing the scope for disagreement over whether retaliation is legitimate.

The United States has had much success in complaints it has initiated against other countries' trade practices. As of September 2003, the United States had filed 63 complaints against other countries. Of the 39 that have been resolved through panel proceedings, the United States lost only 3 in litigation. In turn, the United States was a respondent in 77 cases over the same time period and successfully defended its practices in 4 of the resulting 38 panel proceedings. These statistics suggest that WTO complaints are not brought frivolously, in the sense that complaints, whether by or against the United States, have a high probability of success.

An effective dispute-settlement mechanism that has the confidence of all participants is an important part of the cooperative trading system. A dispute-settlement system can help to ensure that all parties to trade agreements receive the benefits on which they agreed.

Progress Toward Free Trade

The United States has pursued trade liberalization through negotiations at the global, regional, and bilateral levels. This multipronged approach allows for continuing progress even when one avenue for liberalization is blocked or stalled. Due to its global reach, the broadest and most important forum for liberalization is the World Trade Organization. This body now has 148 members. Among the central principles of the WTO is the requirement that the lowest tariff offered to one WTO member must be offered to all members. This principle, known as most-favored-nation (MFN) treatment, ensures that even if cooperative agreements are reached among a smaller group of countries, those countries will extend the benefits broadly to other WTO members. Although WTO rules permit important exceptions to the MFN principle, such as allowing countries to lower barriers with trade-agreement partners and as part of trade preference programs for poor countries, when the MFN principle is observed it creates a "level playing field" of equal tariffs on all trading partners so that countries will buy goods from the most-efficient producer.

The WTO encompasses agreements made under the GATT, as well as agreements on trade in services, intellectual property, and other issues. The WTO is driven by its members. It does not serve as a legislative body and passes no laws. What the WTO provides is a forum for countries to come together to negotiate. When there are decisions to be made, they are reached by consensus of the members rather than by majority vote. The principal task of the WTO Secretariat is to support the work of member countries as they pursue the goal of trade liberalization.

The Administration played a critical role in launching the Doha Development Agenda negotiations in 2001, following the failure of the 1999 Seattle ministerial meeting to initiate new multilateral trade negotiations. Participating nations agreed that the negotiations would focus on the needs of developing countries and their integration into the global trading system. The United States has put forward proposals for liberalization of trade in agriculture, consumer and industrial goods, and services—the three major areas for market access under negotiation. The Administration is committed to a successful completion of the Doha Development Agenda. This would substantially lower barriers to trade in all countries and provide expanded

market access for American goods and services, while boosting economic prospects for developing countries. One study estimates that removal of tariff barriers, production subsidies, and export subsidies could raise annual world income by over $355 billion by 2015. According to another study, a successful round that lowered trade barriers around the world could raise the level of U.S. GDP by $144 billion each year, which translates into additional annual income of $2,000 or more for a family of four.

The WTO operates by consensus, so it takes little to halt progress. While the Administration seeks to continue work on global trade negotiations through the WTO, it has also independently pursued trade liberalization with developed and developing nations through far-reaching bilateral and regional agreements (Table 12-2). These free trade agreements (FTAs) remove substantially all barriers to trade between participants and allow for cooperation in other areas of concern, such as regulation of investments and the protection of intellectual property, the environment, and labor rights. Under WTO rules, countries may undertake preferential liberalization in a free trade agreement, as long as the accord is comprehensive and the liberalization is completed in a reasonable period of time.

TABLE 12-2.— *Status of Free Trade Agreements (FTAs) with the United States*

Country or Region	Status
Israel	In effect since April 22, 1985
Mexico and Canada (NAFTA)	In effect since January 1, 1994
Jordan	In effect since December 17, 2001
Singapore	In effect since January 1, 2004
Chile	In effect since January 1, 2004
Australia	In negotiation as of January 2004
Morocco	In negotiation as of January 2004
Central America (CAFTA)	
El Salvador, Guatemala, Honduras and Nicaragua	Negotiations concluded on December 17, 2003
Costa Rica	Negotiations concluded on January 25, 2004
Dominican Republic	In negotiation
Southern African Customs Union (Botswana, Lesotho,	
Namibia, South Africa, and Swaziland)	In negotiation
34 Western Hemisphere Countries (FTAA)	In negotiation
Bahrain	In negotiation
Thailand	Intentions to negotiate announced
Panama	Intentions to negotiate announced
Columbia, Peru, Bolivia, and Ecuador	Intentions to negotiate announced

Source: U.S. Trade Representative.

For each potential trading partner in a free trade agreement, the United States assesses the economic benefits such an agreement would bring to the United States, the extent to which the country is ready to undertake free trade obligations, and the role that the agreement would play in furthering the broader, worldwide trade-liberalization agenda. Throughout the process of selecting and negotiating with FTA partners, the Administration consults with members of Congress, public-interest groups, and industry representatives. The United States has demonstrated its willingness to liberalize trade with countries from around the world, both developing and developed. These agreements offer the benefits of trade and investment to the United States and our partner countries and help build a coalition of nations interested in achieving progress in multilateral talks.

The United States has worked to rapidly expand its set of FTA partners, while maintaining low trade barriers to goods and services from all countries through our global commitments.

Conclusion

The United States has benefited and continues to benefit enormously from the international exchange of goods and services. Trade allows countries to specialize in those activities that make the best use of their skills and resources, as well as to reap the benefits in terms of imported goods. These gains have increased as lower barriers, better transportation, and easier communication have expanded existing international markets and created new ones.

Another important but often overlooked benefit to the expansion of free trade is the expansion of freedom and democracy. Involvement in the global economy provides incentives for nations to ensure a degree of transparency and stability in order to attract investors and trading partners. It also encourages countries to embrace a more democratic and less corrupt system of government. Economic freedoms can lead to greater political freedoms.

As the complexity of international trade has increased, so too has the complexity of the agreements that govern it. The dispute-settlement mechanism in the WTO has been useful for resolving disagreements between WTO members. The United States has been challenged on certain trade practices, but in turn has used the dispute settlement system to assert its rights and challenge the practices of other countries.

The Administration is committed to an open and unfettered trading system to promote economic growth in the United States and around the world.

International Capital Flows

International capital flows are the transfer of financial assets, such as cash, stocks, or bonds, across international borders. They have become an increasingly significant part of the world economy over the past decade and an important source of funds to support investment in the United States. In 2002, around $700 billion flowed into the United States. Inflows of international capital help to finance U.S. factories, support U.S. medical research, and fund U.S. companies. At the same time, U.S. investors provided nearly $200 billion in capital to other countries for a wide range of purposes.

Around $2 trillion flowed into countries around the world in 2002, equivalent to roughly 6 percent of global GDP (Chart 13-1). Although these world capital flows have dropped from a peak of over 13 percent of GDP in 2000, largely reflecting a global economic slowdown, they remain above the level of the early 1990s.

Chart 13-1 **Global Capital Flows as a Percent of World GDP**
The 1990s saw a surge in global capital inflows. Flows have since declined, but remain above their level in 1992.

Percent of world GDP
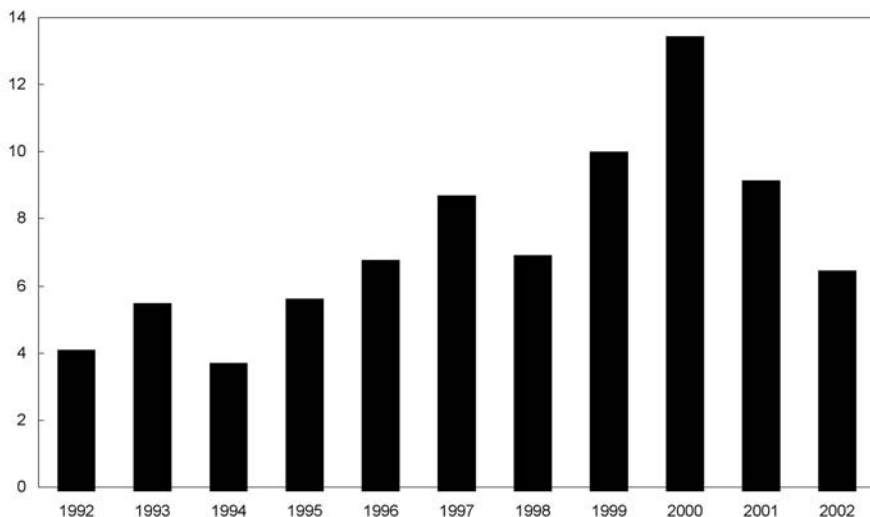
Source: International Monetary Fund.

This chapter describes the various types of international capital flows and discusses their benefits, as well as their risks. The key points in this chapter are:

- Capital flows have significant potential benefits for economies around the world.
- Countries with sound macroeconomic policies and well-functioning institutions are in the best position to reap the benefits of capital flows and minimize the risks.
- Countries that permit free capital flows must choose between the stability provided by fixed exchange rates and the flexibility afforded by an independent monetary policy.

Types of International Capital Flows

Not all capital flows are alike, and there is evidence that the motivation for capital flows and their impact vary by the type of investment. Capital flows can be grouped into three broad categories: foreign direct investment, portfolio investment, and bank and other investment (Chart 13-2).

Chart 13-2 **World Capital Inflows in 2002**
World capital inflows, which include direct investment, portfolio investment, and bank and other investment, totaled $2 trillion in 2002.

World capital inflows = $2,040 billion

Bank and other investment 30%

Direct investment 25%

Portfolio investment 45%
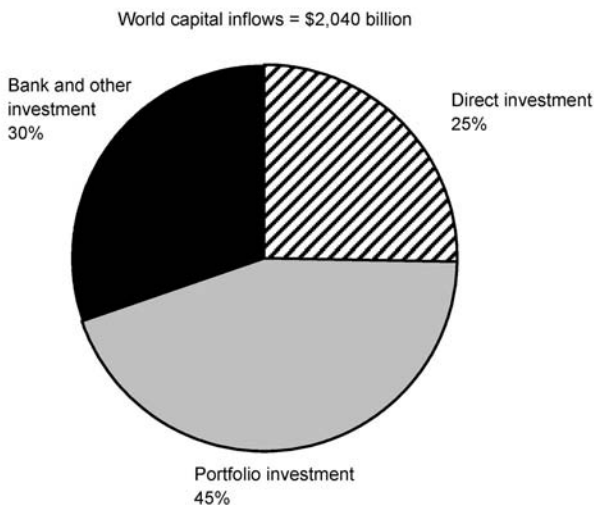
Source: International Monetary Fund.

Foreign Direct Investment

Foreign direct investment occurs when an investor, in many cases a firm rather than an individual, gains some control over the functioning of an enterprise in another country. This typically takes place through a direct purchase of a business enterprise or when the purchaser acquires more than 10 percent of the shares of the target asset.

A number of factors affect the flow of foreign direct investment. Trade links between investor and recipient countries tend to increase foreign direct investment, as demonstrated by the establishment of Japanese auto plants in the United States starting in the 1980s. Proximity to foreign markets also plays a role, as shown by the investment of U.S. companies in China to service Chinese consumers and firms. The political, economic, and legal stability of the recipient country also matters. Investors are reluctant to establish ownership of foreign companies or set up businesses abroad if corruption or political or social instability are likely to jeopardize operations.

In 2002, foreign direct investment made up roughly one quarter of world capital inflows. About 40 percent of these flows went to the major industrial countries—the United States, Canada, the United Kingdom, Japan, and countries in the euro zone. During much of the 1990s, the United States was the largest single recipient of foreign direct investment. Foreign direct investment flows to industrialized countries are driven largely by the desire for better distribution networks and market access. Another 30 percent of total foreign direct investment went to emerging markets. Relative to flows to industrial countries, these investments were driven more by the low production costs and growing markets of Asia, as well as the privatization of state-owned enterprises in many countries in Latin America and Eastern Europe.

Portfolio Investment

Portfolio investment occurs when investors purchase noncontrolling interests in foreign companies or buy foreign corporate or government bonds, short-term securities, or notes. This type of investment accounted for almost half of world capital inflows in 2002.

Economic and financial conditions in the recipient and investor countries are important influences on portfolio investment flows. The market for these assets is typically more liquid than that for direct investments; it is usually easier to sell a stock or bond than a factory. As a result, investors can quickly reshuffle portfolio investments if they lose confidence in their purchases. Not surprisingly, portfolio investment is far more volatile than foreign direct investment. Countries that receive large capital inflows in one year can see a quick reversal of these inflows if economic or political developments cause investors to reevaluate the expected return on their assets.

Sudden and destabilizing reversals of portfolio investment took place in countries such as Korea, Mexico, Russia, Brazil, and Argentina during the second half of the 1990s and early 2000s. These reversals partly reflected the concern that private-sector and government borrowers in emerging market economies might be unable to meet their financial obligations.

In the United States, portfolio investment in U.S. government securities has played an increasingly important role since 2001. Foreign purchases of U.S. government securities rose from 3 percent of total capital inflows in 2001 to 33 percent in the first three quarters of 2003. One of the most important factors explaining this change is a shift in the share of U.S. security purchases by foreign investors from equities into lower-risk assets, such as U.S. government obligations. Another important factor is increased purchases of U.S. government securities by foreign central banks. A decline in the number of mergers and acquisitions in the United States has also led to lower foreign purchases of private assets.

Bank Investment

Bank investment is the third major type of capital flow. Bank-related international investment includes deposit holdings by foreigners and loans to foreign individuals, businesses, and governments. These investments, grouped with a few other miscellaneous types of investments, accounted for over one quarter of total international capital inflows in 2002. For emerging markets, the importance of these bank-related and other investment flows has declined dramatically in the past decade. While these flows represented an average of 28 percent of capital inflows to emerging economies from 1992 to 1996, they represented an average of only 3 percent of inflows from 1997 to 2002. Economic crises in a number of Asian and Latin American countries since the mid-1990s have contributed to reduced bank lending to these regions since 1997, notably from banks in Japan and Europe.

Benefits of International Capital Flows

Capital flows can have a number of important benefits:
- International capital allows countries to finance more investment than can be supported by domestic saving, thereby increasing output and employment.
- Greater access to foreign markets can provide new opportunities for foreign and domestic investors to increase the return and reduce the risk of their portfolios.

- Foreign direct investment can facilitate the transfer of technology and managerial expertise to developing countries, thus improving productivity.
- Better risk management and other management techniques associated with foreign direct investment can help recipients modify their production processes to lower costs and raise productivity.
- Exposure to international capital markets and the resulting increased competition may induce governments and firms issuing assets to improve macroeconomic policy, management, and profitability. These improvements may, in turn, encourage additional foreign investment.
- Improved international access to investment opportunities in the country receiving capital inflows expands the number of potential investors in any domestic project. This will tend to reduce the cost of raising capital.
- Increased capital inflows can spur the development of domestic financial sectors. A well-developed financial sector can lead to greater investment and reduced financial-sector vulnerability.

Empirical evidence suggests that countries that are open to capital flows can enjoy many of these benefits. In the case of foreign direct investment, studies indicate that industries and some developing countries with more foreign direct investment grow faster than those with less foreign direct investment. In addition, extensive research has found that foreign-owned firms tend to have higher productivity and wages than do their domestic counterparts. Finally, for some developing countries, foreign direct investment can help catalyze the adoption of more-advanced technologies and management practices.

Foreign portfolio investment has played a key role in furthering the development of domestic equity and bond markets. In the case of equity markets, one report estimates that opening up to foreign shareholders leads to an almost 40 percent increase in the real dollar value of the stock market. This lowers the cost of equity capital for domestic firms, as a higher stock price means that a smaller portion of a company needs to be sold to raise a given amount of capital. Developing equity markets can help restrain the ability of corporate managers to pursue their own goals and can help align managerial incentives with earnings growth. In the case of debt markets, evidence indicates that foreign investment can widen the investor base and help businesses raise capital. Moreover, developing countries that lack debt markets may rely excessively on bank lending. Studies suggest that this may leave economies more vulnerable to financial crises because banks are less likely to hold well-diversified portfolios than are participants in developed bond markets.

For all of these reasons, financial market liberalization has been linked to greater investment and higher output growth. One study found that equity market liberalization raised annual economic growth by about 1 percentage point per year in the five years following liberalization. In a related study, the same researchers showed that 17 out of a set of 21 countries that opened their equity markets to foreign participation experienced faster average-growth rates than before liberalization.

A foreign banking presence can also have substantial benefits for the host economy. Foreign-owned financial institutions have been shown to improve the standards and efficiency of the domestic banking sector. This can raise the net yield on saving and enhance capital accumulation and growth. In Latin America, studies have shown that foreign banks in the latter half of the 1990s had higher and less-volatile loan growth than the average domestic bank. Foreign banks may also be a stabilizing force during periods of financial stress. This is partly because foreign banks are often better capitalized and have access to financing through their parent companies at times when domestic banks might be unable to raise capital. Because foreign banks are often better managed and less exposed to domestic downturns, they can also provide citizens some insurance against a collapse of the domestic banking sector. Drawing on the experiences of the Asian crises, academic work suggests that the greater the foreign bank presence in a developing country, the less likely the country was to experience a banking crisis. The ability to hold bank accounts in other countries and borrow from overseas financial institutions can also facilitate trade.

Risks of International Capital Flows

Many countries that reduced barriers to capital flows in the 1990s experienced large capital inflows, increased investment, and strong growth. Several of these countries, however, subsequently experienced economic crises. In the majority of these crises, capital outflows were associated with currency depreciations. The governments, firms, and citizens of many of these emerging markets had significant amounts of debt denominated in foreign currency but received income denominated in domestic currency. The currency depreciations therefore greatly impaired the capacity of these borrowers to service their debts. The resulting increase in bankruptcies and, in some cases, government defaults, weakened the banking sectors and other financial institutions in these countries. All of these factors contributed to sharp contractions in output and high unemployment rates. Such "currency crises" occurred in Mexico, Thailand, Korea, Russia, and Argentina from the mid-1990s through 2001. These experiences have led to a more guarded view of the advantages of capital flows.

One lesson learned from these crises is that a strong institutional framework is important if a country is to benefit fully from openness to capital flows. In other words, capital flows are more likely to yield substantial benefits and carry fewer risks in countries where the financial system is strong and well developed; laws and regulations are clear, reasonable, and enforced by the courts and public institutions; and the reporting of financial information is timely and accurate so that investors have a clear understanding of the conditions and strength of the assets in which they are investing. Corruption is also associated with lower foreign investment and weaker growth.

In countries with weak institutions or high levels of corruption, capital inflows may not be channeled to their most-productive uses, dissipating their potential benefits. In these cases, improved access to capital can allow firms and sovereigns to accumulate high levels of debt through purchases of unproductive assets. This can ultimately leave firms and countries vulnerable to changes in investor sentiment, possibly contributing to economic crises.

One approach to limiting these risks when legal and financial institutions are poorly developed is to restrict foreign capital flows. Experience, however, suggests that capital controls impose substantial costs. Controls on the movement of capital can distort firms' investment decisions, increase opportunities for corruption, and discourage foreign direct investment. All of these effects can depress growth (Box 13-1).

Box 13-1: Capital Controls in Emerging Markets

Recent economic crises in several emerging economies that opened their markets to capital flows have renewed debate on the desirability of capital controls. Any benefits of restrictions on capital flows, however, must be weighed against the costs and distortions they impose.

Capital controls can take various forms and can target either capital inflows or capital outflows. Countries may adopt *controls on capital inflows* in an attempt to prevent an appreciation of their currency or to direct foreign investments to longer-term ventures. Experience shows that these controls, regardless of whether they achieve their objective, can create problems, including economic distortions and large administrative fees. For example, in the 1990s, the Chilean government required that a portion of capital inflows be temporarily deposited in a non-interest-bearing central bank account. These restrictions lowered the risk of rapid capital flight, and some analyses show that they lengthened the average period of time that capital inflows remained in Chile. These restrictions, however, also increased administrative costs, especially because the government had to

---

Box 13-1 — *continued*

modify them frequently to close numerous loopholes. Research also shows that these controls on capital inflows caused smaller, public firms to face greater financing constraints than they did before the restrictions. These higher financing costs may have stifled an important source of growth and innovation in Chile.

Countries' experiences with *controls on capital outflows* reinforce the view that controls are difficult to implement and often carry unexpected costs. Controls on capital outflows also take a variety of forms, such as limitations on the amount of domestic holdings of foreign currency and restrictions on the ability of foreign investors to repatriate their earnings. The potential to avert financial crises triggered by capital outflows can make controls appealing in theory. In practice, however, any such benefits tend to be eroded over time as firms and individuals find ways to circumvent the restrictions. Such evasive activity can create additional problems, such as reduced financial transparency and tax compliance, distortions from the unequal impact of the controls (as not all sectors have equal access to the evasive measures), and a general reduction in respect for the law. For example, studies indicate that controls on capital outflows in Russia in the mid-1990s were evaded by exporters, particularly in the energy sector, through the underreporting of earnings.

Finally, capital controls can also distort the behavior of foreign investors. For example, research indicates that American multinational firms invest less in their local affiliates in countries with capital controls. In addition, multinationals tend to alter their investment and payment structure in order to minimize the effect of the restrictions. This distortion is yet another way capital controls can reduce the productivity of the world's stock of capital.

Another approach for developing countries to minimize the risks from opening up to capital movements involves the careful timing, or sequencing, of policies designed to "liberalize" financial markets. One variant of this approach suggests that countries should first achieve macroeconomic stability, in part by implementing sound fiscal and monetary policies. Countries should next strengthen financial market institutions, and only then allow for free capital flows. While this approach may work for some countries under specific economic conditions, the pace and timing of reforms appear to be less important than the consistency of the reforms and the government's commitment to them.

Policy makers increasingly realize that there is no simple rule to best achieve free capital flows, and that country characteristics should be considered. There is some consensus, however, that the benefits of international capital mobility can be substantial and that to best achieve these benefits, countries should implement reforms of domestic financial and legal institutions.

Constraints Imposed by Free Capital Flows

One consequence of allowing capital to flow freely in and out of a country is that this constrains a nation's choice of monetary policy and exchange-rate regime. For important but subtle reasons related to the tendency for capital to flow to where returns are the highest, countries can maintain only two of the following three policies—free capital flows, a fixed exchange rate, and an independent monetary policy. Economists refer to this restriction as *the impossible trinity*. As illustrated by Chart 13-3, countries must choose to be on one side of the triangle, adopting the policies at each end, but forgoing the policy on the opposite corner.

Chart 13-3 "**The Impossible Trinity**"
Countries can adopt only two of the following three policies -- free capital flows, a fixed exchange rate, and independent monetary policy.
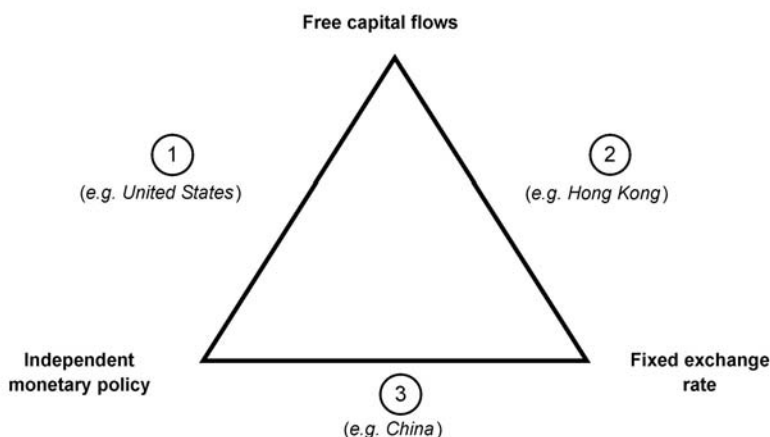
Source: Council of Economic Advisers.

The easiest way to understand this restriction is through specific examples. The United States allows free capital flows and has an independent monetary policy, but it has a flexible exchange rate. (The U.S. government does not attempt to fix, or "peg," the exchange value of the dollar at any particular level against other currencies.) As a simplified example, if the Federal Reserve Board raised its target interest rate relative to foreign interest rates, capital would flow into the United States. By increasing the demand for U.S. dollars relative to other currencies, these capital inflows would increase the price of the dollar against other currencies. This would cause the exchange rate to adjust and the U.S. dollar to appreciate. In the opposite case, if the Federal Reserve Board lowered its target interest rate, net capital outflows would reduce the demand for dollars, thereby causing the dollar to depreciate against foreign currencies.

In contrast, Hong Kong essentially pegs the value of its currency to the U.S. dollar and allows free capital flows. (Hong Kong is a Special Administrative Region of China, but maintains its own currency.) The trade-off is that Hong Kong loses the ability to use monetary policy to influence domestic interest rates. Unlike the United States, Hong Kong cannot cut interest rates to stimulate a weak economy. If Hong Kong's interest rates were to deviate from world rates, capital would flow in or out of the Hong Kong economy, just as in the U.S. case above. Under a flexible exchange rate, these flows would cause the price of the Hong Kong dollar to change relative to that of other currencies. Under a fixed exchange rate, however, the monetary authority must offset these flows by purchasing domestic or foreign currency in order to keep the supply and demand for its currency fixed, and therefore the exchange rate unchanged. The capacity of the government to sustain large purchases and sales of its currency is ultimately limited by several factors, including the amount of foreign exchange reserves held by the government and its willingness to accumulate stocks of relatively low-return foreign currency assets.

Just as in the case of Hong Kong, China pegs its exchange rate to the U.S. dollar. China can operate an independent monetary policy, however, as it maintains restrictions on capital flows. In China's case, world and domestic interest rates can differ, because controls on the transfer of funds in and out of the country limit the resulting changes in the money supply and the corresponding pressures on the exchange rate.

As these three examples show, if a country chooses to allow capital to flow freely, it must also decide between having an independent monetary policy or a fixed exchange rate. Many factors affect how a country makes this crucial decision (Box 13-2).

Box 13-2: Choosing Among a Fixed Exchange Rate, Independent Monetary Policy, and Free Capital Movements

How does a country choose whether to give up a fixed exchange rate, independent monetary policy, or free capital movements? While country-specific factors play a role, experience has shown that these decisions also reflect global trends.

In the late 1920s, many countries, including the United States, adopted an exchange-rate system in which they pegged their currencies to a fixed quantity of gold. This system, which was used previously but was abandoned during World War I, was known as the *gold standard*. It effectively fixed the exchange rates of the currencies for all participating countries. Countries generally coupled this fixed exchange rate with the free movement of capital, relinquishing the ability to influence economic activity at home through the use of independent monetary policy.

This system proved sustainable until the Great Depression of the 1930s, when many governments abandoned exchange-rate stability in order to expand domestic demand by increasing the money supply and lowering interest rates. Following the economic recoveries under this regime, the choice of free capital flows and independent monetary policy remained popular through the end of World War II.

The postwar era, however, saw substantial international integration of markets and increasing cross-border trade. Countries such as the United States wanted to facilitate this increase in trade by eliminating the risks of exchange-rate fluctuations. At a summit held in Bretton Woods, New Hampshire, in 1944, representatives from the major industrial economies designed and implemented a plan that encouraged exchange-rate stability while maintaining autonomous monetary policies. The Bretton Woods system, as it became known, offered countries greater monetary independence while fixing the value of the dollar, yen, deutsche mark, and other currencies. Just as with the previous systems, however, something had to be sacrificed—the Bretton Woods arrangement required capital controls. Capital controls included caps on the interest rates that banks could offer to depositors and limitations on the types of assets in which banks could invest. Further, governments frequently intervened in financial markets to direct capital toward strategic domestic sectors. Though none of these controls alone prevented international capital flows, in combination they allowed governments to restrain the amount of cross-border capital transactions.

In the early 1970s, the Bretton Woods system gave way to a more-diverse set of regimes. Ultimately, as growth in other countries outstripped growth in the United States, demand shifted from the U.S. dollar to foreign currencies, putting downward pressure on the dollar's value. After several negotiated devaluations of the dollar, governments agreed to abandon the system rather than continue to be forced to change domestic interest and inflation rates to keep the dollar's value constant. Furthermore, greater financial sophistication and increasing capital mobility made it more difficult and costly to sustain capital controls in the advanced economies.

Since the end of the Bretton Woods system, countries have chosen a variety of exchange-rate regimes. Countries in the euro zone, for instance, have adopted the euro as a common currency. This is equivalent to fixing the exchange rates among the participating countries. The euro, however, is allowed to move freely against other currencies such as the dollar. Each of the countries within the euro zone has had to give up its own independent monetary policy. The value of the U.S. dollar, on the other hand, floats freely against other currencies. The free movement of capital has been uniformly embraced by the advanced industrial economies and is increasingly being adopted by developing economies.

Encouraging Free Capital Flows

The Administration supports the free flow of capital between the United States and other countries and encourages countries to take steps to open their markets to international investment. Such efforts include the negotiation of Bilateral Investment Treaties, as well as Trade and Investment Framework Agreements. Under these agreements, foreign countries commit to treating U.S. investors fairly and to allowing U.S. corporations to operate in foreign countries in closer accordance with standard U.S. practices and procedures. This protection reduces the risks associated with investing abroad and encourages U.S. multinational companies to expand through foreign direct investment.

Investment measures and protections have also played a central role in free trade agreements negotiated by the United States (these are discussed in Chapter 12, *International Trade and Cooperation*). Recent trade agreements, such as that with Chile, have included investment provisions that protect American investors and ensure their access to foreign investment opportunities.

The United States also encourages countries to undertake the reforms that will help them best reap the benefits of greater investment and capital flows. These reforms include improvements in corporate governance and the distribution of accurate, timely, and complete information on economic conditions, government regulations, and corporate performance. The Administration has focused on reducing the risks of destabilizing capital flows in a number of ways.

One important development in this regard has been the increased inclusion of "collective action clauses" in international bonds issued by emerging market countries—a practice that has been supported and encouraged by the United States. These clauses allow a majority of creditors to bind a minority to key financial terms in the event of a debt restructuring. They also help facilitate ongoing discussions and negotiations between a sovereign and its creditors. By making it easier for issuers and bondholders to agree to changes in bond terms in the event of a default or restructuring, collective action clauses provide a contractual method for improving the resolution of situations where sovereign debt levels are unsustainable. Such improvements to the debt-resolution process should reduce the unnecessary loss of value to creditors and thereby lessen the risk of lending to emerging market countries.

The United States has also endorsed the efforts of the International Monetary Fund and the World Bank to increase the availability, frequency, scope, and quality of the reported data of their member countries. Better and more timely information can assist policy makers and investors to make appropriate decisions. Some of these efforts include:

- The Financial Sector Assessment Program, which involves a rigorous and in-depth analysis of a country's financial system.
- The Special Data Dissemination Standard, which sets certain standards of timeliness and quality for economic and financial statistics to guide countries that have (or desire) access to foreign capital markets.
- The implementation of agreed-upon norms, such as the Code of Good Practices and Fiscal Transparency, which emphasize adherence to certain standards of good practice and promote quality accounting procedures and fiscal transparency.

These programs help investors, public-sector lenders, and governments identify weaknesses and vulnerabilities in firms, sectors, and the economy in general. They also target areas for reform in a country's macroeconomic policy, financial sector, and supervisory systems. This combination of policies should help developed and developing countries take advantage of greater capital market integration, while minimizing the risks.

Finally, the Millennium Challenge Account, a Presidential initiative enacted in January 2004, provides incentives for developing countries to adopt policies that spur economic growth and reduce poverty. First-year funding for the

Millennium Challenge Account is $1 billion. The Administration has requested that this amount rise to $5 billion per year by fiscal year 2006. The Millennium Challenge Corporation, which administers the Millennium Challenge Account, will direct development grants to poor countries that have appropriate economic, political, and structural conditions to benefit from foreign assistance. The Millennium Challenge Corporation will partner with countries that demonstrate a strong commitment to ruling justly, investing in their people, and encouraging economic freedom in order to develop their own strategies for catalyzing economic growth and reducing poverty. The Millennium Challenge Account is designed to provide funding for programs that have clear objectives, a sound financial plan, and measured benchmarks for demonstrating progress in overcoming major obstacles to sustained economic growth. The Millennium Challenge Account will not only improve the ability of recipient countries to fight poverty and to grow more quickly, but will also encourage the international investment that helps to strengthen growth.

Conclusion

Underlying each of the policies promoted by the Administration is the goal of helping countries reap the substantial benefits of the free flow of international capital. Foreign direct investment can facilitate the transfer of technology, allow for the development of markets and products, and improve a country's infrastructure. Portfolio flows can reduce the cost of capital, improve competitiveness, and increase investment opportunities. Bank flows can strengthen domestic financial institutions, improve financial intermediation, and reduce vulnerability to crises. These flows are not without their risks, but such risks can be reduced if countries adopt prudent fiscal and monetary policies, strengthen financial and corporate institutions, and develop the regulations and agencies that supervise such institutions. Such steps allow countries to fully gain from free capital flows.

The Link Between Trade
and Capital Flows

Movements of goods and services across borders are often thought of as distinct from international capital flows. For example, an individual who allocates part of his or her retirement savings to a mutual fund that invests in an international portfolio might not think that this cross-border transaction has an impact on the price of imports, such as foreign cars or food at the supermarket. Yet, for important but subtle reasons, trade flows and capital flows are closely intertwined—indeed, they are two sides of the same coin.

This chapter explores the linkages between trade and capital flows. The key points in this chapter are:

- Changes in a country's net international trade in goods and services, captured by the current account, must be reflected in equal and opposite changes in its net capital flows with the rest of the world.
- The United States has experienced a large net inflow of foreign capital in recent years. Any such inflow must be accompanied by an equally large current account deficit.
- The size and movement of current and capital accounts reflect fundamental economic forces, including saving and investment rates, and relative rates of growth across countries.

The Basic Accounting Identity

The *balance of payments* is the accounting system by which countries report data on their international borrowing and lending, as well as on the flow of goods and services in and out of the country. The balance of payments includes a number of different accounts (Box 14-1). The central relationship of the balance of payments is that the *net* flow of capital into a country, as measured by the financial and capital accounts, must balance the *net* flow of goods, services, transfer payments, and income receipts out of the country, as measured by the current account.

When the current account balance is negative, this means that purchases of foreign goods and services (and other outflows) exceed sales of goods and services to foreigners (and other inflows). This situation is referred to as a *current account deficit*. The *trade balance* is generally the largest component of the current account and captures the net inflows of goods and services. A

positive net flow of capital into the United States means that foreigners are purchasing more U.S. assets than U.S. citizens are purchasing foreign assets. According to the balance of payments, a positive net flow of capital into the United States must be balanced by a current account deficit.
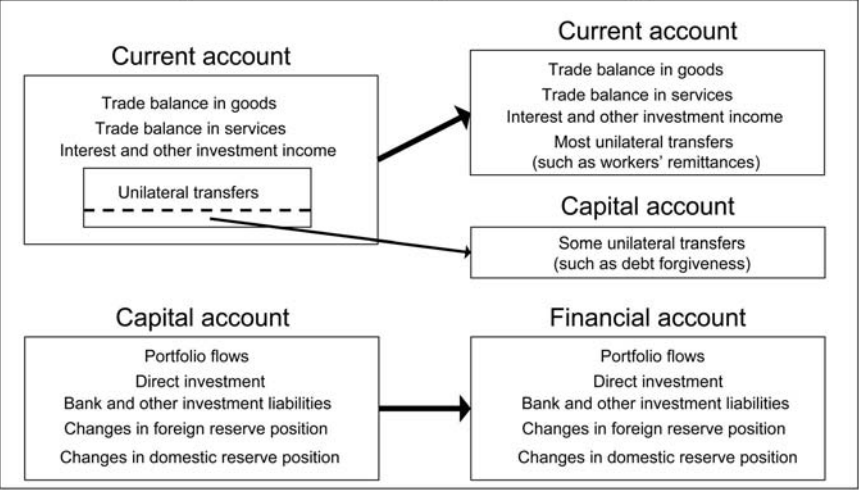

Box 14-1: A New Look for the Balance of Payments

Just as a country's national accounts keep track of macroeconomic variables such as GDP, saving, and investment, a country's balance of payments accounts serve as the bookkeeping for its international transactions, such as exports, imports, and international investment flows. In 1999, the Bureau of Economic Analysis announced that it would adopt new terminology to be consistent with international best practices for balance of payments accounting, as outlined by the International Monetary Fund.

The *old* balance of payments system used two accounts: the *capital account* and the *current account*. The *new* system uses three accounts (Chart 14-1). The new current account includes the trade balance in goods and services, net income receipts, and the balance of most unilateral transfers (one-way transfers of assets, such as pension payments to foreign residents). Some unilateral transfers, including debt forgiveness and the transfer of bank accounts by foreign citizens when immigrating to the United States, have been removed from the *old* current account and are now in a separate account, the *new* capital account. The new capital account represents a very small portion of overall capital flows. Private capital flows and changes in foreign and domestic reserves (formerly in the old capital account) are now in the *financial account*. This new treatment preserves the balance of payments identity that the sum of all the accounts is zero.

To simplify terminology, this *Economic Report of the President* refers to the new capital and financial accounts as *net capital flows*— that is, inflows of capital from foreign countries minus outflows from the United States. Positive net capital flows indicate that more capital is flowing into the United States than out.

Chart 14-1 Changes to the Balance of Payments Terminology in 1999

Current account
Trade balance in goods
Trade balance in services
Interest and other investment income
Unilateral transfers

→

Current account
Trade balance in goods
Trade balance in services
Interest and other investment income
Most unilateral transfers (such as workers' remittances)

Capital account
Some unilateral transfers (such as debt forgiveness)

Capital account
Portfolio flows
Direct investment
Bank and other investment liabilities
Changes in foreign reserve position
Changes in domestic reserve position

→

Financial account
Portfolio flows
Direct investment
Bank and other investment liabilities
Changes in foreign reserve position
Changes in domestic reserve position

To understand how the balance of payments works in practice, consider a consumer in the United States who purchases a scarf from a foreign seller for one dollar. This transaction is recorded as an import and reduces the U.S. current account balance by one dollar. The foreign seller could spend the dollar on U.S. goods or on U.S assets, such as stocks or bonds. If the foreigner purchases U.S. goods, this would be recorded in the balance of payments as a U.S. export in the current account. The U.S. purchase of the foreign scarf and the foreign purchase of U.S. goods would cancel each other out, so there would be no change in the current account and no change in net capital flows. Alternatively, if the foreigner decided to purchase U.S. assets, this would be recorded as a capital inflow into the United States. The increase in net capital flows would balance the decrease in the U.S. current account. In both examples, the resulting change in the current account, if any, exactly balances any change in net capital flows.

Trade in goods can lead to changes in financial balances (such as with the payment for the scarf in the example above), or financial transactions can lead to changes in trade balances. The latter case would occur if a foreigner purchased a U.S. asset, such as a bond, and the American seller of the bond used the proceeds to purchase foreign goods. In both cases, the balance between the current account and net capital flows still holds.

To understand how financial flows can affect trade balances, suppose that at the prevailing rate of return, investors in the United States seek to undertake $200 billion worth of projects. If U.S. savers were willing to provide only $150 billion in capital through saving, then the other $50 billion could come from the rest of the world as $50 billion in capital

inflows. If the U.S. investors choose to spend this capital inflow on foreign goods (perhaps imports of new computers), then net purchases of foreign goods would increase by $50 billion. The resulting $50 billion current account deficit would balance the $50 billion capital inflow. If investors in the United States were not able to obtain the initial $50 billion from abroad, both net capital flows and the current account would equal zero. There would be no current account deficit. Would this be good or bad? One immediate effect would be that the $50 billion gap between desired investment and saving would need to be closed by scaling back investment projects or raising national saving. These changes should be evaluated on their own merits; there is nothing particularly beneficial about having a trade balance or net capital flows exactly equal to zero.

A country's saving and investment decisions are critical to evaluating the implication of any given level of its current account balance. In a world without capital flows, the only funds available for investment come from domestic saving. Capital flows allow a country to finance higher levels of investment by drawing on funds from abroad. This net inflow of funds corresponds to greater net purchases from the world and a decline in the current account balance.

The desirability of positive net capital flows and a current account deficit depend on what the capital inflows are used for. Household borrowing—an excess of household spending or investment over saving—provides a useful analogy. Household debt could reflect borrowing to finance an extravagant vacation, a mortgage to buy a home, or a loan to finance education. Without knowing its purpose, the appropriateness of the borrowing cannot be judged. Similarly for countries, borrowing from abroad can be productive or unproductive. Borrowing from abroad can be justified if it raises the potential output of the economy and this, in turn, generates the resources needed to repay the foreign lenders.

This entire discussion has focused on trade balances and net capital flows with the world as a whole, and not with any individual country. There is no economic basis for concern about trade deficits and the corresponding net capital flows with an individual trading partner when there are many countries in the world (Box 14-2).

Box 14-2: Bilateral Versus Multilateral Balances

A country's aggregate trade deficit matters only to the extent that it reveals information about underlying economic forces, such as relative international growth rates or national saving and investment patterns. In contrast, *bilateral deficits*, such as the U.S. trade deficit with China, reveal *nothing* about underlying economic forces in either country. While trade barriers are a cause for concern, there is no economic sense in which a bilateral deficit is either good or bad. It would be an extraordinary coincidence if all countries had balanced trade with each of their partners. One of the benefits of the international financial system is that it frees countries from these bilateral constraints.

For example, imagine a simplified world that consisted of only the United States, Australia, and China. Suppose the United States ships $100 billion of machine tools to Australia and imports no goods in return. Australia ships $100 billion of wheat to China with no reciprocal goods imports, and China ships $100 billion of toys to the United States. Each country would have $100 billion of exports and $100 billion of imports, so that each would have balanced trade overall. Yet some Americans might complain about their bilateral deficit with China. Some Chinese might complain about their deficit with Australia, and some Australians about their deficit with the United States. All of these complaints would be unfounded; bilateral deficits and surpluses are a natural consequence of a trading world composed of many countries.

Domestic transactions provide a useful analogy. A plumber who spends no more than he earns can still run a bilateral deficit with the local grocer. The plumber can earn money from other sources to pay the grocer and is not constrained to buying only from grocers who have plumbing problems. The bilateral imbalance that exists between the plumber and the grocer is an entirely natural feature of a well-functioning economy with a strong payments system and specialization.

Trends in the U.S. Balance of Payments

The decrease in the U.S. current account balance, from nearly zero in the early 1990s to a deficit of about 5 percent of GDP in the first three quarters of 2003, has been mirrored by a similar increase in net capital flows (Chart 14-2). (The two series in Chart 14-2 are not exact mirror images due to imprecision in the measurement of trade and capital flows.)

Examining the components of the current and financial accounts provides information on the causes of these recent trends in the U.S. balance of payments (Table 14-1). Over the 1990s, a major contributor to the rise in the current account deficit was the increase in imports of foreign goods. The trade balance in goods moved from a deficit of 1.9 percent of GDP in 1990 to a deficit of 4.6 percent of GDP in 2000. Exports of goods increased from 6.7 percent of GDP in 1990 to 7.9 percent of GDP in 2000, but goods imports increased by much more, from 8.6 percent of GDP to 12.5 percent of GDP over the same period. The increase in the current account deficit since 2000 has resulted mainly from lower exports of goods (which fell from 7.9 percent to 6.4 percent of GDP between 2000 and the first three quarters of 2003), rather than increased imports. Imports as a share of GDP actually fell 1 percentage point over the same period (Chart 14-3 and Table 14-1). Most recently, the current account deficit has narrowed from 5.2 percent of GDP in the first quarter of 2003 to 4.9 percent of GDP in the third, reflecting stronger export growth.

Chart 14-2 **Balance of Payments**
The 1990s saw a surge in capital inflows and a corresponding deficit in the current account balance.
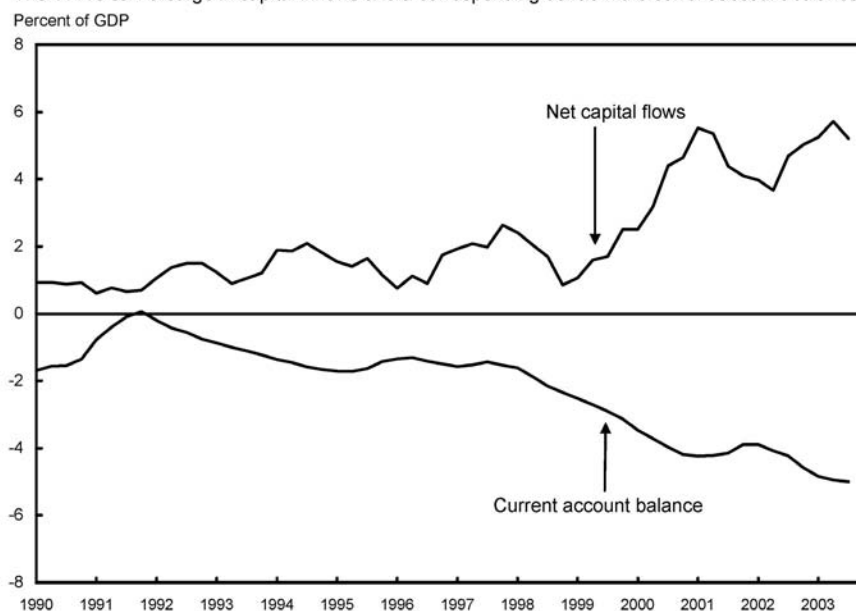Percent of GDP

Note: Data are four-quarter moving totals.
Source: Department of Commerce (Bureau of Economic Analysis).

Table 14-1.— *Current and Financial Account*
[Percent of GDP]

Accounts	1990	2000	2002	2003: Q1-Q3
Current account balance	-1.4	-4.2	-4.6	-5.1
Trade balance in goods	-1.9	-4.6	-4.6	-5.0
Exports	6.7	7.9	6.5	6.4
Imports	-8.6	-12.5	-11.1	-11.5
Services (net)	.5	.8	.6	.5
Other (net)	.0	-.4	-.6	-.6
Net capital flows	.9	4.6	5.0	5.0
Financial account balance	1.0	4.6	5.0	5.1
Direct investment (net)	.2	1.7	-.9	-.5
Portfolio (net)	-.1	3.0	4.2	3.7
Equity securities (net)	-.4	.9	.3	-.8
Debt securities (net)	.3	2.2	3.8	4.5
Other	1.0	.0	1.8	1.8
Capital account balance	-.1	.0	.0	.0
Memo:				
Foreign purchases of U.S. Government securities	.5	-.5	1.6	2.6

Note: Detail may not add to totals because of rounding and seasonal adjustment.

Source: Department of Commerce (Bureau of Economic Analysis).

Chart 14-3 **Exports and Imports of Goods**
Both exports and imports of goods decreased substantially starting in the first quarter of 2001 and have yet to fully recover.

Percent of GDP

Source: Department of Commerce (Bureau of Economic Analysis).

U.S. net capital flows grew from about 1 percent of GDP in 1990 to over 4½ percent of GDP in 2000. This resulted from roughly equal increases in foreign purchases of debt securities, equity securities, and direct investment. This increase in net capital flows into the United States largely reflected the desire of foreigners to participate in higher-return investment opportunities in the United States. The global economic downturn and the collapse of high-tech stock prices and broader equity indices that began in 2000 contributed to a shift in the composition of capital flows in the United States. Foreign investors moved away from foreign direct investment and private equity assets and toward government and corporate bonds. In addition, foreign governments increased their share of these capital flows, although the foreign private sector still accounts for a far greater proportion.

Over the latter half of the 1990s and the early 2000s, the counterpart to the rising U.S. current account deficit has been a growing wedge between U.S. investment rates and U.S. national saving rates (Chart 14-4). The national saving rate in the United States began to decline in 1999, but increased capital inflows allowed U.S. investment rates to remain at a high level through 2000. As discussed in Chapter 1, *Lessons from the Recent Business Cycle*, investment fell substantially after the collapse of the stock market bubble of the late 1990s. In 2001, the decline in investment outpaced a contemporaneous decline in U.S. saving, so that the current account deficit narrowed. U.S. investment has since leveled off while saving remains low, causing a wider U.S. current account deficit. Over the entire period, the availability of foreign investment permitted the United States to maintain higher investment rates than it could have funded relying solely on domestic financing. These capital inflows have helped finance U.S. investments, expand U.S. productive capacity, and strengthen U.S. economic performance.

## Factors that Influence the Balance of Payments

A number of underlying economic factors influence the level of and changes in the balance of payments. One of the most important factors is the differential rate of GDP growth across countries. During the late 1990s, the United States grew faster than many of its major trading partners, such as Japan and a number of major European countries. As a result, capital flowed into the United States, leading to a corresponding trade deficit. Even during the recent business-cycle downturn and recovery, U.S. growth rates have exceeded those of many of our major trading partners. This has contributed to the slow recovery in U.S. exports and has helped to maintain continued capital inflows into the United States.

Chart 14-4 Saving, Investment, and the Current Account Balance

In the late 1980s, a decline in the investment rate led to a reduction in the current account deficit, while a sharp decline in the national saving rate accounted for the most recent expansion of the deficit.

Percent of GDP
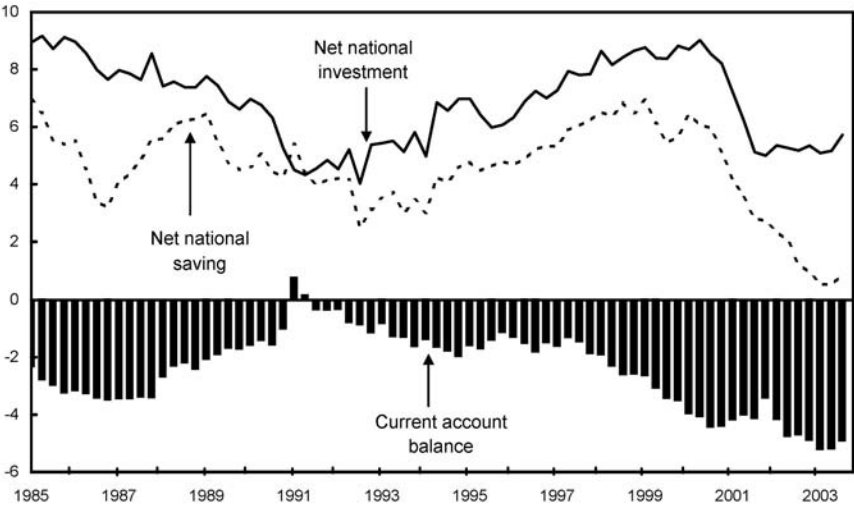
Source: Department of Commerce (Bureau of Economic Analysis).

A second determinant of trade and capital flows is the price of domestic goods relative to foreign goods. Relative prices are influenced by a number of factors, including labor and production costs, labor productivity, and exchange rates. For many manufactured products, for example, labor and production costs in developing countries are often below such costs in the United States. As a result, the prices of these goods produced in developing countries may be substantially lower than the price of similar goods produced in the United States. For other products and projects, such as airplanes and the development of new drugs, the availability of factors of production such as skilled engineers may be more important than the availability of low-skilled workers. Exchange rates can also influence relative prices. A depreciation of a country's currency can make its products cheaper and thus more competitive abroad, even if domestic prices do not change. When a country's currency appreciates, domestically produced goods become relatively more expensive in foreign markets.

A third determinant of the direction and size of capital flows is the relative return that investors expect to make in one country compared with another. This return differential can reflect factors discussed earlier, such as relative output growth, labor costs, or exchange rates. This differential can also

depend on a country's legal framework, accounting and tax systems, infrastructure, culture, and institutions. The flow of capital into the United States likely reflects a view that the expected risk-adjusted, after-tax return on U.S. assets is higher than the return on similar foreign investments.

These factors—growth rates, relative prices, and rates of return—all drive national saving and investment decisions. Those decisions most directly determine the balance of payments. *National saving* is the sum of *private saving* (saving of households and corporations) and *public saving* (the total saving of Federal, State, and local governments, as reflected in their budget balances). When national saving is less than domestic investment, a country must be borrowing from abroad. This borrowing will be reflected in positive net capital flows and a current account deficit.

Although this suggests that the recent increase in the U.S. budget deficit may be related to the recent increase in the U.S. current account deficit, the historical evidence for a relationship between government deficits and trade deficits is mixed. A number of academic studies suggest that other domestic and international factors are more important influences on current account balances than government deficits. The recent U.S. experience supports this. In the 1990s, the large increase in the U.S. current account deficit occurred while the Federal budget surplus was growing (Chart 14-5). From 1997 to 2000, the U.S. current account deficit increased by almost 3 percentage points. Over the same period, the U.S. budget balance went from a slight deficit to a surplus of 2½ percent of GDP. Since 2000, the U.S. budget has moved into deficit by several percentage points of GDP, but the current account deficit has widened by only about 1 percentage point of GDP. These figures show that the current account and Federal budget do not move in lockstep, and that the government deficit is only one of several factors behind the widening of the current account deficit since the mid-1990s.

Chart 14-5 **Budget Deficit and the Current Account Balance**
The relatively steady decline in the current account balance contrasts with the initial reduction and subsequent expansion of the budget deficit.

Percent of GDP
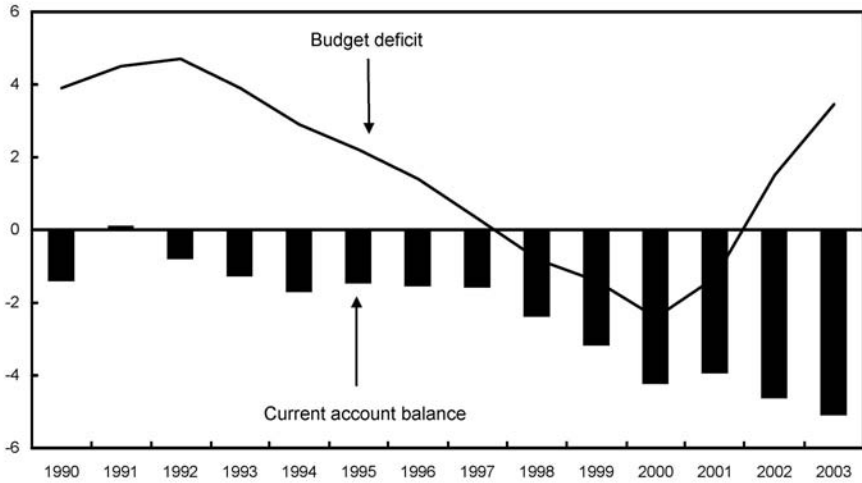
Note: Budget deficit data are for fiscal years ending September 30; current account data are for calendar years. Current account balance for 2003 includes data through the third quarter.
Source: Department of Commerce (Bureau of Economic Analysis).

Possible Paths of Balance of Payments Adjustment

The U.S. current account deficit reached about 5 percent of GDP in the first three quarters of 2003. Historically, many countries with sizable current account deficits have experienced reductions in capital flows and corresponding reductions in their current account deficits. Because the U.S. current account deficit and U.S. capital inflows are balanced by trade and capital flows in other countries, any change in the U.S. balance of payments would involve corresponding changes in other countries' flows of trade and capital. The economic implications of any adjustments depend on how it occurs.

An adjustment in the U.S. trade balance could involve a number of domestic and global factors. For example, faster growth in other countries would be expected to increase demand for U.S. exports and narrow the U.S. current account deficit. Slower growth in the United States relative to its major trading partners would dampen U.S. demand for imports and reduce the U.S. trade deficit. Trade flows could also adjust through changes in the relative prices of U.S. goods and services compared to the prices of foreign goods and services. This relative-price adjustment could occur through changes in nominal exchange rates or through different inflation rates in different countries.

An adjustment in the U.S. balance of payments would also require a change in international capital flows. To reduce net capital flows, foreign investors could buy fewer U.S. assets and/or U.S. investors could buy more foreign assets. This might occur if U.S. national saving were to increase, thereby reducing the need for foreign funds to finance U.S. domestic investment. The U.S. investment rate could also fall, so that the United States required less capital inflow. Lower investment is the least desirable form of adjustment for the balance of payments, however, as it would reduce U.S. productive capacity and lead to slower growth.

It is impossible to predict the exact timing or magnitude of any adjustment in the U.S. current account balance. After a large increase in the U.S. current account deficit in the 1980s, the ensuing adjustments were gradual and benign. Public policies can facilitate changes in the U.S. current account and net capital flows by creating a stable macroeconomic and financial environment, encouraging foreign growth, and spurring increased saving in the United States.

Conclusion

Flows of goods and services across borders are linked to international capital flows through the balance of payments. Changes in the current account (which includes international trade in goods and services) must be balanced by equal and opposite changes in net capital flows with the rest of the world. Similarly, movements in net capital flows require offsetting movements in the current account.

In recent years, the United States has received large net inflows of foreign capital, which have been balanced by large U.S. current account deficits. The U.S. balance of payments is mirrored by trade and capital flows in other countries. Thus, over the same period, the rest of the world as a whole has experienced a current account surplus and capital outflows.

The United States' sizable positive net capital flows and the corresponding trade deficits are neither good nor bad in and of themselves. Instead, they represent underlying economic forces, such as relative GDP growth rates, relative prices of domestic and foreign goods, relative returns on investment, and national saving and investment decisions. Changes in these underlying factors would lead to changes in the U.S. balance of payments and corresponding changes in the international flows of trade and capital.