Flation

William Poole and Robert H. Rasche

"Flation"—not inflation, not deflation—is lifted from the title of a book by Abba P. Lerner.[1] For the past 35 years in the United States and, indeed, in most of the world, policymakers and the public in general have been focused on the issue of *inflation*—that is, the continual upward drift in prices of the overwhelming fraction of goods and services produced in the economy. Sometimes the drift was more of a gallop. For most of this period, the upward trend was also characteristic of the prices of assets such as land, houses, and equities. Inflation, prevalent though it has been in our recent economic experience, has not been the norm for most of U.S. history. In the early 1930s, exactly the opposite experience occurred: *deflation*, or a continual downward drift in the prices of goods, services, and assets.

Deflation has a frightening history. Simultaneously with the deflation of the early 1930s, the U.S. unemployment rate soared to about 25 percent in 1933 at the depth of the Great Depression. Although deflation ended in 1933, the damage to the economy was so great that poor economic conditions persisted until the United States became involved in World War II in 1941. Moreover, the economic history of the 1990s in Japan is characterized by deflation. The Japanese economy has stagnated, and unemployment there has risen today to levels not seen in over 40 years. From these and other episodes around the world, many people associate deflation with "hard times."

The purpose of this analysis is not to get into a discussion of whether a little deflation is compatible with prosperity, although within limits it may be. The more important point is that, without question, substantial deflation is inconsistent with prosperity. Thus, deflation is every bit as serious an issue as inflation; however, the U.S. economy today does not run any significant risk of deflation.

Obviously, not everyone agrees with this judgment. Based on a few recent observations of month-to-month price changes, some commentators have used the "D" word to express their concern about the current state of the U.S. economy. The objective in this paper is to explore this subject and, we hope, make a contribution to public understanding of the issue.

First, we will elaborate on what we believe is the appropriate objective for Federal Reserve policy. Second, we will explain the generally accepted definitions of inflation and deflation, and discuss the fundamental sources of these phenomena. Third, we will review aspects of price behavior in our economy and discuss how data should be interpreted to determine the inflationary or deflationary state of the economy. Finally, although the issue concerns the behavior of the aggregate price level, we will examine some particular sectoral price changes to help better understand the aggregate price level.

THE APPROPRIATE POLICY OBJECTIVE FOR THE FEDERAL RESERVE

Our monetary policy framework is this. First and foremost, the central bank must maintain a commitment to price stability. An operational definition of price stability is an environment in which the inflation rate, properly measured and averaged over several years, is zero. All of our inflation data are subject to measurement errors. Experts in such measurements generally agree that current price indexes, despite statisticians' best efforts, still leave inflation measures that have some upward bias. Hence, in terms of the various inflation indexes, we can say that price stability prevails when broad price indexes exhibit small positive average values for measured inflation and that year-to-year fluctuations around that average are well contained.

If the price level comes unstuck, yielding inflation or deflation, all sorts of other problems will arise. Nevertheless, within the confines of the goal of price stability, the central bank has some flexibility to lean against fluctuations in output and employment. However, the central bank ought not to pursue the goal of stabilizing economic activity so aggressively that it runs any substantial risk of compromising the goal of price stability.

Finally, in leaning against fluctuations in growth and employment, the central bank ought not to

William Poole is the president and Robert H. Rasche is a senior vice president and director of research at the Federal Reserve Bank of St. Louis. This article is adapted from a speech of the same title presented before the International Mass Retail Association Leadership Forum, January 21, 2002. The authors appreciate comments provided by colleagues at the Federal Reserve Bank of St. Louis. The views expressed do not necessarily reflect official positions of the Federal Reserve System.

[1] Lerner, Abba P. *Flation*. New York: Quadrangle Books, 1972.

have goals for levels of the economy's growth and unemployment rates per se. Within a wide range, no one knows what the economy's equilibrium rate of growth is or what rate of unemployment will clear the labor market in the long run. The biggest risks of a major monetary policy mistake occur if a central bank attempts to target the levels of real variables.

Achieving the objective of price stability, as defined above, will yield a highly stable economy. When the market has confidence in Fed policy, short-run changes—that is, over a few months or even a few quarters—in the rate of inflation or deflation will tend to be self-reversing rather than self-reinforcing.

THE DEFINITION AND SOURCES OF INFLATION AND DEFLATION

At the beginning of the great inflation of 1965-80, there was a wide disparity of professional opinion about the fundamental source of inflation or deflation in an economy. One proposition came to be known as the "monetarist view." This view held that sustained inflation or deflation was always a monetary phenomenon; that is, that the only source of long-run positive or negative trends in the general level of prices in an economy is the creation of an excess or insufficient supply of money balances relative to the growth of the productive capacity of that economy. Milton Friedman of the University of Chicago was the most publicly visible proponent of this proposition. The Federal Reserve Bank of St. Louis, in particular the president of the Bank at that time, Darryl Francis, and the Research staff were vocal advocates of this proposition in the policy arena during the late 1960s and early 1970s. A reading of the Memorandum of Discussion of the Federal Open Market Committee (FOMC) for this period makes clear that there were sharp debates over these issues. The FOMC is the Fed's main monetary policy-making body, and the public record of that period shows that Darryl Francis was a vigorous advocate of the monetarist view.

The proposition that the central bank is the source of ongoing inflation or deflation was a distinct minority view 35 years ago. In the FOMC, Darryl Francis was usually the only one expressing this view. The development of economic theory and the economic history of the past three decades have produced a major change in both professional thinking and public attitudes toward the sources of inflation and deflation. Economists are now largely in agreement that if the central bank does not achieve the goal of price stability, no one else can. Many central banks around the world, starting with the Reserve Bank of New Zealand in 1990, have acknowledged this responsibility and have adopted explicit numeric inflation targets.

This view also spread into public thinking about inflation in the United States. Paul Volcker, former chairman of the Board of Governors of the Federal Reserve System, is widely credited for the disinflation that occurred in the United States in the early 1980s. Chairman Greenspan is applauded for the additional progress in the 1990s that brought the U.S. inflation rate to the lowest level in almost 40 years.

Today the Federal Reserve accepts its responsibility for the trend rate of inflation. However, a central bank is not responsible for month-to-month wiggles in the inflation statistics. Nor should a central bank attempt to react to short-run variations, since the sources of such noise are beyond its control and likely to average out over a period of a few months or at most a couple of years. One obvious reason for not reacting to short-run developments is that an unknown part of these changes in the reported inflation rate is purely measurement error, or statistical noise.

Professional opinion has also changed about the source of deflation in the 1930s. It is now widely acknowledged that, at a minimum, the intensity of the Great Depression was magnified by the failure of the Federal Reserve to provide sufficient liquidity to the economy in the face of widespread bank failures. The Federal Reserve in turn has learned from that experience. When the U.S. economy has been threatened by liquidity crises in recent years—such as the stock market crash of 1987, the Asian crises and Russian default of 1998, and the terrorist attack of September 11, 2001—the Fed has moved rapidly to inject large amounts of liquidity into the economy. Liquidity crises have been averted, inflation has remained low and stable, and deflation has not occurred.

Experience elsewhere has not been as benign. Over the period from 1981 through 1990, the Japanese economy grew at an annual rate of 3.7 percent and the inflation rate (measured by the gross domestic product [GDP] price index) averaged 1.5 percent per year. The situation in Japan in the 1990s has been remarkably different. The Japanese economy has struggled in and out of recession, and real growth from 1991 to 2000 averaged only 1.1 percent. Over the same period, very low inflation has turned into deflation. From 1991 to 1996, the Japanese consumption deflator rose at an average annual rate of only 0.5 percent; for 1996 to 2000, the rate was –0.2 percent. Asset prices fell dramatically. The decline of the Nikkei equity price index from a value of close to 40,000 in late 1989 to its recent level of less than 10,000 is common knowledge. What is not as well known outside Japan is that land and real estate prices over the past decade have experienced equally dramatic declines as those seen in equity markets. In April 1993 an index of housing prices in Japan stood at 42.35 million yen. By April 2001 it had fallen to 36.52 million yen, an annual average rate of decline of 1.7 percent.[2] The index of residential land prices reached a peak in March 1991 of 109.7 and fell to 81.7 by September 2001, an annual average rate of decline of 2.4 percent. The decline in commercial land prices was even larger. From a peak of 111.7 in September 1991, the index of these prices fell to 49.1 in September 2001, an annual average rate of decline of 5.6 percent.[3] In terms of the impact on Japan's output and employment, the large deflation of asset prices was probably more important than the gentle deflation of goods prices.

What is responsible for the incredible difference in the performance of the Japanese economy between the 1980s and 1990s? Japan's money stock (using Japan's own preferred measure, M2 + CDs) grew at an average annual rate of 7.9 percent from 1981 through 1990, but only at 2.3 percent per year over the decade from 1991 through 2000. A conclusion consistent with research on this issue is that the ongoing stagnation and deflation that the Japanese economy has experienced in the past decade is likely related to an insufficient supply of liquidity by the Bank of Japan. Slow money growth is not the whole story, but is certainly a significant part of it.

RECENT PRICE BEHAVIOR IN THE U.S. ECONOMY

Public discussion of inflation in the United States generally is focused on the consumer price index (CPI) published monthly by the Bureau of Labor Statistics. The monthly change in the overall CPI is the so-called "headline" inflation number. The CPI is very visible; it has been widely reported for years and is used to construct cost-of-living adjustments in union wage contracts and Social Security benefits.

Sometimes reference is also made to a "core" inflation rate, usually measured by the CPI excluding prices of food and energy products. The rationale for excluding food and energy prices is that they can be quite volatile, and hence longer-term inflation trends can be obscured when they are included.

Starting in 2000, the FOMC chose to focus on a different measure of inflation: changes in the price index for personal consumption expenditures in the national income accounts. This measure of inflation, which for convenience we will call the "consumption price index," is reported monthly by the Bureau of Economic Analysis of the Department of Commerce. Although this index receives less public attention than the CPI, it is preferred by the FOMC because the methodology used in its construction reduces the measurement bias relative to that in the CPI; also, the coverage of goods and services in this index is believed to better represent consumption patterns. For example, prices of medical services are included in the CPI only to the extent that such services are paid directly by consumers. Prices of all medical services are included in the consumption price index whether those services are paid for directly by consumers or are paid for on behalf of consumers by third parties such as insurance companies.

In recent years, inflation as measured by the consumption price index has been lower than that measured by the CPI.[4] Although the following discussion will refer primarily to the consumption price index, no important issues depend on whether the focus is on that index or the CPI.

What should we expect to observe in an economy where price stability prevails? If it were possible to measure the average level of prices with little or no bias in such an economy, then over a period of time an average measured inflation rate very close to zero should be observed. From month-to-month or quarter-to-quarter, positive or negative changes of the inflation index will occur, but over time these would average out to about zero.

[2] The Housing Loan Progress Association. "Price Survey of the Housing Market." < http://jin.jcic.or.jp/stat/stats/ > .

[3] National Land Agency.

[4] In August 2002 the Bureau of Labor Statistics introduced a new measure of consumer prices—the chained consumer price index for all urban consumers (C-CPI-U). Monthly data are available from December 1999. The objective of the new index is to reduce the substitution bias that is present in the CPI-U. Between December 1999 and December 2000 (the only period for which final estimates of the C-CPI-U are available), the inflation rate measured by the C-CPI-U differs from that measured by the consumption price index by only 0.1 percent.

What about prices of individual goods and services under such conditions? There would likely be a dispersion of changes in the prices of individual goods and services around zero. In fact, prices of some goods and services could be continually falling, while prices of other goods and services could be continually rising. It is perfectly normal to experience divergent trends of individual prices under conditions of overall price stability. Thus, trends in the prices of individual goods or services cannot be used to judge whether an economy is experiencing inflation or deflation.

An important influence on inflation data in the United States over the past three years has been the behavior of energy prices on world markets. In 1998, energy prices collapsed as world demand dropped dramatically in response to the crises in Asian economies. Petroleum inventories rose unexpectedly and major producers, including OPEC nations, cut production to stabilize prices and adjust inventories. In 1999 and 2000, energy prices rose sharply as economic activity boomed in the United States and other major industrialized economies at a time when world inventories of oil were particularly low. Leading up to 2002, as the U.S. economy sank into recession and economic growth slowed in Europe, energy demand growth slowed and energy prices on world markets fell again.

The average inflation rate over the four years 1994 through 1997 was 2.7 percent per year as measured by the consumption price index. The average inflation rate over the four years from 1998 through 2001 was 1.7 percent per year. The core inflation component of the consumption price index has fallen from 2.1 percent in the earlier period to 1.6 percent in the latter period. The conclusion from these observations is that there has been a small reduction in trend inflation, whether measured by the total or the core consumption price index, over the past four years.

No estimates of the biases in the index are so large as to suggest that the true rate of inflation is now negative—that is, the U.S. economy is not in a deflationary situation. What, then, is the origin of the "deflation threat" that has been featured in some economic and newspaper commentaries? Some of these discussions appear to concentrate unduly on particular prices and on short-run data collected in the immediate aftermath of the September 11 terrorist attacks. The change in the price index for personal consumption expenditures for September 2001 compared with August 2001 was reported at –0.4 percent. The decline is attributable to falling energy prices and to a statistical artifact of the decision made by the Bureau of Economic Analysis in measuring insurance claim payments as a result of the September 11 attacks. The December 2001 consumption price index showed a decline of 0.2 percent for the month and led to further press speculation about deflation. Again, it is necessary to emphasize that a focus on very short-term movements in the price indexes can lead to misinterpretation of the underlying trends of inflation or deflation in an economy.

CHANGES IN RELATIVE PRICES

One of the great strengths of the U.S. economy is that prices of individual goods and services fluctuate freely. These price changes allow markets to signal how our productive resources can be allocated most efficiently. The disparity among inflation rates for particular goods and services over longer periods of time is significant. From 1980 to 2000, the overall consumption price index rose 95 percent. Consider price behavior in a half-dozen categories within overall personal consumption expenditures. Prices of personal computers and peripheral equipment stand out: such prices are estimated to have fallen by 99 percent since 1980. Note that despite this dramatic price decline, people do not talk about the computer industry suffering from deflation. This is a growth industry, driven by dramatic innovations and increases in efficiency.

Prices of durable goods are estimated to have increased by 20 percent since 1980, considerably slower than the general inflation over this period. Prices of nondurables are estimated to have increased by 65 percent since 1980; nondurable goods prices have risen more than durable goods prices, but still considerably less than the overall rate of inflation. Prices of food and beverages are estimated to have increased 79 percent since 1980, somewhat slower than the overall rate of inflation.

Consider some examples at the other extreme. Since 1980, prices of tobacco and smoking products are estimated to have increased 480 percent and prices of medical services by 197 percent. In the tables that show prices by various sectors, wide differences in experience such as those mentioned here can be seen.

Are falling prices, or prices that increase slowly relative to the general rate of inflation, indicative of "hard times" for particular industries? Sometimes, but certainly not always. Consider personal computers and consumer electronics in general (the latter is included in the durable goods component of the consumption price index). These are goods that have demonstratively high income and price elasticities. What that means is that the amounts consumers buy increase a lot as incomes rise and/or prices fall. Over time, as consumer incomes have increased and prices have fallen, the size of the market for these high-elasticity products has increased dramatically. Color TVs, camcorders, VCRs, DVDs, and personal computers, to name a few such products, are all now common household items in the United States. Many consumers can remember when these products were either unknown or owned by relatively few households.

This is an important point: expansion of the markets for certain products occurred simultaneously with a fall in prices. Price deflation for these goods was not inconsistent with prosperity in the industries producing them. Indeed, declining prices were essential to expanding these markets. The fall in prices was the result of rapid productivity increases from innovations in the production of these items and/or their components. Firms found it profitable to cut prices and expand production. Workers in these industries found their improved productivity rewarded in higher wages. Consumers, workers, and shareholders all have benefited, even though prices have fallen substantially over time.

High-demand elasticities are a critical element in such success stories. In contrast, consider markets for basic agricultural products in the United States. Productivity improvement in U.S. agricultural production over the years has been tremendous. Prices of these products have also fallen relative to goods in general over the long run. However, both income and price elasticities for agricultural products are relatively low. Hence, economic growth and declining prices have not produced large increases in consumption. As a result, fewer and fewer workers have been required over time to produce more than enough output to satisfy both domestic and foreign demand. Farms have gone out of business, the number of people engaged in agricultural production has decreased, and in recent years farm income has been sustained by large "emergency" farm appropriations out of the federal budget. Because of the low price and income elasticities for agricultural goods, deflation in this industry means hard times for many farmers.

Health care provides a really interesting case of relative price changes. In part, the rapid rate of price increase here represents innovation in the form of new products and/or improved procedures. Such price changes really reflect significant quality improvements. Ideally such quality improvements would be incorporated into the measurement of a standardized unit of medical services. With some consumer durables, such as automobiles, statisticians have been quite successful in measuring quality improvement. In other areas, capturing quality change into the measurement of a standard unit of output is difficult if not impossible.

As an example, consider laparoscopic surgery to remove the gall bladder. Not that long ago, gall bladder surgery required a substantial period of hospitalization, during which patient activity levels were significantly restricted. Today, with laparoscopic surgery, the length of the hospital stay is much shorter and patient discomfort much less. Moreover, the patient can resume reasonably normal activity, including going to work, after a short postoperative period. The patient and/or a third-party payer may pay the surgeon substantially more today to remove the gall bladder than 35 years ago, but does this increase mean that the price properly measured is dramatically higher? A well-constructed price index might adjust for the reduction in the pecuniary cost of confinement—fewer hospital days—from the improved technology. However, it is unlikely that any price index would reflect the improved quality of the procedure represented by the reduced nonpecuniary costs of confinement and the shorter recovery time now available. Hence the reported change in the price index for such a procedure certainly overstates the true rate of price change.

FLATION AND THE FED

The Fed's goal is to maintain low and steady inflation, so that expectations of changes in inflation do not enter importantly in the decisions businesses and households make. Using several different measures of inflation expectations, it is clear that long-term expected inflation has changed little in recent years. There is no evidence that changing inflation expectations figure importantly in economic decisions at this time.

Substantial variability in prices of individual goods is consistent with stability in the overall inflation rate. The variability serves to allocate and reallocate resources across different sectors of the economy, according to changes in consumer tastes and differential trends in productivity advancement. Simply put, it is normal that some industries are growing while others are contracting.

A common business problem is to determine a successful pricing strategy. One aspect of pricing strategy is directly relevant to this discussion. When a firm cuts prices to stimulate sales, it may not be successful if its customers believe that even deeper price cuts are around the corner. An expectation of falling prices may, temporarily, reduce rather than increase sales. It is for this reason that generalized deflation can be so dangerous to the economy. A widespread expectation of falling prices may lead to declining demand across much of the economy as people wait for lower prices in the future. Declining demand may force layoffs, which further depress household and business confidence. Conversely, inflation expectations can lead to rising demands and anticipatory buying.

Many analysts seem to view low inflation and high employment as competing goals. That is certainly not the only possible scenario. Maintaining low and stable inflation contributes mightily to overall economic stability. Consider the situation in the weeks following the terrorist attacks of September 11, 2001, when the economic outlook was highly uncertain. The auto industry was successful in selling a record number of cars in October 2001 through price cuts in the form of zero-interest financing. If consumers had reacted by expecting even deeper price cuts and had delayed purchases, the situation in early 2002 would have been very different. Overall, consumers view price cuts in today's environment as a buying opportunity, not as a forecast of further price cuts to come.

Clearly, the stability in the overall price environment—stability in longer-run expectations—is what allows temporary price cuts to work to boost sales and is an important element in stabilizing the general economy. The current U.S. situation does not match cases in the United States and elsewhere that historically have been associated with ongoing deflation. The Federal Reserve pursued an expansionary monetary policy throughout 2001 that has contributed to restoring equilibrium to the U.S. economy. What policy actions will be appropriate going forward will have to be determined as evidence arrives on the strength and durability of the economic expansion. We must be vigilant, but today it is likely that we enjoy flation—no "in" and no "de."

A Case Study of a Currency Crisis: The Russian Default of 1998

Abbigail J. Chiodo and Michael T. Owyang

A currency crisis can be defined as a speculative attack on a country's currency that can result in a forced devaluation and possible debt default. One example of a currency crisis occurred in Russia in 1998 and led to the devaluation of the ruble and the default on public and private debt.[1] Currency crises such as Russia's are often thought to emerge from a variety of economic conditions, such as large deficits and low foreign reserves. They sometimes appear to be triggered by similar crises nearby, although the spillover from these contagious crises does not infect all neighboring economies—only those vulnerable to a crisis themselves.

In this paper, we examine the conditions under which an economy can become vulnerable to a currency crisis. We review three models of currency crises, paying particular attention to the events leading up to a speculative attack, including expectations of possible fiscal and monetary responses to impending crises. Specifically, we discuss the symptoms exhibited by Russia prior to the devaluation of the ruble. In addition, we review the measures that were undertaken to avoid the crisis and explain why those steps may have, in fact, hastened the devaluation.

The following section reviews the three generations of currency crisis models and summarizes the conditions under which a country becomes vulnerable to speculative attack. The third section examines the events preceding the Russian default of 1998 in the context of a currency crisis. The fourth section applies the aforementioned models to the Russian crisis.

CURRENCY CRISES: WHAT DOES MACROECONOMIC THEORY SUGGEST?

A currency crisis is defined as a speculative attack on country A's currency, brought about by agents attempting to alter their portfolio by buying another currency with the currency of country A.[2] This might occur because investors fear that the government will finance its high prospective deficit through seigniorage (printing money) or attempt to reduce its nonindexed debt (debt indexed to neither another currency nor inflation) through devaluation. A devaluation occurs when there is market pressure to increase the exchange rate (as measured by domestic currency over foreign currency) because the country either cannot or will not bear the cost of supporting its currency. In order to maintain a lower exchange rate peg, the central bank must buy up its currency with foreign reserves. If the central bank's foreign reserves are depleted, the government must allow the exchange rate to float up—a devaluation of the currency. This causes domestic goods and services to become cheaper relative to foreign goods and services. The devaluation associated with a successful speculative attack can cause a decrease in output, possible inflation, and a disruption in both domestic and foreign financial markets.[3]

The standard macroeconomic framework applied by Fleming (1962) and Mundell (1963) to international issues is unable to explain currency crises. In this framework with perfect capital mobility, a fixed exchange rate regime results in capital flight when the central bank lowers interest rates and results in capital inflows when the central bank raises interest rates. Consequently, the efforts of the monetary authority to change the interest rate are undone by the private sector. In a flexible exchange rate regime, the central bank does not intervene in the foreign exchange market and all balance of payment surpluses or deficits must be financed by private capital outflows or inflows, respectively.

The need to explain the symptoms and remedies of a currency crisis has spawned a number of models designed to incorporate fiscal deficits, expectations, and financial markets into models with purchasing power parity. These models can be grouped into three generations, each of which is intended to explain specific aspects that lead to a currency crisis.

Abbigail J. Chiodo is a senior research associate and Michael T. Owyang is an economist at the Federal Reserve Bank of St. Louis. The authors thank Steven Holland, Eric Blankmeyer, John Lewis, and Rebecca Beard for comments and suggestions and Victor Gabor at the World Bank for providing real GDP data.

[1] Kharas, Pinto, and Ulatov (2001) provide a history from a fundamentals-based perspective, focusing on taxes and public debt issues. We endeavor to incorporate a role for monetary policy.

[2] The speculative attack need not be successful to be dubbed a currency crisis.

[3] Burnside, Eichenbaum, and Rebelo (2001) show that the government has at its disposal a number of mechanisms to finance the fiscal costs of the devaluation. Which policy is chosen determines the inflationary effect of the currency crisis.

First-Generation Models

The first-generation models of a currency crisis developed by Krugman (1979) and Flood and Garber (1984) rely on government debt and the perceived inability of the government to control the budget as the key causes of the currency crisis. These models argue that a speculative attack on the domestic currency can result from an increasing current account deficit (indicating an increase in the trade deficit) or an expected monetization of the fiscal deficit. The speculative attack can result in a sudden devaluation when the central bank's store of foreign reserves is depleted and it can no longer defend the domestic currency. Agents believe that the government's need to finance the debt becomes its overriding concern and eventually leads to a collapse of the fixed exchange rate regime and to speculative attacks on the domestic currency.

Krugman presents a model in which a fixed exchange rate regime is the inevitable target of a speculative attack. An important assumption in the model is that a speculative attack is inevitable. The government defends the exchange rate peg with its store of foreign currency. As agents change the composition of their portfolios from domestic to foreign currency (because rising fiscal deficits increase the likelihood of devaluation, for example), the central bank must continue to deplete its reserves to stave off speculative attacks. The crisis is triggered when agents expect the government to abandon the peg. Anticipating the devaluation, agents convert their portfolios from domestic to foreign currency by buying foreign currency from the central bank's reserves. The central bank's reserves fall until they reach the critical point when a peg is no longer sustainable and the exchange rate regime collapses. The key contribution of the first-generation model is its identification of the tension between domestic fiscal policy and the fixed exchange rate regime.[4]

While the first-generation models help explain some of the fundamentals that cause currency crises, they are lacking in two key aspects. First, the standard first-generation model requires agents to suddenly increase their estimates of the likelihood of a devaluation (perhaps through an increase in expected inflation). Second, they do not explain why the currency crises spread to other countries.

Second-Generation Models

The second-generation models suggested by Obstfeld (1994), Eichengreen, Rose, and Wyplosz (1997), and others are particularly useful in explaining self-fulfilling contagious currency crises. One possible scenario suggested by these models involves a devaluation in one country affecting the price level (and therefore the demand for money) or the current account by a reduction of exports in a neighboring country. In either case, devaluation in a neighboring country becomes increasingly likely.

Eichengreen, Rose, and Wyplosz (1997) find that a correlation exists between the likelihood of default across countries. That is, the probability of a speculative attack in country A increases when its trading partner, country B, experiences an attack of its own. They estimate that a speculative attack somewhere in the world increases the probability of a domestic currency crisis by about 8 percent. The spillover from one currency crisis into neighboring countries can be attributed to a number of different scenarios. First, an economic event, such as a war or an oil price shock, that is common to a geographical area or a group of trading partners can affect those economies simultaneously; in addition, an individual shock can be transmitted from one country to another via trade. Second, a devaluation or default in one country can raise expectations of the likelihood of a devaluation in other countries. Expectations can rise either because countries are neighboring trade partners or because they have similar macroeconomic policies or conditions (e.g., high unemployment or high government debt). Since the crises are self-fulfilling, these expectations make the likelihood of devaluation increase as well. Lastly, a devaluation can be transmitted via world financial markets to other susceptible countries. Any combination of scenarios can serve as an explanation of the apparent international linkages that are responsible for the spread of speculative attacks from one country to another.

Third-Generation Models

The literature on contagious currency crises has helped clarify the spread of devaluations and their magnitudes. However, the first two generations of models have not provided a policy recommendation for the central bank in the face of a crisis. Indeed, Krugman's first-generation model suggests that a crisis cannot be thwarted—that once a devaluation is expected, it is inevitable. Thus, third-generation

[4] Obstfeld (1986) outlines a multiple equilibrium model in which a currency crisis is brought about when government policy (financing a deficit through seignorage, for example) causes agents to expect a crisis and push the economy to a bad equilibrium.

currency crisis models suggested by Krugman (1999) and Aghion, Bacchetta, and Banarjee (2000, 2001) examine the effects of monetary policy in a currency crisis.

These models argue that fragility in the banking and financial sector reduces the amount of credit available to firms and increases the likelihood of a crisis. They suggest that a currency crisis is brought on by a combination of high debt, low foreign reserves, falling government revenue, increasing expectations of devaluation, and domestic borrowing constraints. Firms' access to domestic loans is constrained by assuming they can borrow only a portion of their wealth (somewhat similar to requiring the firm to collateralize all domestic loans). In these lending-constrained economies, the credit market does not clear: interest rates rise, but not enough to compensate investors for the increase in perceived default risk. Increasing the domestic interest rate, then, does not raise the supply of domestic lending in the normal fashion. Moral hazard, a firm's ability to take its output and default on its loan, forces banks to restrict lending. Therefore, increasing the interest rate reduces the amount of loans as it increases firms' incentive to default.

These third-generation models offer a role for monetary policy (aside from the decision to abandon the exchange rate peg) through a binding credit constraint in an imperfect financial market. If firms' leverage in the domestic market is substantially reduced, they may be forced to accumulate a large amount of foreign-denominated debt. When, in domestic markets, the amount of available lending depends on the nominal interest rate, the central bank can deepen a crisis by further reducing firms' ability to invest. The typical prescription for a currency crisis is to raise interest rates and raise the demand for domestic currency.[5] However, in the third-generation models, an interest rate increase can greatly affect the amount of lending and further restrict firms' access to financial capital. In cases where lending is highly sensitive to the interest rate, an increase in the nominal interest rate can be detrimental, altering the productive capacity of the economy by stifling investment. The perceived drop in output puts additional pressure on the exchange rate, perhaps through actual or expected tax revenue, exacerbating the crisis. In this situation, an alternative strategy for the central bank is warranted: it is actually beneficial to lower the interest rate to spur investment.[6]

These three generations of models suggest four factors that can influence the onset and magnitude of a currency crisis. Domestic public and private debt, expectations, and the state of financial markets can, in combination with a pegged exchange rate, determine whether a country is susceptible to a currency crisis and also determine the magnitude and success of a speculative attack. In the next section, we provide an example of a recent currency crisis, keeping these four factors in mind.

THE RUSSIAN DEFAULT: A BRIEF HISTORY

After six years of economic reform in Russia, privatization and macroeconomic stabilization had experienced some limited success. Yet in August 1998, after recording its first year of positive economic growth since the fall of the Soviet Union, Russia was forced to default on its sovereign debt, devalue the ruble, and declare a suspension of payments by commercial banks to foreign creditors. What caused the Russian economy to face a financial crisis after so much had been accomplished? This section examines the sequence of events that took place in Russia from 1996 to 1998 and the aftermath of the crisis. (For a timeline, see Table 1.)

1996 and 1997

Optimism and Reform. In April 1996, Russian officials began negotiations to reschedule the payment of foreign debt inherited from the former Soviet Union. The negotiations to repay its sovereign debt were a major step toward restoring investor confidence. On the surface, 1997 seemed poised to be a turning point toward economic stability.

- The trade surplus was moving toward a balance between exports and imports (see Figure 1).
- Relations with the West were promising: the World Bank was prepared to provide expanded assistance of $2 to $3 billion per year and the International Monetary Fund (IMF) continued to meet with Russian officials and provide aid.
- Inflation had fallen from 131 percent in 1995 to 22 percent in 1996 and 11 percent in 1997 (see Figure 2).
- Output was recovering slightly.

[5] Flood and Jeanne (2000) argue that increasing domestic currency interest rates can act only to speed devaluation.

[6] The expansionary monetary policy in this case is assumed not to be inflationary since it only alleviates liquidity constraints.

Table 1

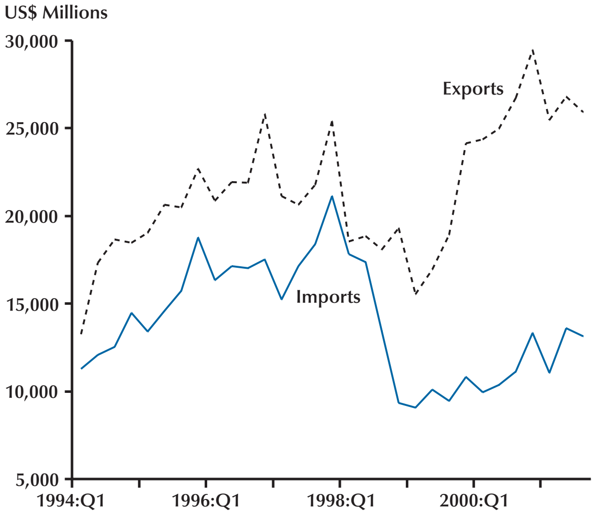
A Timeline of Russian Events

April 1996	Negotiations with the Paris and London Clubs for repayment of Soviet debt begin.
1997	Trade surplus moving toward balance. Inflation around 11 percent. Oil selling at $23/barrel. Analysts predict better credit ratings for Russia. Russian banks increase foreign liabilities. Real wages sagging. Only 40 percent of workforce being paid fully and on time. Public-sector deficit high.
September/October 1997	Negotiations with Paris and London Clubs completed.
November 11, 1997	Asian crisis causes a speculative attack on the ruble. CBR defends the ruble, losing $6 billion.
December 1997	Year ends with 0.8 percent growth. Prices of oil and nonferrous metal begin to drop.
February 1998	New tax code submitted to the Duma. IMF funds requested.
March 23, 1998	Yelstin fires entire government and appoints Kiriyenko. Continued requests for IMF funds.
April 1998	Another speculative attack on the ruble.
April 24, 1998	Duma finally confirms Kiriyenko's appointment.
Early May 1998	Dubinin warns government ministers of impending debt crisis, with reporters in the audience. Kiriyenko calls the Russian government "quite poor."
May 19, 1998	CBR increases lending rate from 30 percent to 50 percent and defends the ruble with $1 billion.
Mid May 1998	Lawrence Summers not granted audience with Kiriyenko. Oil prices continue to decrease. Oil and gas oligarchs advocate devaluation of ruble to increase value of their exports.
May 23, 1998	IMF leaves Russia without agreement on austerity plan.
May 27, 1998	CBR increases the lending rate again to 150 percent.
Summer 1998	Russian government formulates and advertises anti-crisis plan.
July 20, 1998	IMF approves an emergency aid package (first disbursement to be $4.8 billion).
August 13, 1998	Russian stock, bond, and currency markets weaken as a result of investor fears of devaluation; prices diminish.
August 17, 1998	Russian government devalues the ruble, defaults on domestic debt, and declares a moratorium on payment to foreign creditors.
August 23-24, 1998	Kiriyenko is fired.
September 2, 1998	The ruble is floated.
December 1998	Year ends with a decrease in real output of 4.9 percent.

NOTE: CBR, Central Bank of Russia.

Figure 1

Russian Merchandise Trade Balance

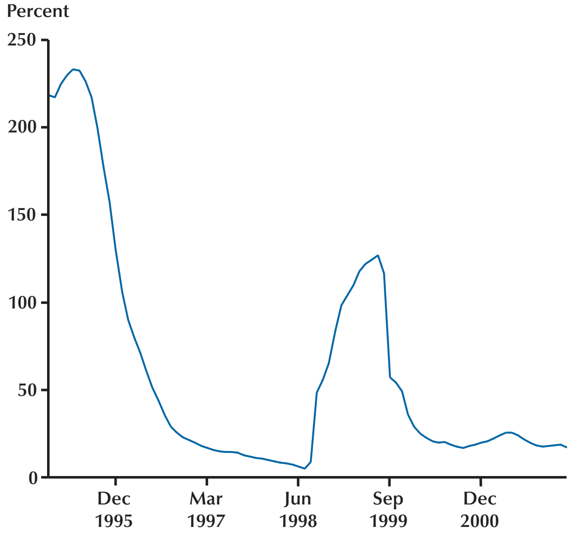


SOURCE: CBR.

Figure 2

CPI Inflation

Percent Change over Previous Year



SOURCE: IMF.

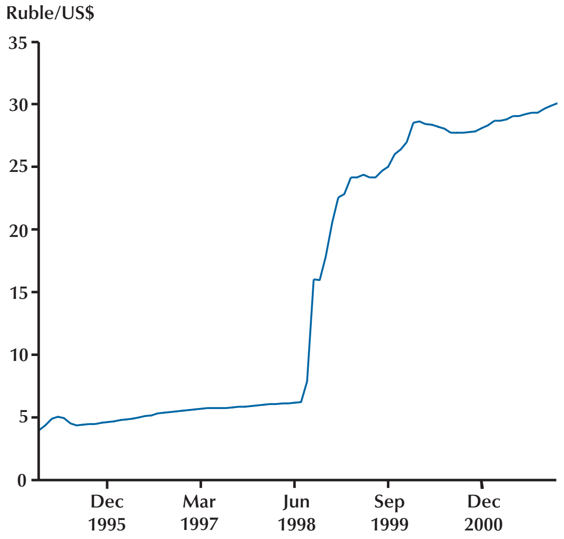
- A narrow exchange rate band was in place keeping the exchange rate between 5 and 6 rubles to the dollar (see Figure 3).
- And oil, one of Russia's largest exports, was selling at $23 per barrel—a high price by recent standards. (Fuels made up more than 45 percent of Russia's main export commodities in 1997.)

In September 1997, Russia was allowed to join the Paris Club of creditor nations after rescheduling the payment of over $60 billion in old Soviet debt to other governments. Another agreement for a 23-year debt repayment of $33 billion was signed a month later with the London Club. Analysts predicted that Russia's credit ratings would improve, allowing the country to borrow less expensively. Limitations on the purchase of government securities by nonresident investors were removed, promoting foreign investment in Russia. By late 1997, roughly 30 percent of the GKO (a short-term government bill) market was accounted for by nonresidents. The economic outlook appeared optimistic as Russia ended 1997 with reported economic growth of 0.8 percent.

Revenue, Investment, and Debt. Despite the prospects for optimism, problems remained. On average, real wages were less than half of what they were in 1991, and only about 40 percent of the work force was being paid in full and on time. Per capita direct foreign investment was low, and regulation of the natural monopolies was still difficult due to unrest in the Duma, Russia's lower house of Parliament. Another weakness in the Russian economy was low tax collection, which caused the

Figure 3

Exchange Rate



SOURCE: IMF (end of period data).

public sector deficit to remain high. The majority of tax revenues came from taxes that were shared between the regional and federal governments, which fostered competition among the different levels of government over the distribution. According to Shleifer and Treisman (2000), this kind of tax sharing can result in conflicting incentives for regional governments and lead them to help firms conceal part of their taxable profit from the federal government in order to reduce the firms' total tax payments. In return, the firm would then make transfers to the accommodating regional government. This, Shleifer and Treisman suggest, may explain why federal revenues dropped more rapidly than regional revenues.

Also, the Paris Club's recognition of Russia as a creditor nation was based upon questionable qualifications. One-fourth of the assets considered to belong to Russia were in the form of debt owed to the former Soviet Union by countries such as Cuba, Mongolia, and Vietnam. Recognition by the Paris Club was also based on the old, completely arbitrary official Soviet exchange rate of approximately 0.6 rubles to the dollar (the market exchange rate at the time was between 5 and 6 rubles to the dollar). The improved credit ratings Russia received from its Paris Club recognition were not based on an improved balance sheet. Despite this, restrictions were eased and lifted and Russian banks began borrowing more from foreign markets, increasing their foreign liabilities from 7 percent of their assets in 1994 to 17 percent in 1997.

Meanwhile, Russia anticipated growing debt payments in the coming years when early credits from the IMF would come due. Policymakers faced decisions to decrease domestic borrowing and increase tax collection because interest payments were such a large percentage of the federal budget. In October 1997, the Russian government was counting on 2 percent economic growth in 1998 to compensate for the debt growth. Unfortunately, events began to unfold that would further strain Russia's economy; instead of growth in 1998, real GDP declined 4.9 percent.

The Asian Crisis. A few months earlier, in the summer of 1997, countries in the Pacific Rim experienced currency crises similar to the one that eventually affected Russia. In November 1997, after the onset of this East Asian crisis, the ruble came under speculative attack. The Central Bank of Russia (CBR) defended the currency, losing nearly $6 billion (U.S. dollars) in foreign-exchange reserves. At the same time, non-resident holders of short-term government bills (GKOs) signed forward contracts with the CBR to exchange rubles for foreign currency, which enabled them to hedge exchange rate risk in the interim period.[7] According to Desai (2000), they did this in anticipation of the ruble losing value, as Asian currencies had. Also, a substantial amount of the liabilities of large Russian commercial banks were off-balance-sheet, consisting mostly of forward contracts signed with foreign investors. Net obligations of Russian banks for such contracts were estimated to be at least $6 billion by the first half of 1998. Then another blow was dealt to the Russian economy: in December 1997, the prices of oil and nonferrous metal, up to two-thirds of Russia's hard-currency earnings, began to drop.

1998

Government, Risk, and Expectations. With so many uncertainties in the Russian economy, investors turned their attention toward Russian default risk. To promote a stable investment environment, in February 1998, the Russian government submitted a new tax code to the Duma, with fewer and more efficient taxes. The new tax code was approved in 1998, yet some crucial parts that were intended to increase federal revenue were ignored. Russian officials sought IMF funds but agreements could not be reached. By late March the political and economic situation had become more dire, and, on March 23, President Yeltsin abruptly fired his entire government, including Prime Minister Viktor Chernomyrdin. In a move that would challenge investor confidence even further, Yeltsin appointed 35-year-old Sergei Kiriyenko, a former banking and oil company executive who had been in government less than a year, to take his place.

While fears of higher interest rates in the United States and Germany made many investors cautious, tensions rose in the Russian government. The executive branch, the Duma, and the CBR were in conflict. Prompted by threats from Yeltsin to dissolve Parliament, the Duma confirmed Kiriyenko's appointment on April 24 after a month of stalling. In early May, during a routine update, CBR chair Sergei Dubinin warned government ministers of a debt crisis within the next three years. Unfortunately, reporters were in the audience. Since the Asian crisis had heightened investors' sensitivity to currency stability, Dubinin's

[7] The requirement of forward contracts was the CBR's way of preventing runs on its foreign currency reserves.

restatement of bank policy was misinterpreted to mean that the Bank was considering a devaluation of the ruble. In another public relations misunderstanding, Kiriyenko stated in an interview that tax revenue was 26 percent below target and claimed that the government was "quite poor now." In actuality, the government was planning to cut government spending and accelerate revenue, but these plans were never communicated clearly to the public. Instead, people began to expect a devaluation of the ruble.

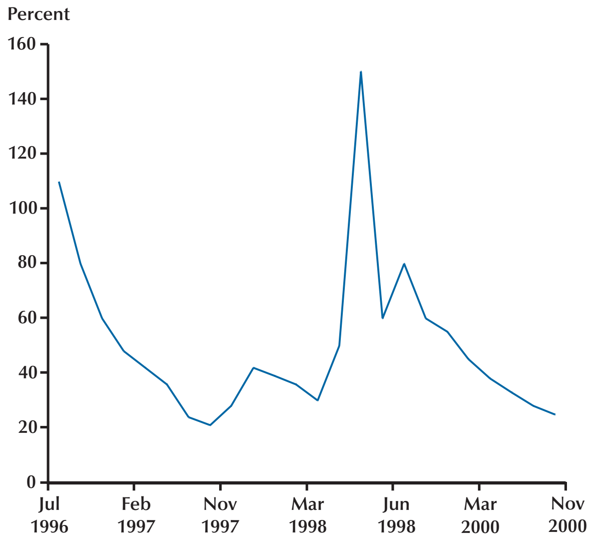
Investors' perceptions of Russia's economic stability continued to decline when Lawrence Summers, one of America's top international-finance officials, was denied a meeting with Kiriyenko while in Russia. An inexperienced aide determined that Summers's title, Deputy Secretary of the Treasury, was unworthy of Kiriyenko's audience and the two never met. At the same time, the IMF left Russia, unable to reach an agreement with policymakers on a 1998 austerity plan. Word spread of these incidents, and big investors began to sell their government bond portfolios and Russian securities, concerned that relations between the United States and Russia were strained.

Liquidity, Monetary Policy, and Fiscal Policy. By May 18, government bond yields had swelled to 47 percent. With inflation at about 10 percent, Russian banks would normally have taken the government paper at such high rates. Lack of confidence in the government's ability to repay the bonds and restricted liquidity, however, did not permit this. As depositors and investors became increasingly cautious of risk, these commercial banks and firms had less cash to keep them afloat. The federal government's initiative to collect more taxes in cash lowered banks' and firms' liquidity.[8] Also, in 1997, Russia had created a U.S.-style treasury system with branches, which saved money and decreased corruption, yet also decreased the amount of cash that moved through banks. The banks had previously used these funds to buy bonds. Also, household ruble deposits increased by only 1.3 billion in 1998, compared with an increase of 29.8 billion in 1997.

The CBR responded by increasing the lending rate to banks from 30 to 50 percent, and in two days used $1 billion of Russia's low reserves to defend the ruble. (Figure 4 shows the lending rate.) However, by May 27, demand for bonds had plummeted so much that yields were more than 50 percent and the government failed to sell enough bonds at its weekly auction to refinance the debt coming due.

Figure 4

Lending Rate



SOURCE: CBR.

Meanwhile, oil prices had dropped to $11 per barrel, less than half their level a year earlier. Oil and gas oligarchs were advocating a devaluation of the ruble, which would increase the ruble value of their exports. In light of this, the CBR increased the lending rate again, this time to 150 percent. CBR chairman Sergei Dubinin responded by stating "When you hear talk of devaluation, spit in the eye of whoever is talking about it" (quoted in Shleifer and Treisman, 2000, p. 149).

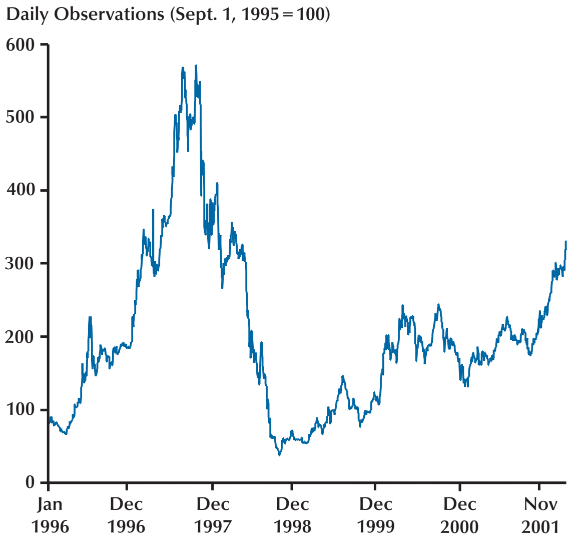
The government formed and advertised an anticrisis plan, requested assistance from the West, and began bankruptcy processes against three companies with large debts from back taxes. Kiriyenko met with foreign investors to reassure them. Yeltsin made nightly appearances on Russian television, calling the nation's financial elite to a meeting at the Kremlin where he urged them to invest in Russia. In June the CBR defended the ruble, losing $5 billion in reserves.

Despite all of the government efforts being made, there was widespread knowledge of $2.5 to $3 billion

[8] As a result of a 1998 elimination of tax-offsets paper issued by government agencies to pay for goods and services, the receipts of which could be used to decrease their tax duties, banks and companies were forced to provide more cash to pay their taxes, thus lowering their liquidity.

Figure 5

The Russian Stock Market



SOURCE: <http://red-stars.com/financial>.

in loans from foreign investors to Russian corporations and banks that were to come due by the end of September. In addition, billions of dollars in ruble futures were to mature in the fall. In July the IMF approved additional assistance of $11.2 billion, of which $4.8 billion was to be disbursed immediately. Yet between May and August, approximately $4 billion had left Russia in capital flight, and in 1998 Russia lost around $4 billion in revenue due to sagging oil prices. After losing so much liquidity, the IMF assistance did not provide much relief.

The Duma, in an effort to protect natural monopolies from stricter regulations, eliminated crucial parts of the IMF-endorsed anti-crisis program before adjourning for vacation. The government had hoped that the anti-crisis plan would bring in an additional 71 billion rubles in revenue. The parts that the Duma actually passed would have increased it by only 3 billion rubles. In vain, lawmakers requested that the Duma reconvene, lowering investors' confidence even further.

Default and Devaluation. On August 13, 1998, the Russian stock, bond, and currency markets collapsed as a result of investor fears that the government would devalue the ruble, default on domestic debt, or both. Annual yields on ruble-denominated bonds were more than 200 percent. The stock market had to be closed for 35 minutes as prices plummeted. When the market closed, it was down 65 percent with a small number of shares actually traded. From January to August the stock market had lost more than 75 percent of its value, 39 percent in the month of May alone. (Figure 5 shows the Russian stock market's boom and bust.) Russian officials were left with little choice. On August 17 the government floated the exchange rate, devalued the ruble, defaulted on its domestic debt, halted payment on ruble-denominated debt (primarily GKOs), and declared a 90-day moratorium on payment by commercial banks to foreign creditors.

The Aftermath

Russia ended 1998 with a decrease in real output of 4.9 percent for the year instead of the small growth that was expected. The collapse of the ruble created an increase in Russia's exports while imports remained low (see Figure 1). Since then, direct investments into Russia have been inconsistent at best. Summarized best by Shleifer and Treisman (2000), "the crisis of August 1998 did not only undermine Russia's currency and force the last reformers from office…it also seemed to erase any remaining Western hope that Russia could successfully reform its economy."

Some optimism, however, still persists. Figure 6 shows Russian real GDP growth, which grew 8.3 percent in 2000 and roughly 5 percent in 2001—lower but still positive. Imports trended up in the first half of 2001, helping to create a trade balance. At the same time, consumer prices grew 20.9 percent and 21.6 percent in 2000 and 2001, respectively, compared with a 92.6 percent increase in 1999. Most of the recovery so far can be attributed to the import substitution effect after the devaluation; the increase in world prices for Russia's oil, gas, and commodity exports; monetary policies; and fiscal policies that have led to the first federal budget surplus (in 2000) since the formation of the Russian Federation.

HOW DO THE THEORIES EXPLAIN THE RUSSIAN CRISIS?

As discussed earlier, four major factors influence the onset and success of a speculative attack. These key ingredients are (i) an exchange rate peg and a central bank willing or obligated to defend it with a reserve of foreign currency, (ii) rising fiscal deficits that the government cannot control and therefore is likely to monetize (print money to cover the deficit), (iii) central bank control of the interest rate in a fragile credit market, and (iv) expectations of devaluation and/or rising inflation. In this section we discuss these aspects in the context of the Russian devaluation. We argue that an understanding of all three generations of models is necessary to evaluate the Russian devaluation. Krugman's (1979) first-generation model explains the factors that made Russia susceptible to a crisis. The second-generation models show how contagion and other factors can change expectations to trigger the crisis. The third-generation models show how the central bank can act to prevent or mitigate the crisis.

The Exchange Rate and the Peg

When the ruble came under attack in November 1997 and June 1998, policymakers defended the ruble instead of letting it float. The real exchange rate did not vary much during 1997. Clearly a primary component of a currency crisis in the models described here is the central bank's willingness to defend an exchange rate peg. Prior to August 1998, the Russian ruble was subject to two speculative attacks. The CBR made efforts both times to defend the ruble. The defense was successful in November 1997 but fell short in the summer of 1998. Defending the ruble depleted Russia's foreign reserves. Once depleted, the Russian government had no choice but to devalue on August 17, 1998.

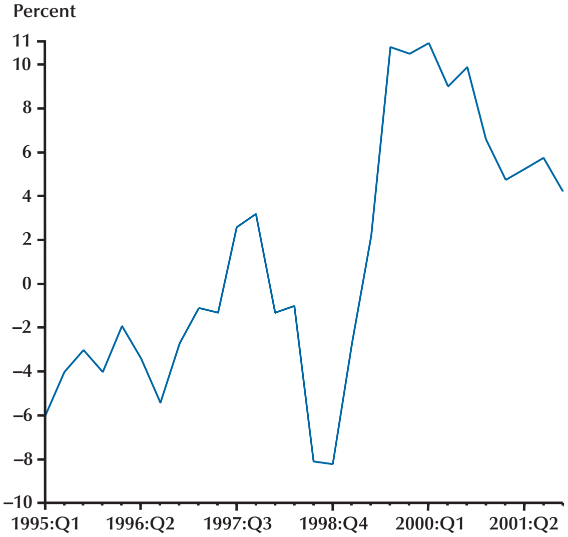
Revenue, Deficits, and Fiscal Policy

Russia's high government debt and falling revenue contributed significantly to its susceptibility to a speculative attack. Russia's federal tax revenues were low because of both low output and the opportunistic practice of local governments helping firms conceal profits. The decrease in the price of oil also lowered output, further reducing Russia's ability to generate tax revenue. Consequently, Russia's revenue was lower than expected, making the ruble ripe for a speculative attack. In addition, a large amount of short-term foreign debt was coming due in 1998, making Russia's deficit problem even more serious. Krugman's first-generation model suggests that a government finances its deficit by printing money (seigniorage) or depleting its reserves of foreign currency. Under the exchange rate peg, however, Russia was unable to finance through seigniorage. Russia's deficit, low revenue, and mounting interest payments put pressure on the exchange rate. Printing rubles would only have increased this pressure because the private sector would still have been able to trade rubles for foreign currency at the fixed rate. Thus, whether directly through intervention in the foreign currency market or indirectly by printing rubles, Russia's only alternative under the fixed exchange rate regime was to deplete its stock of foreign reserves.

Figure 6

Real GDP Growth

Quarterly Change from Previous Year



SOURCE: Russian Statistics Committee and International Bank for Reconstruction and Development staff estimates.

Monetary Policy, Financial Markets, and Interest Rates

During the summer of 1998, the Russian economy was primed for the onset of a currency crisis. In an attempt to avert the crisis, the CBR intervened by decreasing the growth of the money supply and twice increasing the lending rate to banks, raising it from 30 to 150 percent. Both rate hikes occurred in May 1998, the same month in which the Russian stock market lost 39 percent of its value. The rise in interest rates had two effects. First, it exacerbated Russia's revenue problems. Its debt grew rapidly as interest payments mounted. This put pressure on the exchange rate because investors feared that Russia would devalue to finance its non-denominated debt. Second, high government debt prevented firms from obtaining loans for new capital and increasing the interest rate did not increase the supply of lending capital available to firms. At the same time, foreign reserves held by the CBR were so low that the government could no longer defend the currency by buying rubles.

Expectations

Three components fueled the expectations of Russia's impending devaluation and default. First, the Asian crisis made investors more conscious of the possibility of a Russian default. Second, public relations errors, such as the publicized statement to government ministers by the CBR and Kiriyenko's refusal to grant Lawrence Summers an audience, perpetuated agents' perceptions of a political crisis within the Russian government. Third, the revenue shortfall signaled the possible reduction of the public debt burden via an increase in the money supply. This monetization of the debt can be associated with a depreciation either indirectly through an increase in expected inflation or directly in order to reduce the burden of ruble-denominated debt. Each of these three components acted to push the Russian economy from a stable equilibrium to one vulnerable to speculative attack.

CONCLUSION

In this paper we investigate the events that lead up to a currency crisis and debt default and the policies intended to avert it. Three types of models exist to explain currency crises. Each model explains some factor that has been hypothesized to cause a crisis. After reviewing the three generations of currency crisis models, we conclude that four key ingredients can trigger a crisis: a fixed exchange rate, fiscal deficits and debt, the conduct of monetary policy, and expectations of impending default. Using the example of the Russian default of 1998, we show that the prescription of contractionary monetary policy in the face of a currency crisis can, under certain conditions, accelerate devaluation. While we believe that deficits and the Asian financial crisis contributed to Russia's default, the first-generation model proposed by Krugman (1979) and Flood and Garber (1984) and the second-generation models proposed by Obstfeld (1984) and Eichengreen, Rose, and Wyplosz (1997) do not capture every aspect of the crisis. Specifically, these models do not address the conduct of monetary policy. It is therefore necessary to incorporate both the first-generation model's phenomenon of increasing fiscal deficits and the third-generation model's financial sector fragility. We conclude that the modern currency crisis is a symptom of an ailing domestic economy. In that light, it is inappropriate to attribute a single prescription as the prophylactic or cure for a currency crisis.

REFERENCES

Aghion, Philippe; Bacchetta, Philippe and Banerjee, Abhijit. "A Simple Model of Monetary Policy and Currency Crises." *European Economic Review*, May 2000, *44*(4-6), pp. 728-38.

__________; __________ and __________. "Currency Crises and Monetary Policy in an Economy with Credit Constraints." *European Economic Review*, June 2001, *45*(7), pp. 1121-50.

Ahrend, Rudiger. "Foreign Direct Investment Into Russia—Pain Without Gain? A Survey of Foreign Direct Investors." *Russian Economic Trends*, June 2000, *9*(2), pp. 26-33.

Burnside, Craig; Eichenbaum, Martin, and Rebelo, Sergio. "On The Fiscal Implications of Twin Crises." Working Paper No. 8277, National Bureau of Economic Research, May 2001.

Desai, Padma. "Why Did the Ruble Collapse in August 1998?" *American Economic Review: Papers and Proceedings*, May 2000, *90*(2), pp. 48-52.

Economist. "Surplus to Requirements." 8 July 2000, p. 79.

Eichengreen, Barry; Rose, Andrew and Wyplosz, Charles. "Contagious Currency Crisis." March 1997. <http://www.haas.berkeley.edu/~arose/>.

Fischer, Stanley. "The Russian Economy at the Start of 1998." U.S.-Russian Investment Symposium, Harvard University, Cambridge, MA, 9 January 1998.

Flemming, Marcus. "Domestic Financial Policies Under Fixed and Under Floating Exchange Rates." *IMF Staff Papers*, 9 November 1962.

Flood, Robert P. and Garber, Peter M. "Collapsing Exchange Rate Regimes: Some Linear Examples." *Journal of International Economics*, August 1984, *17*(1-2), pp 1-13.

__________ and Jeanne, Olivier. "An Interest Rate Defense of a Fixed Exchange Rate?" Working Paper WP/00/159, International Monetary Fund, October 2000.

Kharas, Homi; Pinto, Brian and Ulatov, Sergei. "An Analysis of Russia's 1998 Meltdown: Fundamentals and Market Signals." *Brookings Papers on Economic Activity*, 2001, *0*(1), pp. 1-67.

Krugman, Paul. "A Model of Balance-of-Payment Crises." *Journal of Money, Credit, and Banking*, August 1979, *11*(3), pp. 311-25.

___________. "Balance Sheets, the Transfer Problem, and Financial Crises." *International Tax and Public Finance*, November 1999, *6*(4), pp. 459-72.

Malleret, Thierry; Orlova, Natalia and Romanov, Vladimir. "What Loaded and Triggered the Russian Crisis?" *Post-Soviet Affairs*, April-June 1999, *15*(2), pp. 107-29.

Mudell, R.A. "Capital Mobility and Stabilization Policy Under Fixed and Flexible Exchange Rates." *Canadian Journal of Economics*, November 1963.

Obstfeld, Maurice. "Rational and Self-Fulfilling Balance-of-Payments Crises." *American Economic Review*, March 1986, *76*(1), pp. 72-81.

___________. "The Logic of Currency Crises." *Cahiers Economiques et Monetaires*, Banque de France, 1994, *43*, pp. 189-213.

Popov, A. "Lessons of the Currency Crisis in Russia and in Other Countries." *Problems of Economic Transition*, May 2000, *43*(1), pp. 45-73.

Russian Economic Trends. Various months.

Shleifer, Andre and Treisman, Daniel. *Without A Map: Political Tactics and Economic Reform in Russia*. Cambridge, MA: MIT Press, 2000.

Velasco, Andrés. "Financial Crises in Emerging Markets." National Bureau of Economic Research *Reporter*, Fall 1999, pp.17-19.

Asset Mispricing, Arbitrage, and Volatility

William R. Emmons and Frank A. Schmid

After nearly four decades, academic economists continue to debate financial-market efficiency as vigorously as ever.[1] The original theoretical arguments put forward in favor of efficient markets were based on the notion of stabilizing speculation in the form of arbitrage (Friedman, 1953). Simply put, arbitrage is "the simultaneous purchase and sale of the same, or essentially similar, security in two different markets for advantageously different prices" (Sharpe and Alexander, 1990). In theory, a perfectly hedged trading position of this sort could be executed at no cost (as the short-sale proceeds are used to finance the long position). Vigilant traders on the look-out for just such arbitrage opportunities would ensure that no one could consistently "beat the market"—the hallmark of efficient markets theory.

The academics' logical case for efficient markets boils down to a pair of simple rhetorical questions: Why would utility-maximizing traders leave unexploited any profitable opportunities (after adjusting properly for risk)? And if no risk-adjusted "free lunches" exist, how could market prices be predictable enough to make money? For several decades, empirical evidence piled up both for and against market efficiency. As of the early 1990s, neither side could claim total vindication. As the 1990s progressed, however, the weight of the evidence seemed to tip toward those who claimed asset prices were, at least to some extent, predictable (Campbell, Lo, and MacKinlay, 1997, Chaps. 2 and 7).

The academic asset-pricing literature today is dominated by attempts to explain why and to what extent the price movements of financial assets are predictable. One potential explanation of stock-return predictability is that markets are efficient ("no free lunch") but expected returns are time-varying, perhaps being linked to the business cycle. For example, expected returns may be highest when economic risks are perceived to be high, such as at or near the bottom of a business cycle. Conversely, expected returns may be lowest when economic risks are perceived to be low, at or near a business-cycle peak. Thus, the simple random-walk model of stock returns may be false, but a relevant notion of market efficiency survives because high returns are earned only by taking large amounts of risk. A different type of explanation of return predictability rejects market efficiency and focuses on market imperfections of various sorts, such as incomplete stock-market participation by households, significant transactions costs, changes in investor sentiment, or limited wealth and liquidity resources to conduct arbitrage (as in the current article).[2]

Whatever its economic explanation, mounting evidence of return predictability leads Campbell, Lo, and MacKinlay (1997, p. 24) to suggest that it is time for financial economists to focus their attention on the "relative efficiency" of a market instead of continuing the all-or-nothing battle of attrition that is characteristic of much of the earlier market efficiency literature.

As we now understand more clearly, the original case for efficient markets probably leaned too heavily on the notion of risk-free, cost-free arbitrage to eliminate all profitable trading strategies immediately. In real markets, arbitrage is neither as easy nor as effective as economists once had assumed. For one thing, financial markets are not complete and frictionless, so arbitrage in general is risky and costly. In addition, it is not realistic to assume that the number of informed arbitrageurs or the supply of financial resources they have to invest in arbitrage strategies is limitless.

This article builds on an important and insightful recent model of arbitrage by professional traders who need—but lack—wealth of their own to trade (Shleifer and Vishny, 1997). Professional arbitrageurs must convince wealthy but uninformed investors to entrust them with investment capital in order to exploit mispricing and push the market back toward the ideal of efficiency. Unfortunately, arbitrageurs cannot prove that they recognize the intrinsic (or "fundamental") values of the assets they claim are mispriced. Even worse, it is possible the assets will

William R. Emmons is an economist and Frank A. Schmid is a senior economist at the Federal Reserve Bank of St. Louis. William V. Bock provided research assistance.

[1] For early statements of the theory of efficient markets and the unpredictability of asset-price movements, see Fama (1965), Muth (1960), or Samuelson (1965). For a recent summary of the evidence for return predictability and its implications for efficient-markets theory, see Campbell, Lo, and MacKinlay (1997, Chap. 2).

[2] Ironically, Keynes (1936, Chap. 12) clearly foreshadowed the recent interest in investor sentiment and liquidity for understanding stock market behavior, but was forgotten for decades as the efficient-markets hypothesis dominated the academic discussion.

become even more mispriced before reverting eventually to their intrinsic values. Having incurred losses, the outside investors may demand their money back at this point even though the expected profit of staying invested actually has increased.

Thus, market efficiency may depend ultimately on the successful resolution of a principal-agent problem that exists between informed but wealth-constrained arbitrageurs and uninformed wealthy investors. The resulting degree of market efficiency may change over time and differ across markets, and it could depend importantly on factors such as the outside investors' use of performance-based ("feedback") strategies when deciding on the possible termination of ongoing investment mandates.

After developing a simple model of wealth-constrained professional arbitrage that departs in several important aspects from the canonical Shleifer and Vishny (1997) model, we calibrate our model to illustrate its qualitative features. We show that the existence of professional arbitrageurs mitigates—but cannot eliminate—mispricing in the market relative to intrinsic values, regardless of how sensitive the outside investors are to arbitrageurs' past performance in deciding whether to remain invested with them. We also show that arbitrage dampens the unconditional volatility of asset returns, which we measure as the expected value of squared returns. Most importantly, the presence of arbitrageurs limits both the degree of increased mispricing and level of volatility during a financial crisis, which we define as a period of heightened volatility and acute shortage of liquidity.[3] This result points out that professional arbitrageurs tend to stabilize markets even when they are wealth-constrained. Other papers show that investors who use "positive feedback" trading strategies—such as portfolio insurers—tend to destabilize markets (Grossman and Zhou, 1996).

We analyze a three-date (two-period) model of an aspiring professional arbitrageur (or "convergence trader" in the language of Kyle and Xiong, 2001, and Xiong, 2001) who must obtain financing from investors less informed than he is about the intrinsic value of a financial asset—that is, its liquidation value at the end of the second period. In addition to these two types of individuals, there are noise traders who have wealth to invest but who misperceive the asset's intrinsic value. It is the noise traders who drive the asset's price away from the intrinsic value.

The investors provide the arbitrageur with funds to invest in an underpriced asset at the outset of the model. The price is observed again at the end of the first period, at which time the investors may "roll over" their funds with the arbitrageur or demand their money back if they have lost confidence in his ability. The asset will assume its intrinsic value at the end of the second period with certainty, although only the arbitrageur knows in advance what that value is. Consequently, the two-period return on the arbitrageur's private information would be both positive and risk-free if he could be assured of financing.

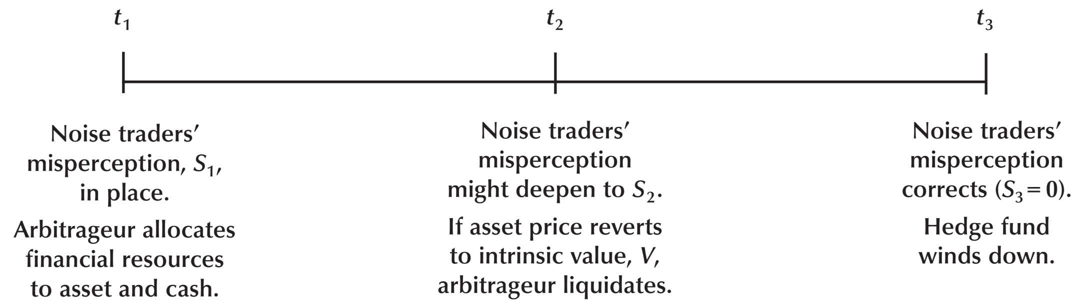
Our set-up highlights the fact that a two-period risk-free arbitrage nevertheless can be risky over a *one*-period horizon in the presence of noise traders and financial constraints on the arbitrageur. The risk arises because the arbitrageur needs outside investors, and these outside investors might revise their beliefs about the arbitrageur's talent at the interim date, based on the return the arbitrageur achieved in the first period. If the investors downwardly revise their beliefs about the arbitrageur's abilities because the fund lost money due to a deepening of the mispricing, they might withdraw their money precisely when the expected return on the arbitrage is at its maximum. One implication is that the arbitrageur will invest "strategically"—that is, he will not invest as much initially as he would in a world without wealth constraints—in order to hedge against the possibility of being unable to exploit even greater mispricing should it occur one period ahead. Of course, this is not a new finding; for papers with similar results, see Grossman and Vila (1992), Shleifer and Vishny (1997), or Gromb and Vayanos (2001).[4] Our paper's contributions in this respect

[3] Myron Scholes (2000) suggests that the global financial crisis of 1997-98 was characterized by an increase in volatility, especially in equity markets, and a flight to liquidity (that is, a preference by many investors for assets whose liquidity was expected to be good). The crisis was accentuated by the "negotiated bankruptcy" of Long-Term Capital Management (LTCM), a hedge fund in which Scholes himself was a partner. According to Scholes, prior to the crisis, LTCM "was in the business of supplying liquidity" and therefore its demise worsened the crisis by eliminating the liquidity it had been supplying. A theoretical analysis relevant to this episode is Xiong (2001).

[4] The first rigorous investigations of the multi-period investment problem were Merton (1971, 1973) and Breedon (1979). Merton concluded that a trader should keep a constant fraction of his wealth invested in the risky asset at all times. The fraction depends on the asset's expected return and risk and the investor's degree of risk aversion. Grossman and Vila (1992) added leverage and solvency constraints to the dynamic trader's problem. Their trader optimally commits more wealth to the risky asset the shorter is the investment horizon and the further from the leverage constraint (not just today but prospectively in the future) the trader finds himself. Campbell and Viceira (1999) is a recent examination of the problem under the assumption that the investor is aware that the probability distributions from which asset returns are drawn change over time.

Figure 1

Timeline of the Model


t_1
t_2
t_3
Noise traders' misperception, S_1, in place.
Arbitrageur allocates financial resources to asset and cash.
Noise traders' misperception might deepen to S_2.
If asset price reverts to intrinsic value, V, arbitrageur liquidates.
Noise traders' misperception corrects ($S_3 = 0$).
Hedge fund winds down.

are a more realistic objective function for the arbitrageur and a set-up in which the arbitrageur's trading significantly affects the asset's price. Our model generates interior solutions and we provide calibrated illustrations of the model's results. While Shleifer and Vishny (1997) assume that the arbitrageur maximizes assets under management, we assume that he maximizes his income. The arbitrageur's income is determined by an incentive scheme that resembles real-world contracts of hedge fund managers.

THE MODEL

There are three types of agents in the model. Noise traders have wealth but misperceive the intrinsic value of a financial asset. Professional arbitrageurs have no wealth or borrowing capacity but know the intrinsic value of the financial asset. Investors have wealth but no insight into the financial asset's intrinsic value. Unlike noise traders, investors *know* that they cannot recognize the asset's intrinsic value. All parties are risk-neutral.

The investors may provide the arbitrageur with funds to invest in an underpriced asset at the outset of the model (see Figure 1). We refer to this arrangement as a hedge fund. Noise traders misperceive the intrinsic value of at least one financial asset in the economy, which generates arbitrage opportunities that so-called "long-short" investment strategies seek to exploit. Asset mispricing implies that there are relatively overpriced and relatively underpriced assets, which means that a portfolio that is long on the relatively undervalued asset and short on the relatively overvalued asset trades below intrinsic value.

We treat a market-neutral long-short portfolio as a single, complex financial asset. Arbitrage is the process of acquiring a long-short portfolio and holding it until its price returns to the portfolio's intrinsic value. The long-short portfolio that any arbitrageur might hold defines a market segment of a larger arbitrage industry. We assume that arbitrageurs are highly skilled people who pursue proprietary trading strategies and therefore enjoy a monopoly in their segment. For simplicity only, we make the assumption that the operating costs in the arbitrage industry are zero.

The risk-free rate of return, and therefore the opportunity cost of capital, is zero. For simplicity, we assume that risky assets trading at fair value—including the stock market index—also have an expected return of zero. This implies that there are no priced systematic risk factors in the economy, that is, there is no equity risk premium.

The asset trades at three moments in time, t ($t = 1,2,3$). We capture the influence of the noise traders' misperceptions of the intrinsic value of the asset at times t_1 and t_2 with the parameters S_1 and S_2, respectively. There is no fundamental risk in the model because the price of the asset will revert to the intrinsic value at a known date (t_3) with certainty (so $S_3 = 0$).

The supply of the financial asset is unity. Noise traders' demand for the financial asset at time t ($t = 1,2,3$) is expressed as

$$(1) \qquad QN_t = \frac{V - S_t}{p_t}, \; 0 \leq S_t < V,$$

where p_t is the price of the financial asset and S_t is the misperception of the noise traders about the intrinsic value of the financial asset. Because the financial asset in question is a long-short portfolio whose value is underestimated by the noise traders, the noise traders demand less than one unit of the financial asset. Without misperception ($S = 0$), the noise traders would be willing to absorb the unit supply of the asset or, in other words, the asset would trade at the intrinsic value ($p_t = V$).

The arbitrageur is compensated in two ways in accord with actual practice—via an up-front "management fee" and an after-the-fact performance-based "incentive fee."[5] At the beginning of each period, he receives a fraction (α) of the assets under management, and at the end of the period he receives a fraction (β) of any positive return on the portfolio. This corresponds to compensation structures in real-world hedge funds, where managers typically collect $\alpha = 1$ percent or $\alpha = 2$ percent of the equity capital, plus $\beta = 20$ percent of any positive return on the fund's equity. We assume that the arbitrageur invests his entire fee income in the fund. This is because he recognizes the profitability of the fund's activities.

The variable F_t denotes the total financial resources available to the arbitrageur at time t ($t = 1,2,3$). The value of F_1 is exogenous, while the quantities F_2 and F_3 are determined in the model. The startup capital, F_1, is provided solely by the investors, while the arbitrageur acquires the share α in F_1 immediately as part of his compensation. The arbitrageur acquires additional equity at t_2 in the amount of a fraction α of the outsiders' share in F_2. Furthermore, the arbitrageur acquires equity in the fund through capital gains on his equity position and through his share β in the capital gains on the outsiders' equity. The quantity F_3 is the fund's liquidation value. Note that the arbitrageur is both the general equity partner of the fund and its manager, receiving compensation from outside investors (limited partners) according to the fee schedule described above.

We assume that the fund raises equity capital only at the outset—at t_1. This assumption prevents the arbitrageur from diluting initial investors' equity stakes later on. Remember that the arbitrageur's compensation depends not only on the return on but also on the amount of the outsiders' equity capital under management. The arbitrageur therefore might have an incentive to raise fresh capital at t_2, particularly if he expects low returns in the second period. This would dilute the fund's existing investors' equity stakes. Thus, we assume (in keeping with typical hedge-fund arrangements) that the fund closes to new and existing investors after raising the initial capital. Reinvested capital gains are consequently the sole source of additional equity capital in the second period.

At time t_2, the price of the asset either reverts to V or it does not. If the asset price is V at t_2, the arbitrageur liquidates the fund and holds cash until t_3. If the asset price does not equal V at t_2, the arbitrageur invests aggressively—albeit not all of the fund's cash—in the underpriced asset. This portfolio then generates a risk-free return because the asset price rises to V at t_3 with certainty.

The arbitrageur's (that is, the hedge fund's) demand for the asset at the interim date, t_2, is given by

$$(2)\qquad QA_2 = \frac{D_2}{p_2},\ 0 \le D_2 \le F_2,$$

where D_2 is the amount of the hedge fund's demand in dollars. The amount $F_2 - D_2 \ge 0$ is held in cash. Because total demand aggregated across noise traders and the arbitrageur must equal the asset supply of one unit ($QN_2 + QA_2 = 1$), the price of the financial asset at t_2 is determined by combining (1) and (2):

$$(3)\qquad p_2 = V - S_2 + D_2,\ \ 0 \le D_2 < S_2.$$

The condition $D_2 < S_2$ implies that the asset still trades at a discount to the intrinsic value at t_2: $p_2 < V$. This assumption recognizes the arbitrageur's incentive *not* to bid up the price all the way to intrinsic value immediately—an implication of the fact that the arbitrageur will be compensated during the last period for achieving a positive return on investment.

As shown by Grossman and Vila (1992), the arbitrageur does not want to invest all of F_1 in the asset at t_1, either. After all, the asset may become even more underpriced at t_2, in which event he will want to increase his investment ("double up"). With D_1 denoting the amount the arbitrageur invests in the asset at t_1, we have

$$(4)\qquad QA_1 = \frac{D_1}{p_1},$$

which implies the initial asset price will be

[5] Amin and Kat (2001) report that hedge funds typically charge their investors a fixed annual "management fee" of 1 or 2 percent of assets under management plus an "incentive fee" of 15 to 25 percent of the fund's realized annual return. The incentive fee is waived if a particular "hurdle rate" has not been achieved, which can be a fixed number or a reference rate such as the T-bill rate plus or minus a spread. Most funds also apply a "high-water mark" provision, which requires the fund to make up any past losses before the incentive fee is paid. We assume the hurdle rate is zero, and we exclude any high-water mark for the sake of simplicity.

$$p_1 = V - S_1 + D_1,\ D_1 < S_1. \tag{5}$$

The condition $D_1 < S_1$ implies $p_1 < V$, which again captures the fact that the arbitrageur will not bid the price all the way up to the asset's intrinsic value because of the incentives built into his compensation schedule.

The investors have prior beliefs about the arbitrageur's talent in exploiting possible asset mispricing, but are not perfectly informed. Investors update their beliefs about the arbitrageur's talent using a simple Bayesian learning rule, which is based solely on the arbitrageur's past performance. When past returns are poor, investors don't know for sure whether the poor returns are due to a random error (noise), a deepening of noise trader misperception (bad luck), or truly inferior investment talent. Pulling some of their money from the hedge fund after the asset mispricing has deepened—that is, when the expected return on the long-short portfolio is highest—is the investor's rational response to the problem of inferring the arbitrageur's (unobservable) talent from data that are ambiguous (that is, observationally equivalent under more than one possible economic structure).

The investor's rule of updating his beliefs about the arbitrageur's talent implies that, if the hedge fund loses money during the first period, the fund faces withdrawals at the interim date, t_2. Specifically, we assume that the withdrawals at t_2 are a multiple of the hedge fund's posted gross return (that is, before management fees) at t_2, denoted R_2, should this return be negative. Remember that, while investors can withdraw capital, they cannot inject additional funds. Thus, the supply of funds in the second period is the following[6]:

$$F_2 = \begin{cases} F_1 \cdot \alpha \cdot (1 + R_2) + F_1 \cdot (1 - \alpha) \cdot (1 + R_2)^{\gamma}, \gamma > 1, & \text{if } -1 \le R_2 < 0 \\ F_1 \cdot (1 + R_2), & \text{if } \ R_2 > 0, \end{cases} \tag{6}$$

where γ is a parameter that determines the responsiveness of the investor to past performance. For $\gamma = 1$, poor first-period returns do not shake the confidence of investors in the arbitrageur's talent. At the other extreme, responsiveness that becomes unboundedly large implies that even a small first-period loss is multiplied into a huge withdrawal of funds. Note that the outside investors may withdraw only what is theirs. This means that, even if the outsiders pull all of their money, the arbitrageur's equity stake remains and the fund can stay in business. The arbitrageur knows that—despite a temporary deepening of the mispricing—the price of the asset will revert to intrinsic value at t_3 for certain, so he will keep his own money invested, come what may.

Our multiplicative feedback rule provides the arbitrageurs with what may be a more realistic incentive structure than the linear feedback rule in Shleifer and Vishny (1997). Our feedback rule does not penalize small negative returns quite as severely for a high degree of responsiveness, γ, as is the case in Shleifer and Vishny. For a responsiveness coefficient of $\gamma = 5$, for instance, a gross return in the first period, R_2, of –1 percent reduces the fund's equity capital by approximately 4 percentage points. A 5 percent loss, on the other hand, leaves the fund with approximately 77 percent of its equity capital at the beginning of the next period. We provide more results from the model below.

The gross return of the hedge fund in the first period, R_2, is given by

$$R_2 = \frac{(F_1 - D_1) + D_1 \cdot \dfrac{p_2}{p_1} - F_1}{F_1} = \frac{D_1 \cdot \dfrac{p_2 - p_1}{p_1}}{F_1}. \tag{7}$$

The fund's first-period return consists of its return on the financial asset, normalized by the total funds available for investment.

For simplicity, we assume a specific form of uncertainty about noise trader sentiment at t_2, S_2. With probability $1 - q$ $(0 < q < 1)$, noise traders recognize the true value of the asset, which implies $S_2 = 0$. In this case, the arbitrageur liquidates at t_2 and holds cash until t_3. Then the arbitrageur's assets under management at t_3 would amount to

$$F_3^{S_2=0} = F_2^{S_2=0} \equiv F_1 \cdot (1 + R_2^{S_2=0}), \tag{8}$$

where

$$R_2^{S_2=0} = \frac{(F_1 - D_1) + D_1 \cdot \dfrac{V}{p_1} - F_1}{F_1}.$$

On the other hand, noise trader misperception deepens to S_2 with probability q, $S_2 = S > S_1 (> 0)$. If noise traders continue to misperceive the intrinsic value of the asset, the hedge fund's assets at t_3 will amount to the following:

[6] Some hedge funds have "lock-up" periods of one to three years, while others allow investors to withdraw money with only a few weeks' notice. As a result of the poor quality of investors' information about the arbitrageur's talent, the arbitrageur's past performance often is a major determinant of the resources he receives to manage, regardless of the actual arbitrage opportunities available to him.

$$(9)\qquad F_3^{S_2=S} = \frac{V}{p_2^{S_2=S}} \cdot D_2 + (F_2^{S_2=S} - D_2)$$

$$= \frac{V}{p_2^{S_2=S}} \cdot D_2 + F_1 \cdot (1 + R_2^{S_2=S})^{\gamma} - D_2,$$

where

$$R_2^{S_2=S} = \frac{(F_1 - D_1) + D_1 \cdot \dfrac{p_2^{S_2=S}}{p_1} - F_1}{F_1}.$$

THE ARBITRAGEUR'S OPTIMIZATION PROGRAM

The arbitrageur's total income consists of management fees and capital gains on reinvested management fees. The expected value of the management fees, MF, equals the sum of the expected values of the management fees collected at $t_1(MF_1)$, at $t_2(MF_2)$, and at $t_3(MF_3)$. The expected value of the capital gains is CG. The arbitrageur's maximization problem therefore is

$$(10)\qquad \underset{D_1, D_2}{Max}\ \{MF_1 + MF_2 + MF_3 + CG\},$$

where the management fees are

$$(11)\qquad MF_1 = \alpha \cdot F_1,$$

$$(12)\qquad MF_2 = MF_2^{S_2=S} + MF_2^{S_2=0},\ \text{and}$$

$$(13)\qquad \begin{aligned} MF_3 &= \beta \cdot q \cdot R_3^{S_2=S} \\ &\cdot (F_2^{S_2=S} - (1 + R_2^{S_2=S}) \cdot MF_1 - MF_2^{S_2=S}) \end{aligned}$$

and where

$$\begin{aligned} MF_2^{S_2=S} = \alpha \cdot q \cdot (F_2^{S_2=S} &- \alpha \cdot [1 + R_2^{S_2=S}] \cdot F_1 \\ &- \beta \cdot \max\{0, R_2^{S_2=S}\} \cdot (1 - \alpha) \cdot F_1) \\ &+ \beta \cdot q \cdot \max\{0, R_2^{S_2=S}\} \cdot (1 - \alpha) \cdot F_1, \end{aligned}$$

$$\begin{aligned} MF_2^{S_2=0} = \alpha \cdot (1 - q) \cdot (F_2^{S_2=0} &- \alpha \cdot [1 + R_2^{S_2=0}] \cdot F_1 \\ &- \beta \cdot R_2^{S_2=0} \cdot (1 - \alpha) \cdot F_1) \\ &+ \beta \cdot (1 - q) \cdot R_2^{S_2=0} \cdot (1 - \alpha) \cdot F_1,\ \text{and} \end{aligned}$$

$$R_3^{S_2=S} = \frac{(F_2^{S_2=S} - D_2^{S_2=S}) + D_2^{S_2=S} \cdot \dfrac{V}{p_2^{S_2=S}} - F_2^{S_2=S}}{F_2^{S_2=S}}.$$

The quantity $MF_2^{S_2=S}$ represents the income the arbitrageur collects at t_2 should the noise traders' misperception deepen in the first period, while $MF_2^{S_2=0}$ is the fee income if the asset price reverts to intrinsic value. The arbitrageur also captures capital gains on the equity he builds from the reinvested management fees. The expected value of the capital gains, CG, equals

$$(14)\qquad \begin{aligned} CG = q \cdot (R_2^{S_2=S} \cdot MF_1 \cdot [1 + R_3^{S_2=S}] + R_3^{S_2=S} \cdot MF_2^{S_2=S}) \\ + (1 - q) \cdot R_2^{S_2=0} \cdot MF_1. \end{aligned}$$

The arbitrageur's choice variables are $D_1(\leq F_1)$ and $D_2(\leq F_2^{S_2=S})$, which are the amounts the arbitrageur invests in the asset at t_1 and t_2, respectively. Unless the asset reverts to intrinsic value at $t_2(p_2 = V)$, the t_2 price of the asset given in equation (3) is a function of the t_2 choice variable, D_2. Similarly, the t_1 price of the asset given in equation (5) is a function of the choice variable, D_1.

SOLUTION TO THE MAXIMIZATION PROBLEM

We solve the maximization problem numerically. We hold constant all of the following: $V = 1$; $F_1 = S_1 = 0.2$; $S_2 = 0.4$; $q = 1 - q = 0.5$; $\alpha = 0.02$; and $\beta = 0.2$. Note that $F_1 = S_1 = 0.2$ means that the arbitrageur has sufficient buying power to eliminate the t_1 mispricing entirely if so desired. Also, note that $0.4 = S_2 > S_1 = 0.2$ means that noise trader misperception may deepen between t_1 to t_2—that is, the asset may become even more mispriced. For the values chosen for S_1, S_2, and q, noise trader misperception, S, is as likely to double as it is to vanish. Thus, the expected value of noise trader misperception in the second period, $q \cdot S_2$, equals the noise trader misperception observed in the first period, S_1.

We vary γ, the responsiveness to past performance of fund withdrawals, from $\gamma = 1$ (no responsiveness by the investors to past investment performance, that is, no withdrawals) to $\gamma = 20$ (extreme responsiveness) with a step length of unity. We use a grid search method to solve the maximization problem. This involves varying D_1 and D_2 independently in very small increments within their bounds, $0 \leq D_i \leq F_i\ (i = 1,2)$, to find the maximum of the objective function.

The findings of the grid search are displayed in Figures 2 through 5. The first important point to make concerns the extent to which the presence of the hedge fund affects asset mispricing. Figure 2 shows that the mispricing is less pronounced in each period than it would be without the hedge fund.

Figure 2

Effect of Investor Responsiveness on Asset Prices

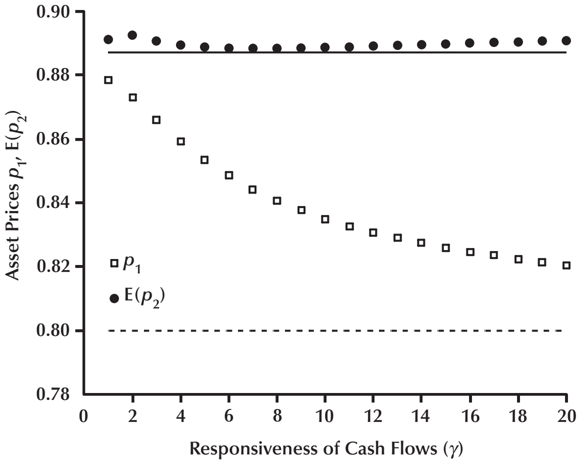
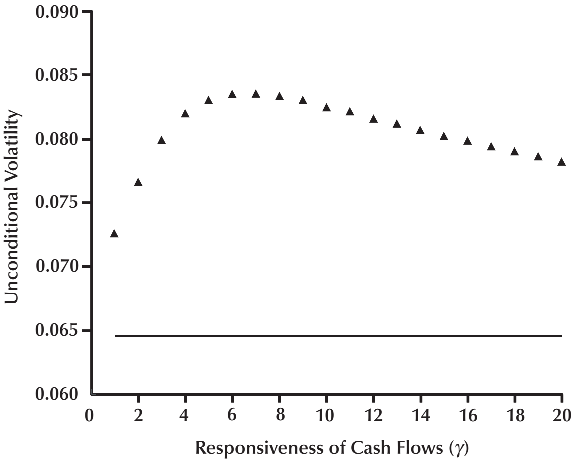


Figure 3

Effect of Investor Responsiveness on Asset Price Volatility



Remember that, without arbitrage, the first-period price, p_1, and the expected value of the second-period price, E[p_2], both would equal 0.8 (shown as a dashed line). On the other hand, without noise traders, the asset would trade at unit value in both periods (not shown). The hedge fund almost halves the difference between the expected value of the second-period price, E[p_2] (shown as solid circles), and the asset's intrinsic, unit value. In fact, the degree of investor responsiveness, γ, has little bearing on E[p_2], which approaches the value of approximately 0.8873 (shown as a solid horizontal line) as γ approaches infinity. By comparison, the degree of responsiveness has a strong impact on the first-period price, p_1 (shown as open boxes). This is because the arbitrageur treads even more cautiously when putting on this trade in the first period when he knows that the investors penalize negative returns with sizeable withdrawals. In fact, the higher is γ, the more cash the arbitrageur holds in the first period, and therefore, the lower is p_1. As the degree of investor responsiveness, γ, goes to infinity, the amount the arbitrageur invests in the first period goes to zero and, consequently, the first-period price, p_1, converges to 0.8—the value the asset would adopt if there were no hedge fund in the market (shown as a dashed line). Thus we conclude that the hedge fund pushes the price of the asset (or its respective expected value) toward the intrinsic, unit value in both periods. This is our first main finding.

Figure 3 shows the unconditional volatility of the asset's returns for various degrees of investor responsiveness, γ. The unconditional volatility is calculated as the expected value of the squared returns over the two periods. For low values of investor responsiveness, volatility increases as γ increases. For high values of responsiveness, a further increase in γ reduces volatility monotonically. As γ goes to infinity, volatility approaches a level (as shown by the solid line) that is lower than the volatility level at $\gamma = 1$ (as indicated by the leftmost symbol), which is the benchmark case of unwavering investor confidence in the hedge fund manager. The reason for this "volatility hump" lies in the existence of two opposite effects. All else equal, the higher γ is, the bigger is the drop in the asset's price from t_1 to t_2 should the noise traders' misperception deepen. On the other hand, the higher γ is, the lower is the price of the asset at t_1 because the arbitrageur puts less money to work. For low values of investor responsiveness, the volatility-increasing effect dominates. For increasingly higher values of γ, this effect becomes progressively weaker until it vanishes for an infinitely large degree of investor responsiveness.

It is important to note that the hedge fund greatly reduces asset price volatility, regardless of the degree of investor responsiveness. The unconditional volatility without the hedge fund runs at 0.5694 (not shown), which is a multiple of the volatility that we observe even at the degree of responsive-

Figure 4

Effect of Investor Responsiveness on Asset Return When Misperception Deepens

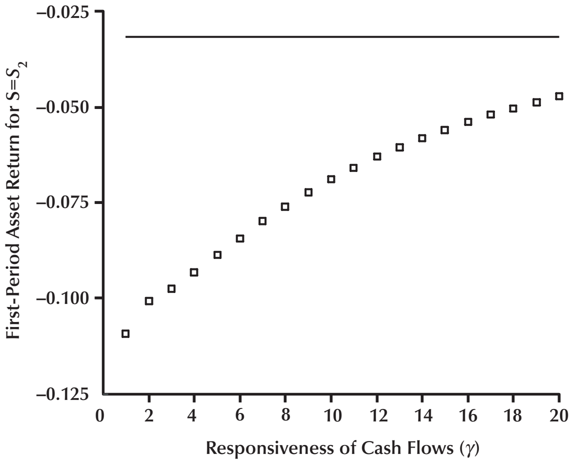
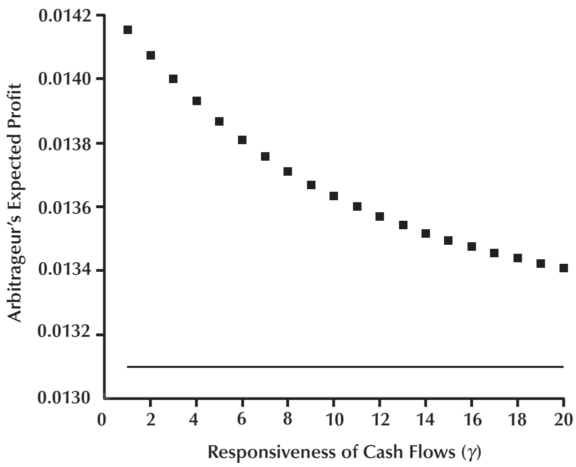


Figure 5

Effect of Investor Responsiveness on Arbitrageur's Profit



ness that generates the highest level of volatility. Thus, we conclude that the hedge fund unambiguously reduces unconditional volatility. This is our second main finding.

Another way to look at the impact of arbitrage on volatility is to ask how the market behaves when asset mispricing deepens. Such an event—if severe—might cause, or occur alongside, a financial crisis. Figure 4 shows, for the case of a deepening noise trader misperception of the asset's intrinsic value, the first-period asset return as a function of investor responsiveness. The absolute value of the percentage decline of the asset price increases with investor responsiveness, γ. For an infinitely high value of γ, the arbitrageur holds cash in the first period and then invests aggressively at t_2, although he does not invest all the cash available. The horizontal line in Figure 4 signifies the first-period return for this borderline case of an infinite degree of responsiveness. Note that, without a hedge fund, the first-period return would amount to a negative 25 percent (not shown), which is more than twice as much (in absolute value) as what is observed even with a degree of responsiveness of zero (that is, γ equal to one). Hence, we conclude that the presence of a hedge fund dampens volatility in the event of a deepening of noise trader misperception, as might occur in a financial panic. This is our third main finding.

Finally, we are interested in the question of how investor responsiveness affects the arbitrageur's profit, that is, his incentive to set up a hedge fund and engage in arbitrage. Figure 5 shows the arbitrageur's profit as a function of γ. Not surprisingly, the profit of the arbitrageur decreases monotonically with increased investor responsiveness to past performance. The monotonic decline in the profitability of arbitrage with increasing investor responsiveness to past performance is a manifestation of the fact that liquidating a hedge portfolio when the expected return from arbitrage is highest is counterproductive—that is, it runs against "the nature of the trade."

CONCLUSION

Even financially constrained professional arbitrageurs may be able to exploit asset mispricing if they can link up with rational but uninformed investors. To achieve this goal, the two parties must overcome—at least to a degree—the problem of asymmetric information about the arbitrageur's talent. The result of such an endeavor is a hedge fund that goes long on (comparatively) underpriced assets and short on (comparatively) overpriced assets. As a byproduct, the impacts of noise trader misperceptions on asset prices and volatility are reduced. This holds for any degree of responsiveness to past performance ("feedback") of the investors' confidence in the arbitrageur's talent.

This article builds on the dynamic-investment literature that reaches back at least to Merton (1971).

Shleifer and Vishny (1997) provided an insightful model of wealth-constrained arbitrageurs that can be, and has been, extended in several directions. We add several realistic features to the professional arbitrageur's problem in the canonical model, including the ability to build an equity stake in his hedge fund over time, and a potentially more realistic multiplicative (rather than linear) investor feedback rule. Like Shleifer and Vishny, we assume that the hedge fund can influence the market price. Hedge funds do, in fact, sometimes move market prices because they operate in specialized market segments that have limited liquidity. It is also true, however, that hedge funds alone cannot prevent asset-price volatility or occasional mispricing—which might deepen before it eventually corrects.

REFERENCES

Amin, Guarav S. and Kat, Harry M. "Hedge Fund Performance 1990-2000: Do the 'Money Machines' Really Add Value?" Working paper, University of Reading (UK), 15 May 2001.

Breedon, Douglas T. "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities." *Journal of Financial Economics*, September 1979, *7*(3), pp. 265-96.

Campbell, John Y.; Lo, Andrew and MacKinlay, A. Craig. *The Econometrics of Financial Markets*. Princeton, N.J.: Princeton University Press, 1997.

__________ and Viceira, Luis M. "Consumption and Portfolio Decisions When Expected Returns Are Time Varying." *Quarterly Journal of Economics*, May 1999, *114*(2), pp. 433-95.

Fama, Eugene F. "The Behavior of Stock Market Prices." *Journal of Business*, 1965, *38*, pp. 34-105.

Friedman, Milton. "The Case for Flexible Exchange Rates," in *Essays in Positive Economics*. Chicago: University of Chicago Press, 1953.

Gromb, Denis and Vayanos, Dimitri. "Equilibrium and Welfare in Markets with Financially Constrained Arbitrageurs." Working paper, London Business School and Massachusetts Institute of Technology, October 2001.

Grossman, Sanford J. and Vila, Jean-Luc. "Optimal Dynamic Trading with Leverage Constraints." *Journal of Financial and Quantitative Analysis*, June 1992, *27*(2), pp. 151-68.

__________ and Zhou, Zhongquan. "Equilibrium Analysis of Portfolio Insurance." *Journal of Finance*, September 1996, *51*(4), pp. 1379-1403.

Keynes, John Maynard. *The General Theory of Employment, Interest, and Money*. New York: Harcourt, Brace & World, 1936.

Kyle, Albert S. and Xiong, Wei. "Contagion as a Wealth Effect." *Journal of Finance*, August 2001, *56*(4), pp. 1401-40.

Merton, Robert C. "Optimal Consumption and Portfolio Rules in a Continuous-Time Model." *Journal of Economic Theory*, December 1971, *3*(4), pp. 373-413.

__________. "An Intertemporal Capital Asset Pricing Model." *Econometrica*, September 1973, *41*(5), pp. 323-61.

Muth, J. "Optimal Properties of Exponentially Weighted Forecasts." *Journal of the American Statistical Association*, 1960, *55*, pp. 299-306.

Samuelson, Paul. "Proof That Properly Anticipated Prices Fluctuate Randomly." *Industrial Management Review*, 1965, *6*, pp. 41-49.

Scholes, Myron S. "Crisis and Risk Management." *American Economic Review Papers and Proceedings*, May 2000, *90*(2), pp. 17-21.

Sharpe, William P. and Alexander, Gordon. *Investments*. 4th Edition. Englewood Cliffs, N.J.: Prentice Hall, 1990.

Shleifer, Andrei and Vishny, Robert W. "The Limits of Arbitrage." *Journal of Finance*, March 1997, *52*(1), pp. 35-55.

Xiong, Wei. "Convergence Trading with Wealth Effects: An Amplification Mechanism in Financial Markets." *Journal of Financial Economics*, November 2001, *62*(2), pp. 247-92.

Regime-Dependent Recession Forecasts and the 2001 Recession

Michael J. Dueker

Business recessions, as a major source of nondiversifiable risk, impose high costs on society. Since firms cannot obtain "recession insurance," they try to foresee recessions and reduce their exposure ahead of time. Consequently, forecasting business cycle turning points has remained an important endeavor. Of course, the difficulty is deriving reliable methods to forecast business cycle turning points. Previous studies found that accurate recession forecasts remain elusive (Filardo, 1999; Del Negro, 2001; Chin, Geweke, and Miller, 2000). Forecasts of economic output are not a good substitute for forecasts of business cycle turning points because less than 20 percent of all months pertain to recessions. Hence, empirical models of output growth focus largely on explaining variation in output growth during economic expansions, since this variation accounts for the lion's share of the sample variance.

Throughout the 1990s, recession forecasting models relied exclusively on the 1990-91 recession for out-of-sample confirmation (Estrella and Mishkin, 1998; Birchenhall et al., 1999; Friedman and Kuttner, 1998). Out-of-sample confirmation is particularly important for recession forecasting because recessions are infrequent, making it tempting to overfit specific episodes in sample. In general, recession forecasting models failed to predict the 1990-91 recession out of sample. The occurrence of the 2001 recession raises the question: Was the 1990-91 recession uniquely difficult to predict or is recession forecasting a failed enterprise? If recession forecasting models were to repeat in 2001 their dismal performance from the 1990-91 recession, then doubts about such models would mount with justification. In this article, I examine the out-of-sample forecasts from recession forecasting models with three levels of sophistication. All three models concur with the previous finding that the 1990-91 recession was hard to predict. The simplest of the three models largely misses the 2001 recession, but two regime-switching models come quite close to predicting the onset date of the recession six months ahead of time. One innovation to recession forecasting introduced here is to allow the critical probability level for a recession to be predicted to depend on the current state of a Markov switching process—hence, regime-dependent recession forecasts.

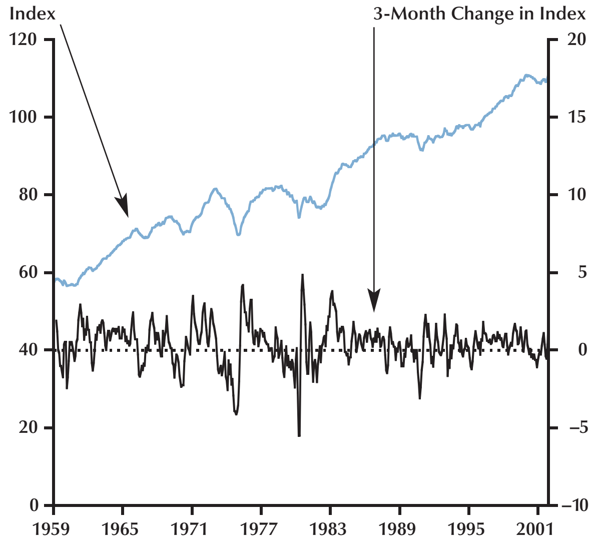
In this way, the forecasts presented here respond to the criticism that economists equivocate too much when it comes to their recession forecasts. When recession forecasts are expressed as probability statements, it is tempting to claim ex post that the ex ante probability of recession from the forecasting model was "close enough" to either one or zero to be considered a correct forecast. For example, if the model suggested that a recession would occur with a probability of 35 percent, then after the fact the model builder could try to justify either outcome: If a recession ensued, the model builder could cite the jump in probability from zero to 35 percent as evidence that a large shift toward recession was detected; if no recession ensued, the model builder could say that the 35 percent probability was far from 100 percent and did not indicate recession. To avoid such ambiguity, economists are often asked to make specific calls as to whether the economy will or will not be in recession six months from now. A yes/no recession signal comes from comparing the forecasted probability of recession with a critical value determined prior to the out-of-sample forecasting.

The dependent variable I use to separate business recessions from expansions is based on the business cycle turning points defined by the National Bureau of Economic Research (NBER). As in Birchenhall et al. (1999), I use the composite index of leading indicators (CLI) from the Conference Board as the explanatory variable in the probit forecasting models. The CLI receives much attention as a harbinger of future business cycle conditions. It has ten components: manufacturing hours, consumer expectations, stock prices, initial unemployment claims, building permits, the money supply, the spread between short- and long-term interest rates on government securities, vendor performance, manufacturing orders for consumer goods, and manufacturing orders for capital goods. Figure 1 shows the upward trend in the CLI and its three-month percentage changes. The decreased volatility in the three-month change since the early 1980s is symptomatic of decreased volatility in the business cycle.

Michael J. Dueker is a research officer at the Federal Reserve Bank of St. Louis. Mrinalini Lhila and John Zhu provided research assistance.

Figure 1

Index of Leading Indicators



The 1990-91 and 2001 recessions can be used to evaluate the out-of-sample performance of the model and the leading indicators. I find that the 1990-91 recession is the anomaly in that the recession signal emanating from the leading indicators largely misses it. The 2001 recession, in contrast, was largely predictable from the leading indicators six months ahead.

The probit model takes the monthly movements in the leading indicators and translates them into precise probability statements regarding the likelihood of recession. When dealing with qualitative events such as recessions, however, it is often desirable to define a specific recession yes/no signal from the probit probabilities. That way, the forecasting method either correctly "calls" the recession or it does not. Birchenhall et al. (1999) use this approach to say that a recession is signaled if the logit probability of recession exceeds a critical value. If one chooses the critical value to maximize the in-sample fit, then out-of-sample confirmation of recession forecasting models becomes particulary important, given Robert Lucas's dictum: Beware of econometricians bearing free parameters. The critical value that defines a recession signal is an example of what Lucas calls a free parameter because—although it is not an inherent part of the econometric model—it combines with the model estimates to suggest that the model fits the data well. Out-of-sample confirmation helps ensure that the free parameters are not simply overfitting the in-sample data.

THE SIMPLE PROBIT MODEL

I study three probit forecasting methods. One is the simple probit model and the other two are based on a Markov switching probit from Dueker (1997). A probit model can be used to predict a qualitative variable such as a recession indicator, R_t, where

$R_t = 1$ if the economy is in NBER recession in period t

$= 0$ otherwise.

One way to think of a probit model is to assume that a normally distributed latent variable, R^*, lies behind the recession indicator:

$$(1) \qquad R_t^* = -c_0 + -c_1 X_{t-k} + u_t ,$$

where u is a normally distributed error team and X is the leading indicator explanatory variable lagged k periods—the forecasting horizon. A probability of recession is associated with each possible value of the latent variable, where the latent variable is assumed to be negative during expansions and positive during recessions. In this case, the forecasted probability of recession is

$$(2) \qquad \text{Prob}\left(R_t = 1\right) = 1 - \Phi\left(c_0 + c_1 X_{t-k}\right),$$

where $\Phi(.)$ is the cumulative standard normal density function.

The log-likelihood function for a simple probit model is

$$(3) \qquad \begin{aligned} L = \sum_t R_t \ln \text{Prob}\left(R_t = 1 \,|X_{t-k}\right) \\ + \left(1 - R_t\right) \ln \text{Prob}\left(R_t = 0 \,|X_{t-k}\right). \end{aligned}$$

COEFFICIENT VARIATION VIA MARKOV SWITCHING IN THE PROBIT MODEL

The log-likelihood function in equation (3) highlights the assumption in the probit model that the recession outcomes conditional on available information are independently distributed from month to month. This assumption is questionable unless the econometric model allows for considerable serial correlation in the recession probabilities. As in Dueker (1997), one way to achieve this degree of serial correlation is to introduce serial correlation in the model's coefficients by making them subject to Markov switching.

The simplest interpretation of Markov switching coefficients in a probit model is that they capture time variation in the variance of the error term u from equation (1). In the low-variance regime the variance would still be normalized to one, but in the high-variance regime the variance would be greater than one:

$$\text{Prob}\left(R_t = 1 \mid \text{High Variance}\right)$$
$$= 1 - \Phi\left(c_0 / \sigma + c_1 / \sigma X_{t-k}\right), \sigma > 1.$$

Variance switching implies that the coefficients are restricted to change by the same percentage between regimes. I do not impose this condition because I do not want to restrict the signs of the coefficients, for example, to be the same in both regimes. Nevertheless, conditional heteroskedasticity helps motivate why the coefficients might be subject to regime switching, since volatility is one aspect of the economy that does vary across the business cycle. Because it is not the only aspect of the economy that varies across the business cycle, however, we keep the model more general by not restricting the regime switching to variance switching.

In a Markov switching model, the parameters change values according to an unobserved binary state variable, S_t, which follows a Markov process:

(4)
$$S_t \text{ equals } 0 \text{ or } 1$$
$$\text{Prob } (S_t = 0 \mid S_{t-1} = 0) = p$$
$$\text{Prob } (S_t = 1 \mid S_{t-1} = 1) = q.$$

In this way, the coefficients take on either of two values and thereby change the magnitude of the shock needed to induce a recession:

(5)
$$R_t = 1 \quad \text{if} \quad u_t > c_0(S_t) + c_1(S_t) + c_1(S_t)X_{t-k}$$
$$= 0 \quad \text{if} \quad u_t \le c_0(S_t) + c_1(S_t)X_{t-k}\ .$$

The transition probabilities, p and q, indicate the persistence of the states and determine the unconditional probability of the state $S_t = 0$ to be $(1-q)/(2-p-q)$. Since the state is unobserved and must be inferred as a probability, allowance for two states means that the expected values of the coefficients can lie anywhere between the high and low values corresponding to the two states. In the estimation, Bayes' rule is used to obtain filtered probabilities of the states in order to sum over possible values of the unobserved states and evaluate the likelihood function, as in Hamilton (1990):

(6)
$$\text{Prob}\left(S_t = 0 \middle| R_t = 0, X_{t-1}\right) = \frac{\text{Prob}\left(S_t = 0 \middle| R_{t-1}, X_{t-2}\right) \text{Prob}\left(R_t = 0 \middle| S_t = 0,\ X_{t-1}\right)}{\Sigma_{s=0}^{1} \text{Prob}\left(S_t = s \middle| R_{t-1}, X_{t-2}\right) \text{Prob}\left(R_t = 0 \middle| S_t = s,\ X_{t-1}\right)}$$

(7)
$$\text{Prob}\left(S_t = 0 \middle| R_{t-1},\ X_{t-2}\right)$$
$$= p \text{ Prob}\left(S_{t-1} = 0 \middle| R_{t-1},\ X_{t-2}\right)$$
$$+ (1-q)\text{Prob}\left(S_{t-1} = 1 \middle| R_{t-1}, X_{t-2}\right).$$

The probability in equation (7) is called the one-period-ahead prior probability because it is not conditional on the recession outcome at time t. For a forecast horizon of several months, we need to use the transition probabilities to derive a k-period-ahead probability of the state variable:

(8)
$$\begin{pmatrix} \text{Prob}\left(S_t = 0 \middle| R_{t-k},\ X_{t-k}\right), \\ \text{Prob}\left(S_t = 1 \middle| R_{t-k},\ X_{t-k}\right) \end{pmatrix}' =$$
$$G^k \begin{pmatrix} \text{Prob}\left(S_{t-k} = 0 \middle| R_{t-k},\ X_{t-k}\right), \\ \text{Prob}\left(S_{t-k} = 1 \middle| R_{t-k},\ X_{t-k}\right) \end{pmatrix}',$$

where G is the transition matrix of the Markov state variable.

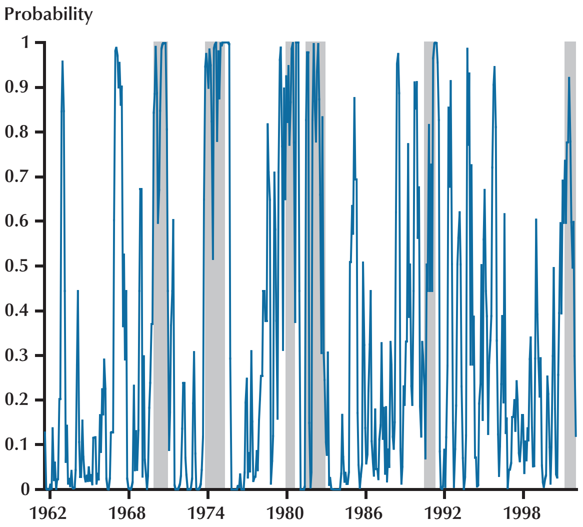
For forecasting k periods ahead, one finds the best-fitting model by maximizing the corresponding likelihood function:

(9)
$$\sum_{t=1}^{T} R_t \ln \begin{pmatrix} \sum\limits_{s=0}^{1} \text{Prob}\left(S_t = s \middle| R_{t-k}, X_{t-k}\right) \\ \text{Prob}\left(R_t = 1 \middle| S_t = s,\ X_{t-k}\right) \end{pmatrix}$$
$$+ \left(1 - R_t\right) \ln \begin{pmatrix} \sum\limits_{s=0}^{1} \text{Prob}\left(S_t = s \middle| R_{t-k}, X_{t-k}\right) \\ \text{Prob}\left(R_t = 0 \middle| S_t = s,\ X_{t-k}\right) \end{pmatrix}.$$

In this forecasting exercise, the forecaster is assumed to know whether the economy is currently in recession when forecasting whether the economy will be in recession six months from now. This assumption is somewhat problematic when forecasting from the early stages of a recession before the NBER has officially declared that the economy entered a recession. For example, when forecasting whether the economy would be in recession in October 2001, it was probably not clear that the economy was in recession in April 2001. The NBER did not announce that the recession had started in March 2001 until November 26, 2001. On the other hand, forecasts of the onsets of recessions are not likely to be clouded by this assumption. In forecasting whether the economy would be in recession in

Figure 2

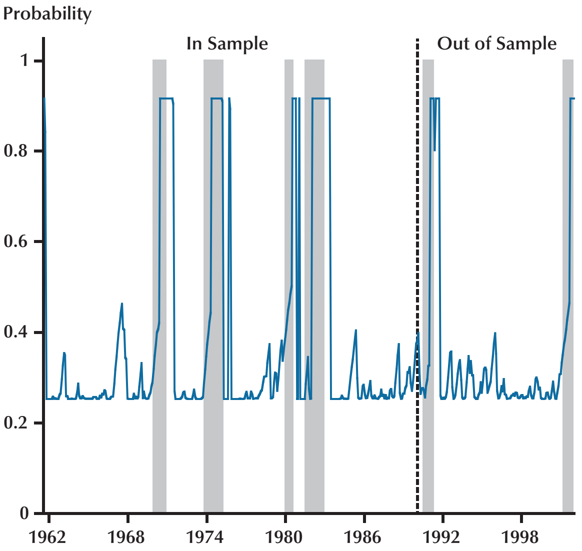
Probability of Recession Conditional on Regime *S*=0



NOTE: Shaded bars indicate recessions.

Figure 3

Six-Month-Ahead Probability of Regime *S*=0



NOTE: Shaded bars indicate recessions.

Table 1

Coefficients from the Markov Switching Probit

Equation	Coefficient value
c_0 (S=0)	0.140 (0.233)
c_1 (S=0)	1.415 (0.336)
c_0 (S=1)	2.397 (0.730)
c_1 (S=1)	–0.062 (0.351)
p	0.950 (0.012)
q	0.984 (0.010)

NOTE: Standard errors are in parentheses.

March or April 2001, no one believed in September or October 2000 that the economy was already in recession. There was no confusion about the current state of the economy. Similarly, in January 1990 everyone knew that the economy was still in an expansion when forecasting whether a recession would start by July 1990. For this reason, one might pay special attention to how the model predicts the onset dates of recessions.

FORECAST RESULTS

The Markov switching probit model was estimated with data from May 1960 to December 1989. The explanatory variable, X, is the three-month percentage change in the leading indicators index shown in Figure 1. The minimum value of the change in the leading indicators index is about –5. If we plug this minimum value and the coefficient estimates from Table 1 into equation (5), we see that a recession is essentially never predicted in the state where $S=1$. Given the values of $c_0(S=1)$ and $c_1(S=1)$, a standard normal shock, u, greater than 2.7 would have to occur at the minimum value of the leading indicators for a recession to occur in that state. In contrast, recessions are often implied in the regime where $S=0$, as shown in Figure 2. (Note that in this and the following charts, the coefficient estimates are also applied to the out-of-sample data from January 1990 to December 2001.) The unconditional probability of the regime $S=0$, however, is only 0.25 and it is forecasted less often than regime $S=1$, as seen in Figure 3. Combining Figures 2 and 3, we see that six-month-ahead forecasts of a high probability of $S=0$ amount to a forecast of recession. Figure 4 combines the two explicitly by plotting the probability of recession after summing across the two states:

(10)

$$\text{Prob}\left(R_t=1\,\middle|\,X_{t-k}\right)$$
$$=\sum_{s=0}^{1}\text{Prob}\left(S_t=s\,\middle|\,R_{t-k},X_{t-k}\right)\text{Prob}\left(R_t=1\,\middle|\,S_t=s,\,X_{t-k}\right).$$

A standard approach—which I call model 1—to deriving explicit recession signals from the forecasted probability of recession, Prob $(R_t = 1|X_{t-k})$, is to choose a critical value, m, such that a recession is signaled if

$$\text{Prob}\left(R_t = 1 \middle| X_{t-k}\right) - m > 0. \quad (11)$$

A key innovation in this paper is to recognize that one can derive an alternative recession signal—called model 2—from regime-specific critical values, m_0 and m_1, such that a recession is signaled if

$$\begin{aligned} &\text{Prob}\left(S_t = 0 \middle| R_{t-k}, X_{t-k}\right) \\ &\qquad \left(\text{Prob}\left(R_t = 1 \middle| S_t = 0, X_{t-k}\right) - m_0\right) \\ &+ \text{Prob}\left(S_t = 1 \middle| R_{t-k}, X_{t-k}\right) \\ &\qquad \left(\text{Prob}\left(R_t = 1 \middle| S_t = 1, X_{t-k}\right) - m_1\right) > 0. \end{aligned} \quad (12)$$

Alternatively, one can rewrite this recession signal as

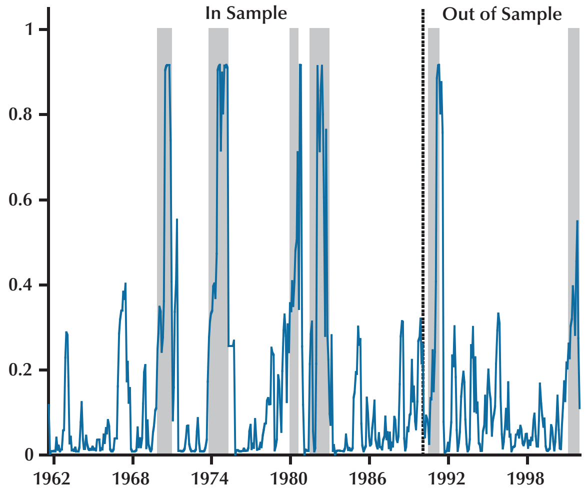
$$\begin{aligned} \text{Prob}\left(R_t = 1 \middle| X_{t-k}\right) > {} & \text{Prob}\left(S_t = 0 \middle| R_{t-k}, X_{t-k}\right) m_0 \\ & + \text{Prob}\left(S_t = 1 \middle| R_{t-k}, X_{t-k}\right) m_1. \end{aligned} \quad (13)$$

Either way, two critical values are used, where the weight given to each depends on the regime probabilities. As shown in Figure 2, the probability of recession is relatively high on average in the regime where $S = 0$, with an average probability of 0.34. In contrast, the average probability of recession in the regime where $S = 1$ is not much above zero. Given the difference between the average probabilities of recession in the two regimes, it seems desirable to have separate critical probability levels for each regime, as in equation (12).

I chose critical probability levels based on the in-sample period through 1989 and examined how well they work in the out-of-sample period. The criterion I used was the greatest number of correct signals, where one point was given to a correct signal during a recession and half a point to a correct signal during an expansion. This point scheme puts greater emphasis on not missing recessions versus supplying false recession signals during expansions. The impetus for this asymmetry in the point scheme is the belief that most firms would be more willing to pay for recession insurance than for a contract that would indemnify them in the case where the economy performed above expectations when a recession was forecast.

Figure 4

Probability of Recession from Six-Month-Ahead Forecasts



NOTE: Shaded bars indicate recessions.

For the in-sample period through 1989, I found that values of $m = 0.28$ (for model 1) and $m_0 = 0.43$ and $m_1 = 0.135$ (for model 2) gave the greatest number of correct signals. It makes sense that the optimal value of m would lie between optimal m_0 and m_1, since it is trying to fill both roles. The critical probability level m_0 lies above the average probability of recession conditional on $S = 0$ (0.34), so that one predicts a recession less than half of the time that $S = 0$.

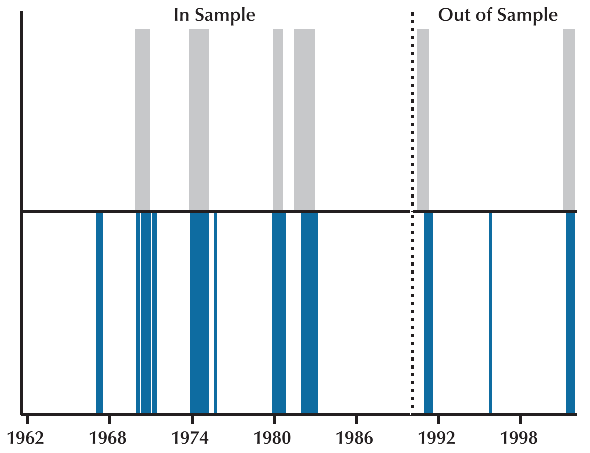
Figure 5 shows the fit of the recession-signaling model, where the signal is based on model 2 with the two regime-dependent critical values. With forecasts from model 2, recessions are generally not missed and the only notable false signal occurred in the 1966 slowdown. Figure 6, in turn, shows the fit of model 1—the signaling procedure that uses only one critical value, as in equation (11). Here, some of the recessions are projected to start earlier and end later than they did and there are many more false recession signals during expansions.

Similarly, Figure 7 shows the fit of the signal from the simple probit model from equation (2). The optimal critical value, m, for the 1960-89 period is 0.24. This approach, model 3, generated even more and longer-lasting false recession signals than the Markov switching model with one critical value (model 1). Based on these results, we do not look

Figure 5

Fit of Recession-Signaling Model

(Model 2: Two Critical Values)

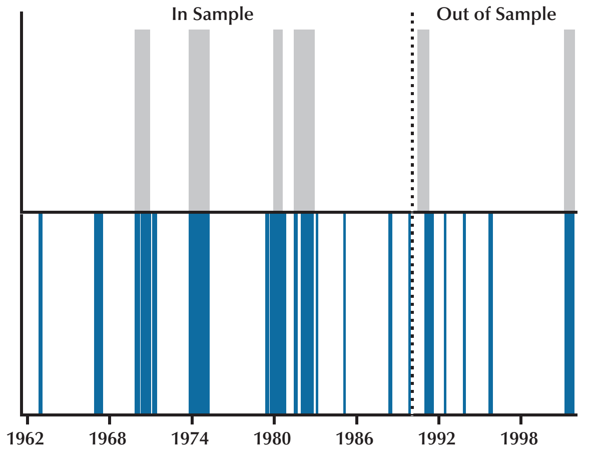


NOTE: Actual recessions (top) and six-month-ahead recession signals (bottom).

Figure 6

Fit of Recession-Signaling Model

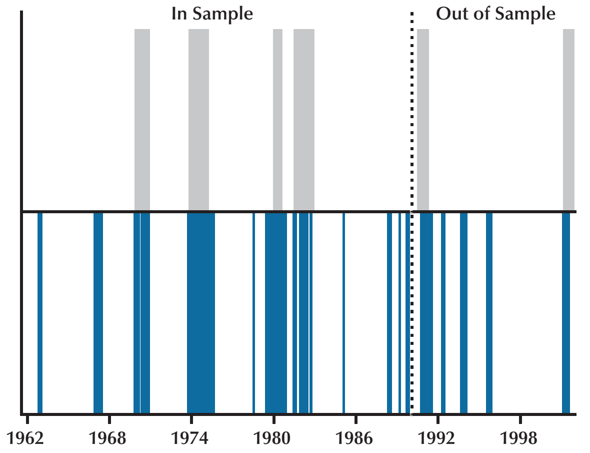
(Model 1: One Critical Value)



NOTE: Actual recessions (top) and six-month-ahead recession signals (bottom).

Figure 7

Fit of Recession Signal from Simple Probit Model (Model 3)

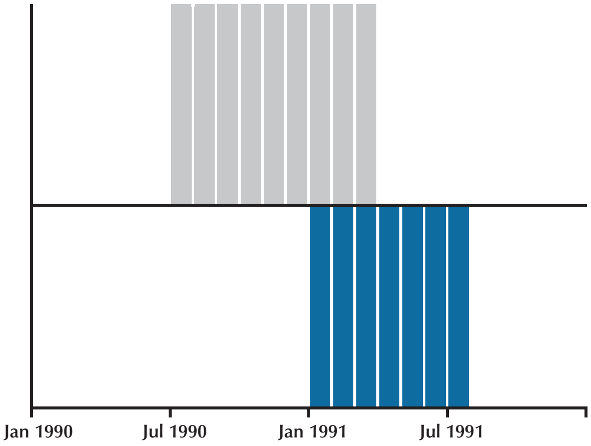


NOTE: Actual recessions (top) and six-month-ahead recession signals (bottom).

Figure 8

1990-91 Recession and Signal

(Signals from Models 1 and 2 Coincide: One or Two Critical Values)



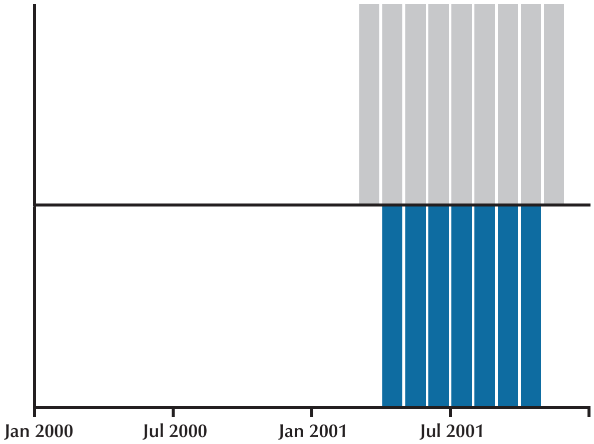
NOTE: Actual recession (top) and six-month-ahead recession signal (bottom).

further at the predictions from the simple probit model.

In comparing the two signals from the Markov switching models, a closer look at the two out-of-sample recessions will help determine whether the use of two critical values in model 2 as free parameters to fit the in-sample data resulted in overfitting. Figure 8 zooms in on the 1990 recession and shows that the recession signals from models 1 and 2 are identical and they both miss the 1990 recession in

Figure 9

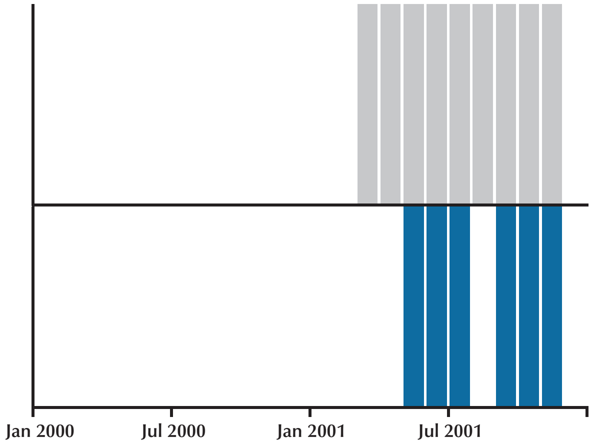
2001 Recession and Signal
(Model 1: One Critical Value)



NOTE: Actual recession (top) and six-month-ahead recession signal (bottom).

Figure 10

2001 Recession and Signal
(Model 2: Two Critical Values)



NOTE: Actual recession (top) and six-month-ahead recession signal (bottom).

the sense that the six-month-ahead signal does not kick in until at least six months too late. This result confirms previous findings that the 1990 recession was difficult to predict out of sample.

Fortunately, the performance of both signal approaches is better in the 2001 recession. Figure 9 shows that the signal from model 1 with one critical value predicted a recession onset in April 2001 using information through October 2000. Thus, this signal did not miss the actual onset date of March 2001 by much. The approach with two critical values—model 2—does slightly worse during the 2001 recession. Figure 10 shows that this signal needed data through November 2000 to predict a recession onset date of May 2001. In addition, it incorrectly projected August 2001 as an expansion month. Nevertheless, one has to keep Figures 5 and 6 in mind before concluding that the signal derived from one critical value is better out of sample than the signal derived from two critical values. Model 2 gave fewer false recession signals than model 1 during the long economic expansion of the 1990s, as seen by comparing Figures 5 and 6.

Looking out from the December 2001 data, both signaling approaches based on the Markov switching model—with either one or two critical values—predict that the recession would have ended by January 2002. Later, the NBER will officially date the end of the recession and then the accuracy of the model's trough prediction will be known.

SUMMARY AND CONCLUSIONS

This article looks at forecasting the 1990 and 2001 recessions out of sample and shows that 1990 appears to be an anamoly in terms of being difficult to predict. Thus, one should not conclude based on the 1990 recession that recession forecasting is a failed enterprise. This article also responds to the exhortation economists receive to provide unequivocal predications about whether or not the economy will be in recession in six months. To translate from a probability of recession to a yes/no recession signal, one compares the probability with a critical value. One innovation in recession signaling that I pursue here is to have regime-specific critical values when the recession probability comes from a regime-switching model. This method of deriving a recession signal reduces the number of false recession signals outside of recession, without impairing the ability to signal the recessions that occur.

REFERENCES

Birchenhall, Chris R.; Jessen, Hans; Osborn, Denise R. and Simpson, Paul. "Predicting U.S. Business-Cycle Regimes." *Journal of Business and Economic Statistics*, July 1999, *17*(3), pp. 313-23.

Chin, Dan M.; Geweke, John F. and Miller, Preston J. "Predicting

Turning Points." Federal Reserve Bank of Minneapolis *Staff Report No. 267*, June 2000.

Del Negro, Marco. "Turn, Turn, Turn: Predicting Turning Points in Economic Activity." Federal Reserve Bank of Atlanta *Economic Review*, Second Quarter 2001, *86*(2), pp. 1-12.

Dueker, Michael. "Strengthening the Case for the Yield Curve as a Predictor of U.S. Recessions." Federal Reserve Bank of St. Louis *Review*, March/April 1997, *79*(2), pp. 41-51.

Estrella, Arturo and Mishkin, Frederic S. "Predicting U.S. Recessions: Financial Variables as Leading Indicators." *Review of Economics and Statistics*, February 1998, *80*(1), pp. 45-61.

Filardo, Andrew J. "How Reliable Are Recession Prediction Models?" Federal Reserve Bank of Kansas City *Economic Review*, Second Quarter 1999, *84*(2), pp. 35-55.

Friedman, Benjamin M. and Kuttner, Kenneth N. "Indicator Properties of the Paper-Bill Spread: Lessons from Recent Experience." *Review of Economics and Statistics*, February 1998, *80*(1), pp. 34-44.

Hamilton, James D. "Analysis of Time Series Subject to Changes in Regime." *Journal of Econometrics*, July/August 1990, *45*(1/2), pp. 39-70.

Investment-Specific Technology Growth: Concepts and Recent Estimates

Michael R. Pakko

The rapid pace of productivity growth since the mid-1990s has been attributed to improvements in technology, particularly in the areas of information processing and communications. From e-mail and cell phones to inventory management and robotic manufacturing techniques, new ways of doing business—facilitated by the use of new types of capital equipment—have transformed the workplace.

However, traditional growth theory and growth accounting techniques—which emphasize the role of disembodied, neutral technological progress—are deficient in explaining the phenomenon of productivity growth driven by technology that is embodied in new types of capital equipment. Consequently, models of "investment specific" technological progress have gained prominence as a framework for evaluating the role of capital-embodied growth.

This article outlines a general model of investment-specific technological change, presents some new estimates, and examines the role that this type of technological progress has in explaining and predicting recent and prospective productivity growth trends.

GROWTH THEORY WITH INVESTMENT-SPECIFIC TECHNOLOGY

The idea that technology can be manifested in new, more efficient types of capital equipment has a long history in economics, dating at least to the "embodiment controversy" of Solow and Jorgenson in the 1960s.[1] The rapid advancement of information-processing and communications technologies has renewed interest in the issue, inspiring the development of general equilibrium models that include investment-specific technological progress.

In this section, I describe a simple neoclassical growth framework—based on the model of Greenwood, Hercowitz, and Krussell (1997)—that incorporates this idea. In addition to balanced, neutral technological progress, the model includes a source of technological change that is associated with improvement in the quality of investment goods that becomes embodied in the productive capital stock. The model differs slightly from Greenwood, Hercowitz, and Krussell in two respects: First, the model in this paper treats equipment and nonresidential structures as two components of a single, composite capital good. In addition, the model described below includes a convex production possibilities frontier.

Our interest is in explaining economic growth—a sustained increase in economic activity per capita. Hence, attention will focus on "steady state" growth paths in which all variables increase at constant (though possibly differing) rates.

A Growth Model with Two Types of Technological Change

A simple model that incorporates both types of technological change can be described as follows: The household sector is modeled as a representative agent who directly controls the production technology and owns the capital stock. Households supply labor inelastically to the production sector and make consumption-saving decisions by maximizing a stream of discounted utility over consumption:

$$(1) \qquad \sum_{t=0}^{\infty} \beta^t u(c_t),$$

where $\beta < 1$ is a constant discount factor and c_t is (per capita) consumption. The momentary utility function is assumed to be of the constant relative risk aversion (CRRA) form $u(c_t) = c_t^{1-\sigma}/1-\sigma$.

Technology is typically incorporated directly into the production function: Output is produced using capital, labor, and the current state of technology. The production function is Cobb-Douglas and technology is specified in labor-augmenting form[2]:

$$(2) \qquad Y_t = K_t^{\alpha}(z_t L_t)^{1-\alpha} .$$

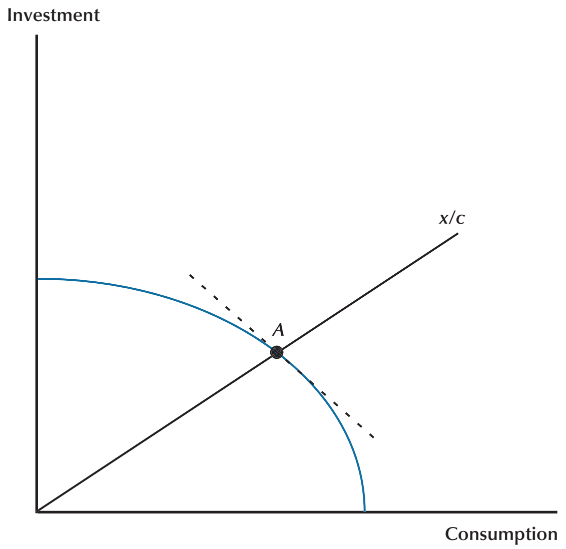
Michael R. Pakko is a senior economist at the Federal Reserve Bank of St Louis. Rachel Mandal, Mrinalini Lhila, and Athena Theodorou provided research assistance.

[1] See, for example, Solow (1960) and Jorgenson (1966). Hercowitz (1998) describes the 1960s controversy in the context of contemporary models of investment-specific technology like the one described here.

[2] Note that with a simple transformation of variables, $\tilde{z} = z^{1-\alpha}$, the production function can be written in the alternative form $Y = \tilde{z} \cdot F(K,L)$, where $\tilde{z}$ is an index of total factor productivity.

Figure 1

Production Possibilities Frontier



Here, z_t is the technology index that directly enhances the productivity of labor, L_t, and indirectly that of capital, K_t, in the production of output, Y_t. Note that equation (2) can be written in terms of labor productivity as

$$y_t = k_t^{\alpha} z_t^{1-\alpha}, \tag{2'}$$

where the ratio of capital and output to labor are represented by lower case variables; that is, $k = K/L$ is the capital/labor ratio and $y = Y/L$ is labor productivity.

Writing equation (2) in terms of log-differences, productivity growth can be expressed as

$$g_y = (1-\alpha)g_z + \alpha g_k, \tag{3}$$

where g_i, for example, denotes the growth rate of variable i. The "growth accounting" equation (3) shows that productivity growth can be decomposed into components representing "total factor productivity," $(1-\alpha)g_z$, and "capital deepening," αg_z.[3]

Investment-specific technology enters the model through the capital accumulation equation,

$$k_t = (1-\delta)k_{t-1} + q_t x_t, \tag{4}$$

which states that the current productive capital stock consists of undepreciated capital from the previous period plus net investment, $q_t x_t$.[4] In equation (4) physical investment measured in consumption units, x_t, is enhanced by an index of the *quality* of newly produced capital goods, q_t. The product $q_t x_t$ represents investment as measured in efficiency units. The improvement in the quality of capital goods reflected in increasing values of q_t is the driving force behind investment-specific technological change.

In the subsequent analysis of the growth properties of these two types of technology, we will assume that the economy's opportunities for producing consumption goods and investment goods is characterized by a nonlinear production possibilities frontier, $H(c_t, x_t)$, that is concave and invariant to scale.[5] Figure 1 illustrates the tradeoff summarized by $H(\cong)$. For a given level of technology and existing capital, the economy is capable of producing any combination of consumption and investment lying on or below the production possibilities frontier (PPF). Points that lie outside the frontier are not feasible given the current state of technology, while points inside the frontier imply inefficient underutilization of resources. The optimal production combination will therefore lie on the frontier itself. The slope of the PPF at any given point shows the trade-off between consumption goods and investment goods—that is, their relative price.

The durability of capital goods means that investment produces a stream of consumption goods into the future. Hence, the location of the optimal point on the PPF will depend on household preferences for substituting consumption between the present and the future (which, given the separable CRRA form of utility assumed, is time invariant in this model).

This combination of consumption and investment can be found from the representative agent's problem of maximizing utility (1) subject to the overall resource contraint,

$$k_t^{\alpha} z_t^{1-\alpha} = H(c_t, x_t), \tag{5}$$

and to the capital accumulation equation (4).

[3] Using the notation from the previous footnote, the growth accounting equation (3) can be written explicitly in terms of the total factor productivity variable, $g_y = g_{\check{z}} + \alpha g_k$.

[4] The accumulation equation is often written so that there is a one-period time to build; that is, capital at time t depends on investment at time t–1. The specification in this paper simplifies the exposition of capital growth and emphasizes the flow concept of investment.

[5] Formally, the $H(\cdot)$ function is assumed to be homothetic. The nonlinear PPF can be thought of as shorthand for a more detailed model in which consumption goods and investment goods are produced in separate sectors, with costly transfer of factors between sectors.

For a given level of technology, the representative household's maximization problem yields an optimal investment/consumption ratio, implying a specific equilibrium such as that shown as point *A* in Figure 1. The slope of the dotted line shows the price of consumption goods relative to investment goods implied by this equilibrium.

Case 1: Neutral Technological Change and Balanced Growth

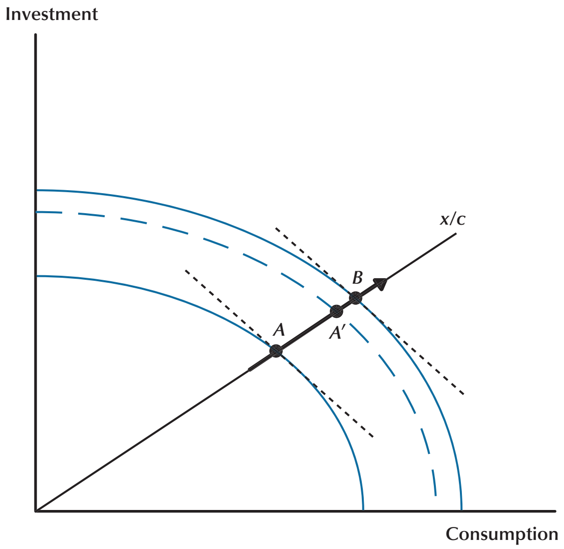
Suppose that the sole source of technology growth is z_t, the index of labor-augmenting technological progress. For given quantities of labor and preexisting capital, an increase in z_t shifts the production possibilities frontier outward, as shown in Figure 2.

Because the expansion takes the form of a radial outward shift, both consumption and investment expand at the same rate as total output; that is, $g_c = g_x = g_y$.[6] Moreover, with q constant, the capital accumulation equation (4) implies that capital grows at the same rate as investment, $g_k = g_x$. Hence, this type of growth is often referred to as "balanced," based on "neutral" technological progress. With investment and consumption growing at the same rate, the economy's growth path will be characterized by a constant x/c ratio, as shown by the growth path running through points *A* and *B* in Figure 2. Along this growth path, the slope of the PPF, representing the relative price of consumption and investment goods, is also constant.

From the growth accounting relationship (3), the shift in the PPF includes the direct effect of the increase in z_t (represented in Figure 2 as movement point *A′*) as well as a component associated with capital growth (accounting for the remaining shift to point *B* in Figure 2). However, the role of capital deepening for this type of technological expansion is distinctly secondary. The direct effect of technology growth is an expansion of investment, which gives rise to a commensurate growth rate of capital. Indeed, substituting the relationship $g_k = g_y$ into the growth accounting equation (3), we find that the rate of output growth (as well as of consumption, investment, and capital growth) is equal to the rate of labor-augmenting technical progress. Although the growth accounting decomposition shows a role for capital deepening, there is no sense in which technological progress is "embodied" in capital growth. Rather, the capital component represents a passive response to "disembodied" technological progress and does not comprise a truly independent source of economic expansion.[7]

Figure 2

Balanced Growth



Case 2: Investment-Specific Technological Change and Capital-Embodied Growth

Growth associated with investment-specific technological progress differs from neutral technology growth in several respects. First, note that q_t does not appear directly in the economy's resource constraint, (5). Instead, the investment-specific technology index appears in the capital-accumulation equation and therefore operates through the capital-deepening component of the growth accounting equation.

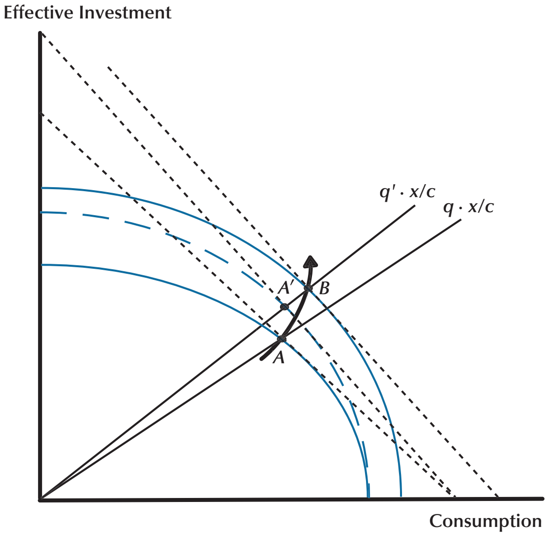
Investment-specific technological progress can be illustrated using a modified PPF framework, as shown in Figure 3. The vertical axis now measures "effective" investment, $q_t x_t$, incorporating the notion of improvement in the quality of investment goods. In Figure 3, the direct effect of an increase in q_t is

[6] This outcome is ensured by the scale-invariance property that is implied by the assumed homotheticity of utility and the PPF function.

[7] Capital deepening does play an important role in the adjustment dynamics of the model. That is, when the economy is not on its steady-state growth path (King and Rebelo, 1993) or is in the transition between steady-state paths (Pakko, 2002b), capital deepening is the mechanism that moves the economy toward long-run equilibrium.

Figure 3

Investment-Specific Growth



shown by the rotation of the PPF to the dashed line passing through point A'. This twist in the PPF represents a change in the tradeoff between consumption and capital accumulation. The movement from point A to point A' represents no change in c or x (or their ratio), but is simply a measure of the growth of "effective" capital that is made possible by the increase in q.

For this reason, the selection of an appropriate numeraire is important. If output were to be measured in terms of constant prices, the shift in the PPF attributable to the increase in q would imply that output had risen for fixed inputs of labor and capital. Hence, growth accounting would incorrectly attribute part of the change to an increase in neutral technology, z. This mismeasurement would be even more severe if total output were measured in units of investment goods.

When the consumption good is taken as numeraire, total real income—as measured in consumption units along the horizontal axis—is left unchanged by the direct effect of growth in q_t. Appropriate measurement of investment-specific versus neutral technology growth therefore requires that the data be expressed in consumption units. In practice, this means that for growth accounting in the presence of investment-specific technical progress, nominal output and investment data should be deflated by a consumption price index.[8]

From the accumulation equation, an increase in q_t has the effect of increasing the effective capital stock. In fact, when improvement in the quality of capital goods is accounted for, the growth rate of the capital stock will be the sum of the growth rates of physical investment and quality improvement,

$$g_k = g_x + g_q . \tag{6}$$

Hence, the indirect impact of investment-specific technology growth will be reflected in effective capital stock growth that shifts the PPF in Figure 3 outward. As was the case for neutral technological progress, the growth component of investment-specific technological progress will be represented by a radial outward shift of the PPF that is characterized by a constant x_t/c_t ratio and a common growth trend for output, consumption, and physical investment.

Substituting equation (6) and the relationship $g_y = g_x$ into (3), we obtain a relationship between productivity growth and the two sources of technology growth:

$$g_y = g_z + \frac{\alpha}{1-\alpha} g_q . \tag{7}$$

In the presence of investment-specific technological progress, total economic growth will be equal to the rate of labor-augmenting technical change plus a component reflecting improvement in the quality of capital goods. Hence, investment-specific growth represents a channel through which technological progress is manifested through "embodiment" in productive capital.

Two features of the growth path passing through points A and B in Figure 3 are important for evaluating the role of investment-specific technology in the data. First, investment—when properly measured to include improvements in the quality of new capital goods—is predicted to grow faster than consumption along a steady-state growth path. In addition, the nature of the change in the tradeoff between consumption and investment, represented by the twist in the PPF, implies that the relative price of investment goods should be falling over time relative to consumption goods.

Figure 4 shows that these trends are, in fact, a characteristic of the data in the National Income and Product Accounts (NIPA).[9] The ratio of investment

[8] Greenwood, Hercowitz, and Krusell (1997) emphasize this point.

[9] In Figure 4, "investment" corresponds to total nonresidential fixed investment and "consumption" is measured as nondurables plus services less housing services.

to consumption has risen persistently over the past half-century and has appeared to accelerate sharply in the past decade. Simultaneously, the price of investment relative to consumption has followed a clear downward trend since at least the late 1950s, with the rate of decline increasing since the 1980s.

ESTIMATES OF INVESTMENT-SPECIFIC TECHOLOGICAL CHANGE

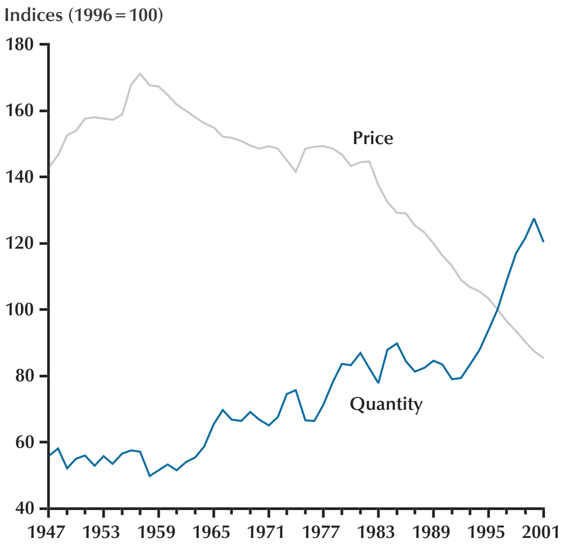
The data presented in Figure 4 suggest that investment-specific technology growth has been an important feature of post-WWII trends in productivity growth. In order to quantitatively evaluate the role of investment-specific technology, however, it is important to carefully examine the issue of quality improvement for investment goods.

The measurement of quality change has always been important in the construction of the NIPA data. Quality characteristics of newly introduced goods are routinely incorporated into the data using so-called "matching models" that compare the attributes of new and existing products. In recent years, the BEA has implemented several revisions to its methodologies in order to account for the rapid rate of innovation in information processing, communications, and other high-tech sectors. In particular, so-called "hedonic regression techniques" have been applied to construct quantity and price indices that adjust for changes in quality over time. Among the more important applications of this approach, the BEA incorporates hedonic indices for computer equipment and purchased software, telephone switching equipment, cellular services, and video players, among others.[10] Moreover, the BEA has even changed its aggregation methodology to more accurately measure the contribution of quality change to GDP growth: the adoption in 1996 of a chain-weighting methodology was intended to allow aggregates to track quality improvement better over time.

Nevertheless, some economists contend that a significant amount of quality change goes unmeasured in the official statistics, particularly in cases where quality improvement is more incremental. In a seminal 1990 study, *The Measurement of Durable Goods Prices*, Robert Gordon undertook to quantify the extent of this unmeasured quality change. Drawing data from a variety of sources, including special industry studies, *Consumer Reports*, and the Sears catalog, Gordon compiled a data set of more than 25,000 price observations. He constructed quality-adjusted price indexes for 105 different product categories, then aggregated the data to correspond to the individual components of the BEA's measure of spending on producers' durable equipment. For each of 22 major categories of investment, Gordon calculated "drift ratios" representing the cumulative deviation of his adjusted price measures from the official data. The adjusted price components were then used to deflate nominal series, with the resulting real series aggregated to create a new quality-adjusted series for investment in durable equipment.

Figure 4

NIPA Investment and Consumption: Relative Prices and Quantities



The bottom line of Gordon's study was that the official NIPA data (as constructed at the time) understated the true growth rate of real investment spending by nearly 3 percentage points per year over the period 1947-83. This quality adjustment for real investment spending is mirrored in the price deflator: the finding that quality-adjusted real investment spending is undermeasured implies that increases in the price of investment goods have been overstated. Unfortunately, Gordon's data set extends only through 1983.

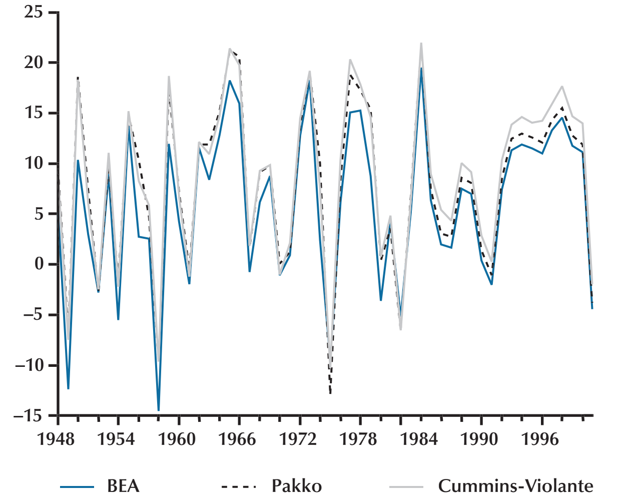
Quality Improvement for Equipment and Software Investment

Previous estimates of investment-specific technology growth have been based on extrapolation of

[10] Landefeld and Grimm (2000) report that 18 percent of GDP is estimated using hedonic methods.

Figure 5

Growth Rates for Equipment and Software Investment



Gordon's aggregate data for series producers' durable equipment. For example, Greenwood, Hercowitz, and Krusell extended the Gordon data through 1990 by adding 1.5 percent to the growth rates of real investment spending for all categories except computers. Hornstein (1999) invoked a similar procedure to extend the estimate through 1997.

As BEA definitions and methodologies are updated and as relative shares of the components of equipment investment change over time, however, the simple extrapolation of Gordon's aggregate data becomes less satisfactory. Ideally, one would like to have extended data series at the disaggregated level of Gordon's original study. A less ambitious alternative is to extrapolate the drift ratios for each of Gordon's 22 major investment categories independently—accounting for changes in BEA definitions and methodology—then aggregate the extrapolated data to calculate a new, extended series.[11]

Two recent studies have followed variants of this procedure. Cummins and Violante (2002) estimate a simple time-series model that relates Gordon's quality-adjusted estimates and official BEA time series data for each of the individual investment categories.[12] After estimating the coefficients of the model, Cummins and Violante extrapolate out-of-sample estimates of quality-adjusted price levels for the period 1984-2000.[13] Pakko (2002a,b) uses a simpler extrapolation technique: recognizing that the measurement bias documented by Gordon is larger in the earlier years of the sample period than the latter period, the Pakko estimates are based on a linear extrapolation of Gordon's drift ratios for the period 1973-83. The drift ratios were then applied to the official BEA price data to create extended quality-adjusted series.

Both sets of estimates were then aggregated to create a quality-adjusted measure of equipment investment for the period 1947-2000. Recent changes in BEA definitions and methodology complicate this procedure. One important innovation made in 1996 was the inclusion of software as an investment component. Gordon's data set did not include software, so both Pakko and Cummins and Violante used the official BEA measure for this component. Similarly, the BEA has devoted considerable effort to accurately measuring quality change for computers and peripheral equipment; hence, both studies assume that the bias found by Gordon in the vintage data has been eliminated in contemporary time series estimates for that component.

Figure 5 shows annual growth rates of these quality-adjusted series for aggregate equipment and software investment, along with the corresponding BEA measure. The two adjusted series track each other closely during the 1947-83 period, since both are based on Gordon's original data.[14] The main source of divergence between the estimates over this period is the difference in aggregation methodologies: Cummins and Violante use the Törnqvist index approach advocated by Gordon, while Pakko uses the Fisher-ideal chain-weighting approach that has subsequently been adopted by the BEA.[15]

During the post-1983 period, the Cummins-Violante series displays more rapid growth than the

[11] A disaggregated approach is preferable to a simple extrapolation of the aggregate trend for two reasons: First, several changes in the BEA's definitions and methodology have, for some components, eliminated or at least mitigated the measurement problems suggested by Gordon's study. In addition, the procedure of re-aggregating the quality-adjusted components using a chain-weighting methodology allows the role of changing expenditure shares over time to be appropriately accounted for.

[12] The model posits that the adjusted price index is a function of a constant, a time trend, current and lagged values of the BEA time series, a cyclical indicator (lagged GDP growth), and an error term.

[13] Giovanni Violante was kind enough to provide the data from Cummins and Violante (2002).

[14] The growth rates of both adjusted measures exceed the official BEA growth rate by an average of about 2.75 percent per year over this period.

[15] The measures also differ in that the Pakko aggregate includes net sales of scrap equipment (excluding autos), as measured by the BEA.

Table 1

Growth Rates and Contributions to Growth of Nonresidential Fixed Investment

Source of equipment and software data	Nonresidential fixed investment			Equipment and software			Nonresidential structures		
	1950-2001	1950-1975	1976-2001	1950-2001	1950-1975	1976-2001	1950-2001	1950-1975	1976-2001
BEA	5.17 (100)	4.59 (100)	5.75 (100)	6.26 (80.3)	5.30 (70.9)	7.23 (87.8)	2.96 (19.7)	3.48 (29.1)	2.44 (12.2)
Adjusted 1	6.85 (100)	6.65 (100)	7.05 (100)	8.35 (80.0)	8.02 (74.0)	8.68 (85.7)	3.95 (20.0)	4.48 (26.0)	3.42 (14.3)
Adjusted 2	7.28 (100)	6.61 (100)	7.95 (100)	8.97 (81.2)	7.95 (73.8)	9.98 (87.3)	3.95 (18.8)	4.48 (26.2)	3.42 (12.7)

NOTE: Numbers in parentheses refer to percent contributions to NFI growth.

Pakko series—due largely to assumptions regarding quality change in communications equipment. In 1997, the BEA introduced a quality-adjusted price index for telephone switching equipment and carried back these revisions to 1985 in the 1999 comprehensive revision of the national accounts.[16] Because this component (the largest single component in the communications equipment category) was the predominant source of quality bias in Gordon's study, Pakko considers that the updated BEA data accurately measure quality change in that sector. On the other hand, Cummins and Violante note that the quality of other types of telecommunications equipment has been improving rapidly, so they opt to use their extrapolated estimate of quality bias from the Gordon data set (amounting to a drift ratio of nearly 7 percent). The two studies also differ somewhat in their treatment of automobiles, instruments and photocopy equipment, and office equipment other than computers.[17] The effect of these differences in assumptions and methodology is that, for the 1984-2000 period, the Cummins-Violante series displays an average annual growth rate that is 2.7 percent higher than the official BEA data, while the growth rate of the Pakko series exceeds the BEA measure by only 1.1 percent per year.[18]

Incorporating Quality Change for Nonresidential Structures

In addition to equipment and software, another important component of the capital stock is the structures component—accounting for approximately 35 percent of nominal nonresidential fixed investment in the period 1948-2001. Gort, Greenwood, and Rupert (1999) examined the measurement of quality improvement in nonresidential structures and estimated that the official NIPA data understates real, quality-adjusted growth by approximately 1 percent per year.

To account for this source of investment-specific technology growth, I construct an adjusted measure of nonresidential structures by adding 1 percentage point to each year's growth rate in real nonresidential structures over the sample period of 1947-2001 (deducting 1 percent growth annually from its price index). The resulting real investment series and price index are then aggregated by chain-weighting with the adjusted measures of equipment and software spending to produce quality-adjusted decompositions for total private nonresidential fixed investment (NFI).

Table 1 shows the growth rates for these estimates of quality-adjusted NFI, along with the contribution of equipment and software and nonresidential structures to total growth.[19] Two measures of quality-adjusted data are included: The first corresponds to the equipment and software data from Pakko. The second measure uses the Cummins-Violante data series. Both measures incorporate the quality improvement in structures suggested by Gort, Greenwood, and Rupert.

For the period 1950-2001, equipment and software spending accounted for more than 80 percent of the growth in total nonresidential investment. The relative contributions to growth have not been constant over time, however. During the first half

[16] Moulten and Seskin (1999).

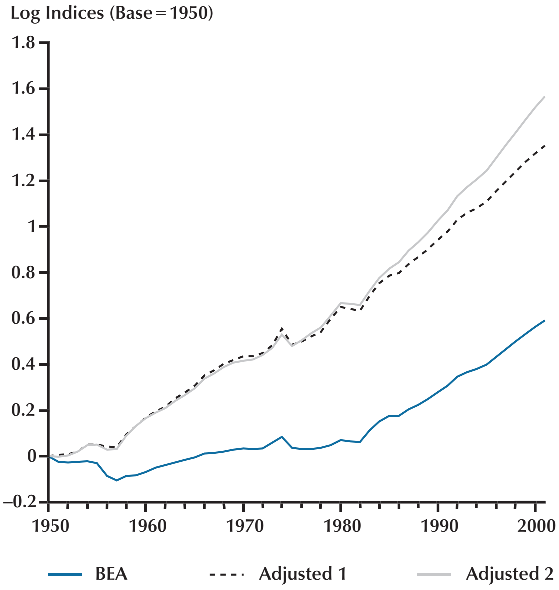
[17] For more detail, see the appendix to Pakko (2002a).

[18] These growth spreads are used to extrapolate each of the adjusted series for the final growth observation from 2000 to 2001 (which was not included in either of the original series).

[19] The contributions to growth in Table 1 are calculated using the BEA's current methodology, as described in Moulten and Seskin (1999).

Figure 6

Investment-Specific Technology (*q*)



of the sample period, equipment and software investment accounted for less than 75 percent of total NFI growth, but accounted for 85 to 90 percent during the second half of the sample.

In previous literature, estimates of investment-specific technology growth have treated the equipment and structures components separately. The estimates in this article use chain-weighted aggregates of both components, allowing flexibility to account for the shifting growth-shares, suggested by the pattern of growth contributions shown in Table 1.

Growth Accounting with Investment-Specific Technical Progress

The data for quality-adjusted investment and associated price indices form the basis for estimating the contribution of investment-specific technology to productivity growth. The first step is to calculate the index of investment-specific technology, *q*, as the price of consumption goods relative to (quality-adjusted) investment goods:

$$q_t = P_c / \widetilde{P}_i , \tag{8}$$

where $\widetilde{P}_i$ is a quality-adjusted price index for investment and P_c is a consumption price index. Following the practice common in previous literature, the consumption price index used for this calculation covers nondurables and non-housing services.[20] Durable goods are excluded from the consumption measure so as to avoid issues of quality improvement in that component.

Figure 6 shows this measure of *q* for each of the three measures of investment prices constructed in the previous section. The data are indexed to a base year of 1950 in order to show their cumulative growth. The two quality-adjusted measures track each other closely through 1983, exceeding the growth rate of the unadjusted NIPA relative price by an average of 1.9 percent. For the period 1984-2001, the two adjusted series exceed the NIPA-based series by 1.0 percent (estimate 1, Pakko) and 2.0 percent (estimate 2, Cummins and Violante) per year.

The estimates of *q*, along with associated data for real investment, *x*, can be used to construct adjusted measures of the capital stock that account for embodied technological progress. The model suggests that real physical investment corresponds to nominal investment deflated by the consumption price index, $P_i I / P_c$. Effective investment, *qx*, is therefore given by $P_i I / \widetilde{P}_i$. In the NIPA data, with $P_i = \widetilde{P}_i$, this is simply the real investment series, so the BEA's data for private nonresidential fixed assets is an appropriate measure of the capital stock. For each of the adjusted investment series, *qx* is the quality-adjusted real component from which a quality-adjusted measure of the capital stock can be derived.

The procedure used to construct quality-adjusted capital stock measures is as follows: First, I use the accumulation equation (4) and the NIPA series for investment and capital to back out a series of implied depreciation factors, $(1–\delta_t)$.[21] These factors are then used to construct synthetic capital-stock series using a perpetual-inventory method—that is, by reconstructing the capital stock using equation (4) with the quality-adjusted investment data. Starting values for capital stocks in the base year used for these calculations, 1950, are initialized using the accumu-

[20] The non-housing services data are constructed by chain-weighting PCE services with the additive inverse of the housing services component. The resulting series is then chain-weighted with nondurables consumption.

[21] The BEA constructs measures of net stocks for individual components, then uses chain-weighted aggregation to build aggregates. The use of these annual depreciation factors approximately adjusts for changes in the composition of the capital stock and total depreciation that arise from this procedure. For more information about the construction of the BEA's fixed-assets series, see Katz and Herman (1997).

Figure 7

Capital Stock Growth Rates

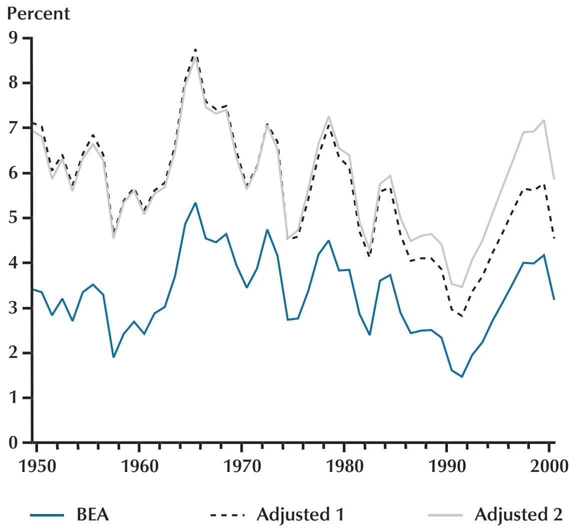
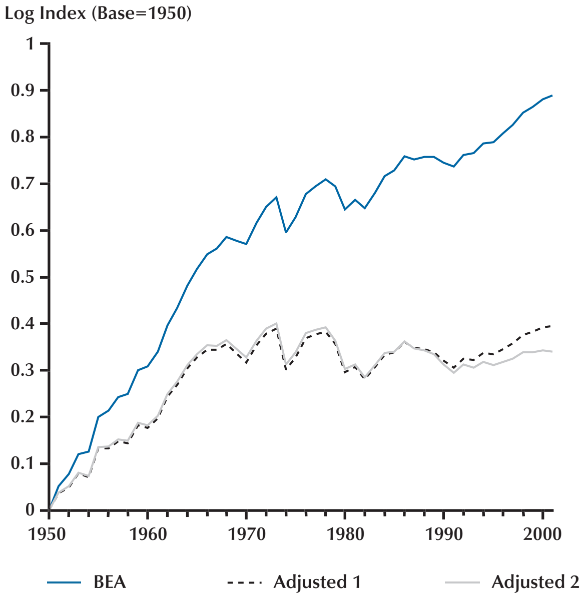


Figure 8

Neutral Technology (*z*)



lation equation (4) to relate investment/capital ratios to the BEA data.[22] The growth rates for these adjusted capital stock series, shown in Figure 7, exceed the official BEA measure by about 2.2 percent (Adjusted 1) and 2.5 percent (Adjusted 2) per year on average over the entire sample period.

Completion of the growth accounting exercise requires data for output and labor. In order for the data to correspond to a broad measure of labor productivity, output is taken to be gross domestic business product.[23] Business sector hours—from the BLS labor productivity accounts—is used to measure labor input. These data series, along with the series for capital growth, can be used to back out measures of labor-augmenting technological change from the growth accounting equation (3).

Figure 8 shows measures of "neutral" technology that are derived from this procedure, where the series are expressed in log levels relative to a 1950 base in order to illustrate cumulative growth. Each of the series displays a clear decelleration beginning in the early 1970s, corresponding to the widely cited "productivity slowdown" that has prevailed for much of the subsequent period. For the two measures of *z* derived from quality-adjusted data, the slowdown is particularly distinctive. After growth associated with investment-specific technology has been accounted for, the indexes of neutral technological progress have been nearly flat since 1970.

Estimates of Neutral and Investment-Specific Technology

Table 2 provides a summary of the two sources of growth, reporting the contributions to total potential growth provided by neutral and investment-specific technological change as in equation (7). For comparability with previous studies, the data coverage for this decomposition begins with 1954. Over the entire period from 1954 to 2001, the quality-adjusted measures show that the role of investment-specific technological change has been considerable, accounting for 60 to 68 percent of total potential growth. Even the unadjusted NIPA data show a contribution of investment-specific technology that is 25 percent of the implied total.

The relative contributions of the two types of technology growth have not been constant over

[22] In particular, the accumulation equation implies that $qx/k = (1-\delta)/(1+g_k)$. The properties of the adjusted investment series and growth rates of investment-specific technology can therefore be used to relate the initial adjusted capital stock levels to the BEA data. This calculation yields initial values for the adjusted capital series of about one-third the level of the official data. Results are not very sensitive to small changes in the assumptions generating this relationship, however.

[23] In keeping with the model's implications for price measurement, real output is calculated by deflating nominal business sector GDP using the consumption price deflator.

Table 2

Sources of Technological Progress, 1954-2001

	1954-2001	1954-1977	1978-2001
BEA			
Neutral	1.65	2.46	0.83
Investment-specific	0.55	0.10	1.01
Total	2.21	2.56	1.85
Adjusted 1			
Neutral	0.69	1.29	0.09
Investment-specific	1.20	0.91	1.50
Total	1.89	2.20	1.59
Adjusted 2			
Neutral	0.57	1.33	–0.18
Investment-specific	1.39	0.93	1.86
Total	1.97	2.26	1.67
Actual productivity growth	1.76	2.52	1.00

the entire sample period, however. In the first half of the sample, 1954-1977, the quality-adjusted measures show that investment-specific technology contributes only about 38 percent to total potential. For the NIPA numbers, the measured contribution of investment-specific technology during this period is negligible. During the second half of the sample period, investment-specific technology overwhelmingly predominates. The second of these two adjusted measures (derived from Cummins and Violante) shows a *negative* contribution for neutral technology growth, while the first measure (based on Pakko) measures the contribution of neutral technology at only about 6 percent.

Implications for Potential Productivity Growth

The final line in Table 2 shows the actual growth rate of output per worker for the relevant sample periods. Over the entire period from 1954 to 2001, all three measures of technology change overpredict actual growth. As demonstrated in the breakdown between the first half and the second half of the sample, however, the overprediction is attributable to the more recent span of years. For the period 1954-77, the two adjusted measures slightly *underpredict* actual productivity growth. For the period 1978-2001, all three overpredict actual growth.

Of course, the measures of potential growth derived from the estimated technology series are approximated measures of long-run relationships, so it is not surprising that they do not precisely replicate actual growth over any given finite sample. During the period from the mid-1970s to the present, however, the magnitude and prevalence of the discrepancy suggest more than measurement or approximation error.

Recent research on the economic effects of introducing new technologies help to explain the apparent gap between measures of technology growth and productivity growth. The data suggest a rather dramatic change in the pattern of technology trends: the period of slow productivity growth in the 1970s and 1980s is associated with a change in the composition of technological progress from neutral to investment-specific technology.

Many economists have suggested that changes in trend technology growth— particularly for capital-embodied technologies—are associated with long transition periods during which productivity lags the rate of technological advance. Indeed, both Cummins and Violante (2002) and Pakko (2002b) focus on the adjustment of productivity growth to technological innovations. Cummins and Violante calculate that the “technology gap”—the gap between the productivity of the best technology and average productivity—rose from 15 percent in 1975 to 40 percent in 2000. This finding is in the spirit of “technology diffusion” models (e.g., Hornstein and Krusell, 1996; Jovanovic and MacDonald, 1994; Greenwood and Yorukoglu, 1997; Andolfatto and MacDonald, 1998; Hornstein, 1999), which posit that learning about the full potential of new technologies can generate long implementation lags as resources are channeled into the process of adapting new technologies into existing production structures.[24] Pakko (2002b) shows that even in the absence of explicit diffusion lags, the adjustment of the capital stock to changes in technology growth trends give rise to long lags between technology and productivity—particularly when technology growth is investment-specific. These findings can be interpreted as suggesting that a great deal of the potential productivity improvement has yet to be fully incorporated into measured actual productivity growth.

CONCLUSIONS

A great deal of attention has recently been paid to the notion that rapid technological innovation

[24] Another class of general growth models addressing the adaptation of “general purpose technologies” (e.g., Helpman, 1998) suggests similar lags.

has been the driving force behind recent gains in U.S. productivity growth. However, the nature of these technology advances—being embodied in entirely new types of high-tech capital equipment—is not well explained by classical growth theory. This paper has reviewed a class of economic models featuring "investment specific" growth that explicitly describe a process in which new technologies are capital-embodied.

Recent estimates of the magnitude of this type of technology growth reported in this article suggest that over 60 percent of potential productivity growth over the past half-century can be attributed to investment-specific technology. Since at least the mid-1970s, the estimates suggest that the importance of investment-specific technology has increased sharply, accounting for practically *all* of the implied potential productivity gains.

However, measured productivity growth has fallen short of these estimates of potential. Recent research on the process of adapting new technologies to existing production frameworks gives reason for optimism about this finding. To the extent that rapid growth of investment-specific technological innovation has yet to be fully exploited, as the data suggest, technology-related gains in productivity should be expected to continue well into the future.

REFERENCES

Andolfatto, David and MacDonald, Glenn M. "Technology Diffusion and Aggregate Dynamics." *Review of Economic Dynamics*, April 1998, *1*(2), pp. 338-70.

Cummins, Jason G. and Violante, Giovanni L. "Investment-Specific Technical Change in the United States (1947-2000): Measurement and Applications." *Review of Economic Dynamics*, April 2002, *5*(2), pp. 243-84.

Gordon, Robert J. *The Measurement of Durable Goods Prices*. Chicago: University of Chicago Press, 1990.

Gort, Michael; Greenwood, Jeremy and Rupert, Peter. "Measuring the Rate of Technological Progress in Structures." *Review of Economic Dynamics*, January 1999, *2*(1), pp. 207-30.

Greenwood, Jeremy; Hercowitz, Zvi and Krusell, Per. "Long-Run Implications of Investment-Specific Technological Change." *American Economic Review*, June 1997, *87*(3), pp. 342-62.

___________ and Yorukoglu, Mehmet. "1974." *Carnegie-Rochester Conference Series on Public Policy*, June 1997, *46*(0), pp. 49-95.

Helpman, Elhanan, ed. *General Purpose Technologies and Economic Growth*. Cambridge, MA: MIT Press, 1998.

Hercowitz, Zvi, "The 'Embodiment' Controversy: A Review Essay." *Journal of Monetary Economics*, February 1998, *41*(1), pp. 217-24.

Hornstein, Andreas. "Growth Accounting with Technological Revolutions." Federal Reserve Bank of Richmond *Economic Quarterly*, Summer 1999, *85*(3), pp. 1-22.

___________ and Krusell, Per. "Can Technology Improvements Cause Productivity Slowdowns?" *NBER Macroeconomics Annual 1996*. Cambridge, MA: MIT Press, 1996, pp. 209-59.

Jorgenson, Dale W. "The Embodiment Hypothesis." *Journal of Political Economy*, February 1966, *74*(1), pp. 1-17.

Jovanovic, Boyan and MacDonald, Glenn M. "Competitive Diffusion." *Journal of Political Economy*, February 1994, *102*(1), pp. 24-52.

Katz, Arnold J. and Herman, Shelby W. "Improved Estimates of Fixed Reproducible Tangible Wealth, 1929-95." *Survey of Current Business*, July 1997, *77*(5), pp. 69-92.

King, Robert G. and Rebelo, Sergio T. "Transitional Dynamics and Economic Growth in the Neoclassical Model." *American Economic Review*, September 1993, *83*(4), pp. 908-31.

Landefeld, J. Steven and Grimm, Bruce T. "A Note on the Impact of Hedonics and Computers on Real GDP." *Survey of Current Business*, December 2000, *80*(12), pp. 17-22.

Moulton, Brent R. and Seskin, Eugene P. "A Preview of the 1999 Comprehensive Revision of the National Income and Product Accounts: Statistical Changes." *Survey of Current Business*, October 1999, *79*(10), pp. 6-17.

Pakko, Michael R. "The High-Tech Investment Boom and Economic Growth in the 1990s: Accounting for Quality." Federal Reserve Bank of St. Louis *Review*, March/April 2002a, *84*(2), pp. 3-18.

___________, "What Happens When the Technology Growth Trend Changes?: Transition Dynamics, Capital Growth and the 'New Economy'." *Review of Economic Dynamics*, April 2002b, *5*(2), pp. 376-407.

Solow, Robert M. "Investment and Technical Progress," in Kenneth J. Arrow, Samuel Karlin, and Patrick Suppes, eds., *Mathematical Methods in the Social Sciences*. Palo Alto, CA: Stanford University Press, 1960.