Competitiveness and Price Setting in Dealer Markets

LUCY F. ACKERT AND BRYAN K. CHURCH

Ackert is a senior economist in the financial section of the Atlanta Fed's research department. Church is an associate professor in the DuPree College of Management at the Georgia Institute of Technology.
They thank Stephanie Mannis for research assistance and Jerry Dwyer, Saikat Nandi, Larry Wall, and Madeline Zavodny for helpful comments.

THE NATIONAL ASSOCIATION OF SECURITIES DEALERS AUTOMATED QUOTATIONS (NASDAQ) SYSTEM IS AN ELECTRONIC MARKET FOR OVER-THE-COUNTER (OTC) STOCKS. IT IS THE SECOND-LARGEST SECURITIES MARKET IN THE UNITED STATES. ALLEGATIONS THAT DEALERS COLLUDE TO WIDEN BID-ASK SPREADS HAVE LED TO SWEEPING CHANGES IN THE RULES GOVERNING TRADING IN THE NASDAQ STOCK MARKET.

Bid and ask quotes are prices at which dealers or market makers are willing to transact. A market maker is an individual or firm that risks its own capital to provide investors with immediacy of supply and demand. The bid-ask spread represents the cost to investors of transacting with the market maker. Investors prefer a narrow spread because it reduces trading costs and improves liquidity (Amihud and Mendelson 1986). Bid-ask spreads, like other transaction costs, significantly affect the efficiency of capital markets (Bhushan 1994; Kim and Verrechia 1994). In an efficient market, prices quickly reflect new information so that the information cannot be used to derive abnormal trading profit. Stock markets are thought to be more efficient when spreads are narrow because information is disseminated more quickly. Yet, at the same time, dealers must receive adequate compensation for making a market in a security, or the market's liquidity is threatened.

Regulators and investors have asserted that Nasdaq dealers conspire to widen bid-ask spreads in order to in-crease their profit at investors' expense. Academics have amassed a substantial body of evidence relating to the Nasdaq scandal. Yet there is no consensus concerning whether dealers collude to fix prices and widen bid-ask spreads.[1] Observed spreads may result from institutional features particular to the Nasdaq market rather than collusion among market makers.

This article explores the Nasdaq pricing controversy in light of economic theory and evidence of alleged collusion, including evidence contained in U.S. Department of Justice and Securities and Exchange Commission reports (1996). The following section examines the important role that securities markets play in promoting a stable economy. Then the discussion reviews specifics of the two organizational structures commonly adopted—auction and dealer markets. These initial sections provide a foundation for understanding the significance of the Nasdaq controversy. Subsequently the article considers the sources and economic consequences of divergence in spreads. Finally, it elaborates on what constitutes collusive behavior and summarizes the case against Nasdaq.

Functions of Securities Markets

Regulation of securities markets in the United States changed dramatically following documented abuses and market irregularities in the 1920s and early 1930s. At that time there was little confidence in the stability of the U.S. market. The stock market crash of 1929 and poor economic conditions brought the role of stock market to the forefront of the policy debate. New federal laws resulted, and a powerful federal agency, the Securities and Exchange Commission (SEC), was established to enforce those laws. According to the Securities Exchange Act of 1934, a national securities exchange must provide a "free and open market" that "protect(s) investors and the public interest" (15 U.S.C. 78f[b][5] and 78o-3[b][6]).

It is clear that well-functioning financial markets are vital in a thriving economy. A primary function of these markets is the allocation of capital and financial resources. Financial markets move capital from savers to those with productive uses for the capital (that is, those with good investment opportunities). In so doing, a well-functioning securities market maintains continuous and active trading, which allows investors to enter and exit when necessary. Market participants want to receive the best price with speedy execution at low cost. Investors have greater confidence in a market that is fair, open, and orderly and offers low transaction costs. At the same time, an effective securities market facilitates price discovery so that prices quickly reflect information and reveal this news to the market's observers and participants. Financial markets also permit individuals to transfer consumption across time. Well-being improves when individuals are permitted to smooth consumption over their life cycle. Designing a securities market that meets these objectives involves trade-offs because these goals typically conflict (Ganley and others 1998).

Dealer versus Auction Markets

The rules governing securities trading vary across organizational structures. Economists debate the merits of two commonly adopted organizational forms: auction and dealer market structures. In an auction market an investor buys or sells at a price set by another investor's limit order. A limit buy order specifies the maximum price that an investor is willing to pay, whereas a limit sell order specifies the minimum price that an investor is willing to receive. By comparison, in dealer markets investors trade with market makers who simultaneously quote prices at which they are willing to buy (the "bid" price) and sell (the "ask" price) a particular security. The best bid (highest) and ask (lowest) prices determine the inside spread.

The New York Stock Exchange (NYSE) is an auction market that maintains a specialist system, wherein one dealer maintains a market in a particular stock. The specialist enters offers to buy at bid prices and sell at ask prices in order to provide liquidity and continuous trading. Investors place market orders to sell or buy at prevailing market prices, and the specialist fills the orders at the inside (best) bid and ask prices. The specialist also maintains a record or limit order book of investors' unexecuted limit orders. Although a specialist in a sense has a monopoly franchise in a particular stock, the presence of one market maker does not necessarily lead to excessive bid-ask spreads because the limit order book provides competition for order flow. Execution costs are expected to be low because the inside spread is often determined by customers' limit orders and investors can trade directly with each other. Investors get the best available prices, whether the prices are from the specialist or from other investors' limit orders.

Regulators and investors have asserted that Nasdaq dealers conspire to widen bid-ask spreads in order to increase their profit at investors' expense.

In contrast to the NYSE, the Nasdaq is a multiple-dealer market, where several dealers maintain a market in a particular stock. Other important dealer markets include most bond and foreign currency markets, as well as the Chicago Board Options Exchange and the London Stock Exchange. Traders in the Nasdaq market do not gather in one location as in an organized exchange but rather are connected electronically through a computer system. To make a market, dealers simultaneously quote prices at which they are willing to buy and sell a particular stock. Because each Nasdaq stock has at least two market makers, a dealer's spread is not necessarily the inside spread. However, investors' market orders get the "best execution" in that orders to sell or buy at the current market price are filled at the inside bid or ask price, whether or not the dealer receiving the order issued that particular price quote. Prior to the recent Nasdaq rule changes (discussed below), limit orders were not revealed to all market participants and were filled by a dealer when the dealer's quote reached the limit price. The presence of multiple market makers is designed to produce narrow bid-ask spreads through competition for order flow among individual dealers. In addition, dealer markets are more flexible and can handle different types of orders from different types of customers.

Why Might Spreads Differ?

A large body of literature examines the determinants of bid-ask spreads and the effects of institutional structure on pricing in securities markets (Benston and Hagerman 1974; Stoll and Whaley 1990;

1. There is not even agreement on whether Nasdaq spreads are wider than those of stocks listed on other U.S. exchanges. See, for example, Kleidon and Willig (1995) and Woodward (1997).

Neal 1992). The width of the spread reflects the costs of inventory, order processing, and trading with informed agents (Glosten and Milgrom 1985; Stoll 1985; Amihud and Mendelson 1986). By standing ready to buy or sell, the market maker provides a useful service. However, while providing immediacy to investors the market maker is exposed to the risk of a market movement that results in a decrease in the value of inventory held. Additional risk arises because some investors' trades are motivated by private information, and the dealer may trade (unknowingly) with an investor who has superior knowledge about a stock. A market maker will quote a wider spread if the chance of trading with an informed investor is greater. Thus, market makers can pass the cost of trading with informed traders on to uninformed traders through the bid-ask spread.

The average bid-ask spread appears to be smaller in specialist markets like the NYSE than in dealer markets such as Nasdaq. Huang and Stoll (1996) found that in 1991 the average quoted spread for a sample of 175 Nasdaq stocks was $0.50 whereas a carefully matched sample of 175 NYSE stocks had an average spread of only $0.26. The difference in spreads does not appear to be generated by differences in inventory, order processing, or asymmetric information costs across the two markets. However, these stock markets have different institutional features that can affect the bid-ask spread, in addition to having divergent pricing systems. Disentangling the effects of various factors is difficult, if not impossible, so the competitiveness of the Nasdaq market continues to be debated.

An institutional factor that must be considered when comparing spreads across markets is the handling of commissions. On the NYSE all traders are charged explicit commissions whereas on Nasdaq commissions are frequently included in the stock's price, lowering the bid price and raising the ask price. For this reason alone, one should expect to find wider spreads on Nasdaq. However, commissions cannot fully explain the difference in spreads across markets because small traders usually pay explicit commissions on Nasdaq as well as on the NYSE (Huang and Stoll 1996).

Another factor that clearly affects how orders are processed, and in turn the bid-ask spread, is the handling of limit orders. As discussed above, on the NYSE limit orders narrow the spread because limit prices can determine the inside spread. Hence, the best prices are available to investors, whether these prices come from the specialist or from other investors through limit orders. On Nasdaq, prices are set by market makers. Prior to the rule changes made effective last year, limit orders were recorded by individual dealers and did not determine the inside spread. Because of this procedural difference, the measured spread on the two exchanges came from different sources. The reservation prices of market makers and investors also differ, leading to further differences in quoted prices and spreads. Dealers derive earnings by recycling stock rather than through long-run speculation. Dealers' earnings come from buying stock and reselling it at higher prices. Investors, on the other hand, generally trade for the long run and buy or sell based on anticipated increases or decreases in a security's value. Despite a recognition that the treatment of limit orders affects the spread, it appears to provide only a partial explanation for wider Nasdaq spreads (Demsetz 1997).

Other institutional arrangements that affect pricing in securities markets are agreements between brokers and dealers to direct order flow, either internally or externally (Godek 1996; Kandel and Marx 1997). When an order is internalized, a dealer trades with a customer at the inside price quote for the dealer's own account, even if the dealer did not issue the best price quote. When an order is preferenced, a dealer forwards the order to another market maker, who fills the order at the best price quote. The dealer who receives a preferenced order is not necessarily the market maker who issued the best price quote. Internalization and preferencing lead to interdependencies across dealers and limit their incentives to narrow spreads because they do not compete over incoming orders through their price quotes.[2] Experimental economics methods have been used to provide insight into the effect of order preferencing on quoted spreads in dealer markets (Ackert and Church 1998; Bloomfield and O'Hara 1998).[3] These studies conclude that preferencing has striking effects on pricing behavior, even if dealers are not permitted to communicate overtly.

Besides recognizing that the ability to direct customer order flow has important effects on quoted spreads, Dutta and Madhavan (1997) argue that dealers compete for orders along dimensions other than price. Nonprice competition for order flow can take the form of research services or agreements with brokers in which dealers pay brokers for order flow.[4] Because these other inducements reduce the per share value of order flow to the dealer, conclusions about the competitiveness of markets are complex and cannot be based simply on price.[5] Empirical evidence suggests that dealers will compete for order flow using methods other than price (Ackert and Church 1998).

Finally, spreads in dealer markets may be wider than in other market structures if market makers conspire to fix prices. Proponents of dealer markets argue that competition among dealers will produce narrow spreads. With a large number of competitive dealers, cooperative agreements may be difficult to design and enforce. However, Dutta and Madhaven (1997) argue that even dealers who behave noncooperatively can set spreads that exceed the competitive level. They show that self-interested dealers can accrue abnormal profit despite acting noncooperatively. Institutional arrangements, like preferencing, result in abnormal profit levels because these arrangements reduce dealers' incentives to compete for order flow using price. From a public policy standpoint this result is important because dealers are not explicitly cooperating to fix prices; that is, excess spreads can arise without explicit collusion.

Collusion in Securities Markets

In the Nasdaq market, more than thirty dealers are involved in the pricing of an actively traded issue, so it is likely that competitive pressures will come to bear. Collusion may be difficult because of the absence of explicit barriers to entry (Grossman and others 1997). In addition, the "product" or service provided is not necessarily homogeneous because market makers may offer cash payments for order flow and other noncash services. However, it is difficult to ignore the words of the dealers themselves (see Box 1). Their testimony, from depositions taken during the Department of Justice investigation, and audiotaped conversations suggest that Nasdaq market makers followed an industrywide practice or quoting convention that fixed transaction prices. The practice of violating the industry's quoting convention, referred to by traders as making a Chinese market, was actually viewed within the industry as unethical and unprofessional conduct.

In understanding recent U.S. securities market experience, it is important to consider what sorts of behavior are deemed anticompetitive. Collusion to raise prices is certainly not a practice or a concern of recent origin. According to Adam Smith, "People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices" ([1776] 1994, 148). The dictionary defines collusion as a "secret agreement or cooperation for an illegal or deceitful purpose." American law on overt price fixing is clear. Such behavior is illegal per se. However, in many cases there is no explicit agreement to fix prices. Under the Sherman Act, the U.S. courts developed the conscious parallelism doctrine. The Supreme Court explained the doctrine as follows: "No formal agreement is necessary to constitute an unlawful conspiracy. Often crimes are a matter of inference deduced from the acts of the person accused and done in pursuance of a criminal purpose. . . . The essential combination or conspiracy in violation of the Sherman Act may be found in a course of dealings or other circumstances as well as in an exchange of words. . . . Where the circumstances are such as to warrant a jury in finding that the conspirators had a unity of purpose or a common design or understanding, or a meeting of minds in an unlawful agreement, the conclusion that a conspiracy is established is justified" (American Tobacco Co. et al. v. the U.S. 328 U.S. 781 [1946]).

The Case against Nasdaq

Serious questions about the competitiveness of the Nasdaq market surfaced in two widely publicized studies by Christie and Schultz (1994) and Christie, Harris, and Schultz (1994). The allegations led to investigations by the Department of Justice and the SEC, as well as numerous on-going civil lawsuits (see Box 2). Christie and Schultz report that odd-eighth price quotes are nearly nonexistent for many Nasdaq stocks and suggest that market makers implicitly collude to widen spreads by avoiding odd-eighth price quotes.[6] According to the Justice Department, this pricing convention existed for at least three decades. However, following the publicity of the first study, Christie, Harris, and Schultz report a sudden decline in the spreads for several actively traded issues and a concomitant increase in the use of odd-eighth price quotes for those stocks. In fact, the inside spreads for Amgen Inc., Cisco Systems, and Microsoft Corporation fell by almost 50 percent immediately after newspapers reported the results of the first Christie and Schultz study. Average spreads for these stocks fell from between \$0.25 and \$0.45 to between \$0.151 and \$0.175.

The United States brought a civil action under the Sherman Act with the claim for relief justified as follows: "Beginning at least as early as 1989, and continuing to the date of this Complaint, a common understanding arose among the defendants and other Nasdaq market makers concerning, among other things, the manner in which

2. Although orders are preferenced and internalized on the NYSE, the arrangement is more prevalent on Nasdaq (Huang and Stoll 1996).

3. Experimental economics methods allow the researcher to conduct investigations that cannot be conducted in naturally occurring markets and complement studies using traditional archival data. In the laboratory the experimental researcher can control factors that are extraneous to the investigation. For example, Ackert and Church (1998) are able to directly examine how dealers' spreads are affected by order preferencing while controlling the overt communication among dealers.

4. Competition for order flow from brokers may result in order flow payments to the brokers that reduce market makers' profits and can be viewed as a way for dealers to share their profits with brokers. The extent to which brokers, in turn, pass these earnings on to individual investors is unclear.

5. Another complication when using price quotes to assess competitiveness arises because many transactions are negotiated and occur between the best bid and ask price (Bessembinder 1997).

6. In June 1997 the NYSE followed the AMEX and Nasdaq and permitted trading in increments of one-sixteenth. Historically, most stocks listed on large U.S. exchanges were quoted in increments of one-eighth, though moving to decimalization is debated. See, for example, Angel (1997).

Making a Chinese Market

As reported in the Department of Justice's *Competitive Impact Statement,* the traders' testimony provides insight into the degree of interdependence in the Nasdaq market and the entrenchment of the pricing convention. According to the market makers, those who attempted to "break the spread" by violating the pricing convention created a "Chinese market." For example, the following trader's testimony suggests that creating a Chinese market was not only considered unprofessional but traders were actually trained to conform to the convention:

> **Q:** And through the period December '93 through December of '94, do you observe the market makers entered very—relatively few odd-eighths. And by that I mean, with perhaps one or two exceptions *under 10 percent of their quotes were odd eighths* in McCormick.
> **A:** Yes, ma'am.
> **Q:** *And again, is that, in your professional opinion, because those market makers had three-quarter point dealer spreads and did not want to enter what were termed "unprofessional markets"?*
> **A:** *Yes, ma'am.*
> **Q:** How is it that all of the market makers knew that entering an odd eighth quote could be unprofessional?
> **A:** *Young traders were trained over the years not to put in unprofessional markets, "Chinese markets." . . . This was part of the—of the traditional and ethical on-the-job training that all of us got, and it encompasses not only that you don't put in unprofessional-looking "Chinese markets," it . . . grew out of a self-imposed industry standard of ethics and conduct.* So that's my answer as to why everybody seems to be doing this, because most of the people were trained the same way. (1996, 21; italics in original)

In fact, the widely held belief that making a Chinese market was unethical was reflected in the Security Traders Association of New York's (STANY) newsletter in 1989. The Security Traders Association is the largest national trade organization for security traders. In reporting on a speech given at an "Ethics Conference" the newsletter misreported a speaker's comments. The correction was as follows:

> In the recently issued STANY NEWSLETTER, *we are certain that you will realize that **** was grossly misquoted* when a portion of his speech was extracted for publication. A corrected copy is featured below.
>
> *As *** and you are all aware, it is clearly UNETHICAL to make a Chinese Market* or to run ahead of an order. (22–23; italics in original)

Most of the communication between Nasdaq traders is on the telephone. Phone calls were used to ensure compliance with the pricing convention as the following audiotape excerpt suggests:

> **Trader 1:** *Who trades CMCAF in your place* without yelling it out?
> **Trader 2:** . . . Sammy
> **Trader 1:** Sammy who?
> **Trader 2:** It may be the foreign department . . .
> **Trader 1:** What?
> **Trader 2:** The foreign didn't realize they had to trade it.
> **Trader 1:** Well, he's trading it in an eighth and he's embarrassing . . .
> **Trader 2:** . . . foreign department
> **Trader 1:** *He's trading it in eighths and he's embarrassing your firm.*
> **Trader 2:** *I understand.*
> **Trader 1:** You know. *I would tell him to straighten up his [expletive deleted] act and stop being a moron.* (24; italics in original)

Additional testimony and taped conversations revealed that when firms continued to violate the pricing convention they were punished in other ways, including the refusal of other market makers to execute deals. The Department of Justice's investigation uncovered other anticompetitive conduct such as "moves on request." A move on request is made when one market maker agrees to change a price quote when requested to do so by another, the purpose being to influence the market in a stock.

bids and asks would be displayed on Nasdaq (the 'quoting convention'). Under the quoting convention, stocks with a dealer spread of 3/4 point or greater are quoted in even-eighths (quarters). Under the quoting convention, market makers used odd-eighth fractions in their bid and ask prices only if they first narrow their dealer spread in the stock in question to less than 3/4 of a point."

The quoting convention has two aspects. Under the first part, stocks with spreads that exceed three-quarters could not be quoted on odd-eighths. This practice ensures that the inside spread of the stock is at least one-quarter because off-eighth quotes are eliminated from the set of possible price quotes. Hence only quarter points (for example, 1/4, 1/2, 3/4) remain and the inside spread, which is the difference between two prices at quarter points, could not be less than a quarter point. Under the second part of the convention, dealers could only use odd-eighth quotes if they narrow their spread to less than $0.75. Market makers are reluctant to narrow their spreads to less than three-quarters of a point because at narrower spreads they are exposed to greater trading risk. In general, at any point in time, a dealer has greater interest in either buying or selling so that a single market maker's quotes do not normally constitute both sides of the spread. Together the two parts of the pricing convention allowed dealers to increase their earnings.

The Department of Justice and twenty-four major dealers reached a settlement on July 17, 1996. The Department of Justice did not assert that dealers had an explicit agreement to collude. However, there was a "conscious commitment to a common scheme," and such agreement is condemned by the Sherman Act. The department's order, which was designed to prevent and detect adherence to the pricing convention, required the firms to tape traders' telephone conversations.

The SEC conducted a concurrent investigation of the Nasdaq market and concluded that the NASD failed to properly oversee trading in the Nasdaq market and enforce compliance with its own rules (1996). The SEC's goal is to promote price competition, and the recommendations in its proposal reflect this goal. Prices should be determined by supply and demand forces and customer order interaction. The proposal included requirements for order handling and execution designed to enhance price competition. Specifically, the SEC ordered Nasdaq dealers to publicly display investor limit orders that are at least 100 shares but not more than 10,000 shares. Dealers were also directed to notify the public of the best available prices.

Despite its conclusion that Nasdaq market makers engaged in abusive practices that suppressed competition, the SEC recognized that various institutional features also could affect the width of spreads. Its report acknowledged the importance of preferencing, internalization, and payment for order flow, concluding that these practices lead to price interdependencies and reduce price competition. Because market makers have a stake in each other's quotes, nonprice forms of competition for order flow provide economic incentives to engage in price fixing. Although direction of and payment for order flow were not prohibited, dealers were strongly chastised for improper behavior, including price fixing and intimidation of rival dealers.[7] The SEC summarized its position as follows: "Vigorous price competition is a hallmark of a free and open market and is critically important to the efficient functioning and regulation of a dispersed dealer market. Because Nasdaq market makers trade securities which are otherwise fungible, price should be a principal means of competition in the Nasdaq market. Any significant hindrance to price competition impedes the free and open market prescribed by the Exchange Act. The investigation found that certain activities of Nasdaq market makers have both directly and indirectly impeded price competition in the Nasdaq market" (1996, 13).

Conclusion

The behavior of security dealers has been closely scrutinized in the 1990s. Recent investigations of the NASD and the Nasdaq market by the Justice Department and SEC suggest that prior to 1996 market makers colluded to fix prices and widen bid-ask spreads. At a minimum, market makers appeared to have adopted a quoting convention that can be viewed as anticompetitive behavior. The purpose of this practice was to increase dealers' profits at investors' expense.

The results of recent academic studies also shed insight into dealer markets and pricing behavior. Important findings suggest that spreads may be large on Nasdaq because dealers had little incentive to compete using price and to narrow the spread. In addition to collusion, institutional features such as preferencing may limit competition for order flow, the effect of which is to produce spreads that are wider than observed in a purely competitive setting.

Through the bid-ask spread market makers are compensated for providing immediacy and liquidity to investors. These dealers also provide other services to their customers such as research. Because they compete along nonprice dimensions, a judgment regarding the competitiveness of the Nasdaq market based solely on the width of the bid-ask spread is problematic. However, the Department of Justice and SEC clearly state that competition on

7. Other evidence of price fixing is reported by the Justice Department and the SEC. For example, price quotes on Instinet, a private electronic market, differed from Nasdaq quotes for the same stocks. Instinet is a proprietary system accessible to the institutional investors and dealers who are subscribers. Price quotes on Instinet may not be directly comparable to those on Nasdaq for several reasons (Woodward 1997). For instance, Instinet prices do not generally include commissions whereas Nasdaq prices do.

BOX 2

The Nasdaq Investigation: A Chronology

May 24, 1994: Approximately 100 security traders meet in New York at the offices of Bear Stearns & Company and are urged to narrow spreads.

May 26–27, 1994: Newspapers report the results of an academic study of the behavior of Nasdaq dealers by Professors William G. Christie and Paul H. Schultz. Christie and Schultz report that market makers attempt to widen spreads by avoiding odd-eighth price quotes. They conclude that the most plausible explanation for this behavior is implicit collusion. The results of the study were released to the press on May 24.

May 31, 1994: Within one week after the release of Christie and Schultz's results, dealer spreads on four prominent Nasdaq stocks narrowed and market makers began entering odd-eighth prices quotes in those stocks. Christie, Harris, and Schultz later reported the change in behavior.

July 1994: Civil lawsuits are filed against thirty-three major dealers alleging collusion.

October 1994: The Justice Department begins an investigation of antitrust law violations.

November 1994: The Securities and Exchange Commission launches an investigation into the NASD's self-regulatory activities.

September 15, 1995: The Rudman Committee submits its report to the NASD. The NASD Board of Governors appointed the committee in November 1994 to review NASD governance and oversight structure. The committee made several recommendations intended to separate the regulatory and oversight functions of the NASD. These recommendations were later implemented.

July 17, 1996: The United States files a complaint alleging that twenty-four major dealers fixed prices, in violation of federal antitrust acts. The same day, the Justice Department settles with the dealers who agree to random taping of trading-desk telephone calls but neither admit nor deny wrong-doing.

August 7, 1996: The SEC concludes that the NASD violated the Exchange Act of 1934, citing deficiencies in market oversight and failure to enforce NASD and federal securities laws. In its settlement with the SEC, NASD agrees to spend $100 million over five years on additional market surveillance.

January 20, 1997: The SEC's new order-handling rules for the Nasdaq market take effect. Market makers are required for the first time to show investors the size and prices for certain orders. The SEC also directs the market to open previously exclusive electronic systems, including Instinet and SelectNet.

December 24, 1997: Thirty securities firms settle a class-action suit alleging price-fixing for $910 million. The agreement is believed to be the largest civil antitrust settlement in U.S. history. Six other firms had previously settled individually for a total of $98.9 million.

Currently: The SEC continues to investigate individual traders in connection with price fixing, and additional civil suits remain unsettled.

price is essential for protecting the public interest. Policymakers can, and in the Nasdaq case did, encourage price competition by removing institutional obstacles.

New rules approved by the SEC and recently implemented in the Nasdaq market, including an open book of limit orders, should enhance price competitiveness. If orders are exposed to the entire market, dealers have greater incentive to improve inside price quotes. However, as dealers focus on price, they may compete less on non-price dimensions and offer fewer services to their clients. Finally, stern warnings and scrutiny from regulators and investors are likely to dampen dealers' incentives to engage in collusive arrangements, whether explicit or implicit. Recent changes in the Nasdaq market will lead to narrower spreads and, in turn, improved market efficiency.

REFERENCES

Ackert, Lucy F., and Bryan K. Church. 1998. "Bid-Ask Spreads in Multiple Dealer Settings: Some Experimental Evidence." Federal Reserve Bank of Atlanta Working Paper 98-9, June.

Amihud, Yakov, and Haim Mendelson. 1986. "Asset Pricing and the Bid-Ask Spread." *Journal of Financial Economics* 17:223–49.

Angel, James J. 1997. "Tick Size, Share Prices, and Stock Splits." *Journal of Finance* 52, no. 2:655–81.

Benston, George J., and Robert L. Hagerman. 1974. "Determinants of Bid-Asked Spreads in the Over-the-Counter Market." *Journal of Financial Economics* 1:353–64.

Bessembinder, Hendrik. 1997. "The Degree of Price Resolution and Equity Trading Costs." *Journal of Financial Economics* 45:9–34.

Bhushan, Ravi. 1994. "An Informational Efficiency Perspective on the Post-Earnings Announcement Drift." *Journal of Accounting and Economics* 18:45–65.

Bloomfield, Robert, and Maureen O'Hara. 1998. "Does Order Preferencing Matter?" *Journal of Financial Economics,* forthcoming.

Christie, William G., Jeffrey H. Harris, and Paul H. Schultz. 1994. "Why Did NASDAQ Market Makers Stop Avoiding Odd-Eighth Quotes?" *Journal of Finance* 49 (December): 1841–60.

Christie, William G., and Paul H. Schultz. 1994. "Why Do NASDAQ Market Makers Avoid Odd-Eighth Quotes?" *Journal of Finance* 49 (December): 1813–40.

Demsetz, Harold. 1997. "Limit Orders and the Alleged Nasdaq Collusion." *Journal of Financial Economics* 45:91–95.

Dutta, Prajit K., and Ananth Madhavan. 1997. "Competition and Collusion in Dealer Markets." *Journal of Finance* 52 (March): 245–76.

Ganley, Joe, Allison Holland, Victoria Saporta, and Anne Vila. 1998. "Transparency and the Design of Securities Markets." Bank of England *Financial Stability Review* (Spring): 8–17.

Glosten, Lawrence R., and Paul R. Milgrom. 1985. "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics* 14:71–100.

Godek, Paul E. 1996. "Why Nasdaq Market Makers Avoid Odd-Eighth Quotes." *Journal of Financial Economics* 41:465–74.

Grossman, Sanford J., Merton H. Miller, Kenneth R. Cone, Daniel R. Fischel, and David J. Ross. 1997. "Clustering and Competition in Dealer Markets." *Journal of Law and Economics* 40:23–60.

Huang, Roger D., and Hans R. Stoll. 1996. "Dealer versus Auction Markets: A Paired Comparison of Execution Costs on Nasdaq and the NYSE." *Journal of Financial Economics* 41, no. 3:313–57.

Kandel, Eugene, and Leslie M. Marx. 1997. "NASDAQ Market Structure and Spread Patterns." *Journal of Financial Economics* 45, no. 1:61–90.

Kim, Olivier, and Robert E. Verrechia. 1994. "Market Liquidity and Volume around Earnings Announcements." *Journal of Accounting and Economics* 12:41–67.

Kleidon, Allan W., and Robert Willig. 1995. "Why Do Christie and Schultz Infer Collusion from Their Data?" Cornerstone Research and Princeton University Working Paper.

Neal, Robert. 1992. "A Comparison of Transactions Costs between Competitive Market Maker and Specialist Market Structures." *Journal of Business* 65:317–34.

Smith, Adam. [1776] 1994. *An Inquiry into the Nature and Causes of the Wealth of Nations.* Reprint, New York: Modern Library.

Stoll, Hans R. 1985. *The Stock Exchange Specialist System: An Economic Analysis.* Monograph Series in Finance and Economics, no. 1985-2. New York: Salomon Brothers Center for the Study of Financial Institutions.

Stoll, Hans R., and Robert E. Whaley. 1990. "Market Structure and Volatility." *Review of Financial Studies* 3, no. 1:37–71.

U.S. Department of Justice. Antitrust Division. 1996. *United States v. Alex. Brown & Sons Inc., et al.—Competitive Impact Statement.* Washington, D.C.: U.S. Department of Justice. Available on-line at <http://www.usdoj.gov/atr/cases/f0700/0739.htm> [July 16, 1998].

U.S. Securities and Exchange Commission. 1996. *Report Pursuant to Section 21(a) of the Securities Exchange Act of 1934 regarding the NASD and the Nasdaq Market.* Washington, D.C.: U.S. Securities and Exchange Commission. Available on-line at <http://www.sec.gov/news/extra/21a.txt> [May 7, 1998].

Woodward, Susan E. 1997. *Price Fixing at Nasdaq? A Reconsideration of the Evidence.* Report commissioned by the Special Studies Division of the Congressional Budget Office. July.

How Powerful Is Monetary Policy in the Long Run?

MARCO A. ESPINOSA-VEGA

The author is a senior economist in the macropolicy section of the Atlanta Fed's research department.
He thanks Frank King for comments on an earlier version of this article and Steve Russell for detailed and thoughtful editorial suggestions.

PRESS REPORTS ABOUT THE STATE OF THE ECONOMY OFTEN GIVE READERS THE IMPRESSION THAT MONETARY POLICY AND THE PEOPLE WHO DIRECT IT ARE QUITE POWERFUL. FOR EXAMPLE, AN ARTICLE IN THE *WASHINGTON POST* IN MARCH 1997 ASSERTS THAT "SECOND TO THE PRESIDENT, ALAN GREENSPAN IS ARGUABLY THE NATION'S MOST POWERFUL PERSON. AS CHAIRMAN OF THE FED, HE GUIDES U.S. MONETARY POLICY, ADJUSTING SHORT-RUN INTEREST RATES."[1]

Many prominent academic economists seem to agree that monetary policy is quite powerful. In reviewing the monetary policy experience of the 1970s, Nobel Laureate James Tobin wrote, "In one respect demand-management policies worked as intended in the 1970s. . . . the decade is distinguished by its three recessions, all deliberately induced by policy. Likewise the expansionary policies adopted to reverse the first two recessions, beginning in 1971 and 1975 respectively, promoted recoveries and in 1977 expansion . . . The major turns in direction conformed to the desires and intentions of the managers of aggregate demand" (1980, 20–21).

Monetary policymakers themselves often describe their role as powerful. Consider, for example, Federal Reserve Chairman Alan Greenspan's testimony before Congress in July of last year. Attempting to explain the influence of monetary policy on the current state of the economy, he stated that "the preemptive actions of the Federal Reserve in 1994 contained a potentially destabilizing surge in demand, short-circuiting a boom-bust business cycle in the making" (1997). Without attempting to explain the full meaning of Greenspan's statement here, it is clear from his language that he believes the Federal Reserve System is powerful enough to have a profound influence on the course of economic activity.

Both Greenspan's statement and Tobin's comments focus on the short-run effects of monetary policy. One might suspect that if Greenspan really is the second most powerful person in the United States then the policy tools he controls must have some long-run influence on the U.S. economy. Ironically, however, although many academic economists and most Federal Reserve policymakers believe that monetary policy is quite powerful in the short run, they also believe that it is virtually powerless in the long run.

Although opinion on this subject is far from uniform, most economists seem to believe that monetary policy can affect the level of real (inflation-adjusted) economic activity—that is, economic variables such as real interest rates, real gross domestic product (GDP) and the unemployment rate—over periods of one or two years. For example, the Fed can create economic recessions or strengthen cyclical expansions. It can do so, according to the conventional view, by increasing the growth rate of the money supply if it wants the economy to grow faster and reducing it if it wants the pace of economic activity to slow. However, increases in the money supply growth rate eventually cause the inflation rate to rise, and decreases in the money growth rate have the opposite effect. When policymakers discuss the short-run effects of monetary policy they usually describe some version of this trade-off between higher inflation, which almost everyone considers undesirable, and desirable changes in other economic variables: higher inflation vs. lower interest rates, lower unemployment, or faster growth in real GDP.

As indicated, however, most economists believe that the long-run effects of changes in monetary policy are very different from their short-run effects. Federal Reserve Governor Meyer has clearly stated the nature of this difference in beliefs, commenting that "there is, to be sure, no trade-off and hence no inconsistency between full employment and price stability in the long run" (1997, 19).

A pair of simple diagrams illustrates the conventional views about the short- and long-run effects of monetary policy. Chart 1 depicts a negatively sloping curve that describes a short-run trade-off between inflation and unemployment. In contrast, Chart 2, which displays a vertical line at the level of full employment, illustrates a scenario in which there is no trade-off between the level of unemployment and the rate of inflation. A low rate of inflation (price stability) is compatible with full employment, but so is a high rate of inflation. If Chart 2 accurately describes the long-run relationship between unemployment and inflation, then changes in monetary policy that lead to changes in the inflation rate have no effect on the long-run levels of unemployment or real output. In the jargon of economists, this diagram describes a situation in which money is superneutral in the long run.[2]

Although the view that monetary policy has real effects in the short run but is superneutral in the long run is widely accepted by academic economists, business economists, and economic policymakers, these groups are not in complete agreement about the ultimate real effects of monetary policy. One source of differences involves the magnitude of the short-run effects. Business economists and policymakers tend to believe that the short-run effects of monetary policy are very large, but most of their academic counterparts see these effects as rather tame and inconsequential. A related difference involves the questions of whether any short-run power that the Fed may have can survive repeated use. Most nonacademics seem to believe that the Fed can use its policy tools as often as it wishes without fear that they will lose their short-run effectiveness. On the other hand, most academic economists believe that repeated, systematic efforts to use the Fed's power to affect real economic activity will grow less and less effective over time.

This article reviews the development of the consensus view that monetary policy can have short-run effects but that it is long-run superneutral. The discussion emphasizes the fact that the basis for this view is the assumption that the only way monetary policy can affect real economic activity is via "money illusion"—that is, by creating changes in the price level that are misunderstood by households and firms and cause them to make bad economic decisions. If monetary policy can affect real economic activity by means other than money illusion then it may be possible for money to be nonsuperneutral in the long run.

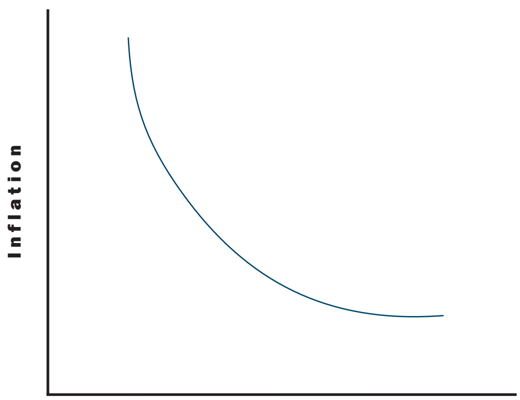
Most economists seem to believe that monetary policy can affect the level of real (inflation-adjusted) economic activity over periods of one or two years.

This article hopes to challenge economists and policymakers to devote more attention to investigating alternative explanations for the real effects of monetary policy—explanations that may imply that money is not long-run superneutral. In order to develop these alternative explanations it is necessary to make very explicit assumptions about the role of money in an economy, how it interacts with real variables and how economic decisionmakers react to the changes in its supply. Different assumptions will turn out to have very important implications for both the nature and the magnitude of the results of policy changes. This point is illustrated in the review of the small but growing branch of the academic literature on the real effects of monetary policy literature that studies models in which money may not be long-run superneutral. In these models the ultimate source of the real effects of monetary policy is the credit markets. By linking monetary policy with the supply of credit these models can analyze an alternative

1. *Linton Weeks and John Berry, "The Shy Wizard of Money: Fed's Enigmatic Greenspan Moves Easily in His Own World,"* Washington Post, *March 24, 1997, sec. A.*

2. *Money is said to exhibit long-run* neutrality *if permanent changes in the level of the supply of money have no long-run effects on real interest rates or the growth rate of real output. In this case, the levels to which prices and other nominal variables will increase are postulated to vary one for one with changes in the level of the money supply. Similarly, an economy is said to display long-run superneutrality if permanent changes in the rate of growth of the money supply have no long-run effects on either real interest rates or the rate of output growth, and the rates of inflation and other nominal variables are postulated to vary one for one with changes in the rate of growth of the money supply.*

CHART 1
A Short-Run Inflation-Unemployment Trade-Off



mechanism for evaluating the long-run effects of monetary policy that does not rely on money surprises.

Another important message of this article is that the very idea that monetary policy is powerful in the short run but powerless in the long run may be internally inconsistent.[3] If monetary policy is indeed as powerful as many informed people seem to believe, then theories of its real effects that rely on money illusion may have to be replaced by theories in which money is not superneutral in the long run.

The Precursors

This section briefly reviews the evolution of two prominent views on the neutrality of money: the Keynesian view and the monetarist view. The discussion begins with a look back at the classical theory that preceded Keynesianism and monetarism. It concludes by describing the clash between the Keynesians and the monetarists and the resulting "unilateral synthesis" of the 1970s.

The Early Quantity Theory. Classical macroeconomic theory, which developed during the late nineteenth and early twentieth centuries, was characterized by its focus on economic fundamentals ("real" economic conditions) such as individuals' propensity to save, the state of technology, and so on. In the classical view monetary policy played no long-run role in determining real economic activity. In particular, it had no long-run effect on the level of real interest rates. Classical theorists acknowledged that monetary policy might have a minor influence over economic activity (particularly interest rates) in the short run. In the long run, however, they viewed money as having a direct influence only on prices.

This early view of the influence of money on prices came to be known as the quantity theory of money. Like many economic concepts, the quantity theory has a rich history of reinterpretations. One of the earliest statements of the theory in its modern form was presented by Fisher (1926). According to Fisher, an economy's general price level is a function of the quantity of money in circulation, the economy's efficiency, or velocity, of circulation (the average number of times a year money is exchanged for goods), and the volume of trade (the quantity of goods purchased with money). Notationally, Fisher expresses the equation of exchange as:

$$M \times V = P \times T,$$

where M is the supply of money, V is the velocity of money, P is the general price level, and T is the total value of transactions or trade. Fisher held that in the long run there was a "natural" level of real economic activity determined by economic fundamentals that could not be affected by increases in the amount of money in the economy. In his words, "An inflation of the currency cannot increase the product of farms and factories . . . The stream of business depends on natural resources and natural conditions, not on the quantity of money" (1926, 155). This hypothesis that there was a natural long-run level of real economic activity, together with the assumption that the only role of money is to serve as a unit of account, formed the basis of Fisher's quantity theory analysis that prices varied proportionately to changes in the quantity of money. According to this equation, if velocity of money and the value of transactions were fairly stable—at least in the long run—as the economy approached its natural level, then changes in the quantity of money would be met with proportional changes in the price level.

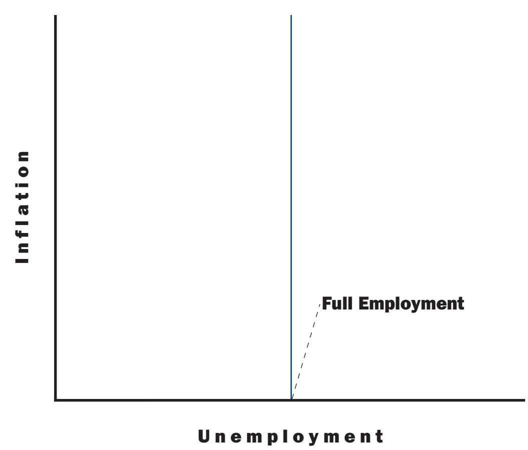
Fisher conceded that monetary policy might have some temporary effects on real economic activity, commenting that "the 'quantity theory' will not hold true strictly and absolutely during transition periods" (1926, 161). In his mind, however, these effects were mainly due to temporary changes in velocity. If velocity was fairly stable in the long run, though, as he assumed, then it had to be the case that prices varied proportionately with the supply of money.

As this description indicates, the basic current consensus on the short- and long-run effects of monetary policy can be traced to the early quantity theorists. According to their view, as represented by Fisher, monetary policy could have temporary real effects but it would be superneutral in the long run. However, even the theory's adherents understood that the theory needed further refinement.[4] This task was undertaken a few years later by Milton Friedman (see below). By the time of the Great Depression, moreover, classical theory had lost much of its popularity and a new, completely different macroeconomic theory was appearing.

The Keynesians and Money: The First Time Around. The first nonclassical macroeconomic theory was the creation of John Maynard Keynes and is laid out in his *General Theory* (1936). One of Keynes's principal goals was to identify the causes of the persistently high rates of unemployment that were afflicting virtually the entire world during the Great Depression. He also sought to identify government policies that could help reduce these high levels of unemployment. Although Keynes's theory discussed the long-term implications of government policies, his focus was on the short run. And although monetary factors played a role in determining real economic activity in his theory (unlike in classical theory), Keynes's analysis emphasized fiscal policy. Keynes believed fiscal policy was the most powerful tool a government could use to lift the economy out of a recession or depression. In fact, his theory predicted that under certain conditions increases in the money supply would be unable to drive interest rates down low enough to stimulate economic activity by generating additional demand for credit. This situation was known as the liquidity trap. In liquidity trap situations money was superneutral even in the short run.

Nonneutrality of Money in the Long Run: The Chicago School. For many years after the Depression and the world war that followed it the question of the long-run implications of government policies received very little attention. One of the early assessments of the long-run effects of monetary policy came from, of all places, the University of Chicago. The university's department of economics—which was and remains perhaps the world's most influential collection of academic economists—has always been associated with the economic principles of the classical economists. The Chicago economics department was instrumental in the development of monetarism, which is usually considered to be a direct descendant of classical macroeconomic theory. In 1951, however, Lloyd Metzler, a prominent member of the economics faculty at Chicago, published a paper describing the long-run implications of central bank open market operations in which he asserted that under some circumstances monetary factors could interact with real variables in such a way as to help determine the level of real economic activity in both the short run and the long run. Metzler wrote that "by purchasing or selling securities, the banking authorities can alter not only the temporary interest rate which prevails while the open-market transaction is taking place but also the rate at which the system will return to equilibrium after the bank's transactions in securities have ceased" (1951, 107). He continued, "By purchasing securities, the central bank can . . . [cause] the system to attain a new equilibrium at a permanently lower interest rate and a permanently higher rate of capital accumulation" (112).

CHART 2
An Inflation-Unemployment Relationship with No Trade-Off



It is important to note that Metzler's conclusion that monetary policy–induced changes in the government's portfolio of liabilities could potentially have long-run real effects does not rely on the monetary authority's ability to produce inflation surprises or on workers' or firms' inability to correctly appraise conditions in the labor market. His analysis is therefore very different from most modern analyses, which view monetary surprises and their impact on naive participants in labor markets or short-run market frictions as the main channel by which monetary policy can affect real economic variables.

Although Metzler's analysis is less than fully rigorous by modern standards, it is worth recalling because it represents one of the first careful descriptions of a mechanism through which monetary policy can have long-run real effects. As noted above, Metzler's conclusions ran counter to the classical tradition of the Chicago school.

3. *If monetary policy has real effects only because of money illusion then it is likely these effects will be very limited in scale. On the other hand, if monetary policy derives its real effects from other sources, then its short-run effects may be relatively large. Thus, limited short-run real effects may go hand in hand with long-run superneutrality while deviations from long-run superneutrality may produce powerful short-run effects.*
4. *In particular, as stated, this description of the quantity theory is really more of an accounting identity than a theory that qualitatively relates money to relevant macroeconomic variables. An accounting identity does not specify what is a given in the analysis and how different variables will change as a result of alternative policy changes. A theory or model, on the other hand, is specific about what is assumed to be exogenous to the model as well as what is determined within it and how different variables react to exogenous changes. A number of complimentary assumptions were really necessary for this equation to spell the list of properties Fisher attached to the quantity theory.*

Perhaps for this reason Metzler's ideas failed to stimulate a research program at Chicago (or elsewhere). Instead, Chicago's monetarists pressed ahead with refinements of the quantity theory of money.

The Monetarists and the Modern Quantity Theory. During the 1950s the monetarists attempted to recover the popularity that classical theory had lost as a result of the Great Depression and the development of Keynesian macroeconomics. The most prominent monetarist was (and remains) Milton Friedman, an economist at the University of Chicago. One of Friedman's first major contributions to monetarism was a refinement of Fisher's quantity theory.

In the classical view monetary policy played no long-run role in determining real economic activity.

As Friedman pointed out, both the size of the money supply and the general level of prices can be considered public knowledge. However, for a theory to be able to make predictions regarding the effects of changes in monetary policy or changes in the price level, it is necessary to establish some assumptions about what differentiates the behavior of money supply from the behavior of money demand. In Friedman's words, "The quantity theory is in the first instance a theory of the demand for money. It is not a theory of output, or of money income, or of the price level. Any statement about these variables requires combining the quantity theory with some specifications about the conditions of the supply of money . . . " (1956, 4).

Friedman's version of the quantity theory is based on the postulate that there is a stable demand for real money balances—that is, for purchasing power in monetary form. He assumes that in the long run the level of money demand depends on economic fundamentals such as real income, the interest rate, and the nature of the technology for conducting transactions. Under this assumption, changes in the nominal supply of money engineered by the Fed have no long-run impact on the real demand for money and consequently lead inevitably and exclusively to changes in the price level. This observation is true both for one-time changes in the money supply and for changes in the rate at which the money supply is growing, which would result in changes in the inflation rate but not in the levels or growth rates of real variables. Thus, one implication of Friedman's restatement of the quantity theory of money is that changes in monetary policy would have no real effects in the long run—that is, money would be long-run superneutral.

Money in Keynesian Analysis: The Second Time Around. As the discussion has shown, early Keynesians focused their attention on fiscal policy. They believed that under normal circumstances changes in the general price level would be both infrequent and relatively inconsequential. As a result, for many years after the Second World War monetarists enjoyed a virtual monopoly over monetary analysis. This situation changed in the mid-1960s, when Keynesians developed a strong interest in the role of monetary policy.

Keynes himself rejected the quantity theory approach to determining the price level. For Keynes, the magnitude of the money supply in the economy was only one of a number of factors affecting the general level of prices. Another important factor was the level of employment. In Keynes's view it was impossible to determine the ultimate impact of a change in the quantity of money on the price level without considering the economy's overall level of employment. More specifically, Keynes believed that "an increase in the quantity of money will have no effect whatever on prices, so long as there is any unemployment" (1964, 295). Since Keynes saw persistent unemployment as the central problem facing industrialized economies, he did not think it would be unusual for economies to go for extended periods of time without observing significant changes in the price level. During the 1950s the general price level was indeed fairly stable. This circumstance lent credence to the Keynesian view that focusing on fiscal policies that might help solve chronic unemployment problems was likely to be more fruitful than devoting a lot of energy to analysis of price level determination.

As the postwar era wore on, inflation began to pick up in both the United States and Western Europe. This development stimulated interest in analyzing the causes of and cures for inflation. In 1958 British economist A.W. Phillips published an empirical analysis of historical data for the U.K. labor market. He hoped to find empirical support for the Keynesian hypothesis that the rate of wage inflation depended on the tightness of the labor market. Phillips found that from 1861 to 1957 the growth rate of nominal wages was negatively related to the rate of unemployment. This "Phillips curve" seemed to link the real side of the economy (the rate of unemployment) to the nominal side (nominal wages). And since wages are the biggest single component of firms' costs, most economists were willing to assume that persistent increases in wage rates would eventually force firms to begin increasing their prices, producing economywide price inflation.[5]

Although Phillips's findings were empirical in nature, they have had a profound and lasting effect on the development of economic theories about the relationship between inflation and real economic variables. As the discussion has shown, Keynesian theory holds that it is possible to use fiscal or monetary policy to increase or decrease the level of aggregate demand and through it the level of employment. The Phillips curve created a link between the level of aggregate demand and the rate of inflation. As a result economic policymakers began to think of demand management policies as involving a trade-off between the unemployment rate (and, more generally, the level of real economic activity) and the inflation rate. And if the Phillips curve was stable over time then this trade-off would exist in both the short run and the long run.

Long-Run Nonsuperneutrality of Money: The Keynesian School. The first attempt to formalize the Keynesian view about the long-run real effects of monetary policy was presented by James Tobin. Unlike the classical economists (but like Metzler), Tobin saw real economic activity in general, and real interest rates in particular, as being determined jointly by economic fundamentals and by monetary policy—even in the long run. In Tobin's words, "Keynes gave reasons why in the short run monetary factors and portfolio decisions modify, and in some circumstances dominate, the determination of the interest rate and the process of capital accumulation. I have tried to show here that a similar proposition is true for the long run. The equilibrium interest rate and degree of capital intensity are in general affected by monetary supplies and portfolio behavior, as well as by technology and thrift" (1965, 684).

Tobin's analysis resembled Metzler's in abstracting from labor markets and concentrating on portfolio adjustments as the channel by which monetary policy could have long-run real effects. According to Tobin's theory, both money and physical capital were elements of an individual's portfolio of savings. For a given real rate of return on capital, an increase in the rate of inflation would make money less attractive and capital more attractive, inducing individuals to reduce their holdings of money in favor of holdings of physical capital. As a consequence, one would observe additional accumulation of capital, a higher capital stock, and a higher output level in the long run.

Long-Run Superneutrality of Money: The Monetarist School. What was the monetarist reaction to Keynesians' claim about the long-run effects of monetary policy? Monetarists did not address Tobin's arguments directly. Instead, they attempted to provide a theoretical underpinning for empirical work of the type conducted by Phillips (1958) that analyzed the relationship between nominal and real variables. Once the theoretical framework was in place, the monetarists used it to explain why monetary policy–induced changes in real economic activity would be short-lived.

While Phillips's statistical evidence involved nominal wages, standard economic theory assumes that households and firms base their employment decisions on real (inflation-adjusted) variables such as real wages, real interest rates, real profits, and so forth. Thus, additional assumptions were needed to reconcile standard economic theory with Phillips's findings. Ironically, the point of departure for this reconciliation was Keynes's observation that "every trade union will put up some resistance to a cut on money wages, however small, but no trade union would dream of striking on every occasion of a rise in the cost of living" (1964, 14–15). Friedman (1968) and Phelps (1967) used Keynes's observation in an attempt to extract some economic content from the statistical relationship discovered by Phillips. Their explanation for the behavior Keynes described was based on two assumptions—one about the nature of monetary policy, and the other about economic decisionmaker's responses to the effects of monetary policy. The first assumption is that increases in the money supply often cause "monetary surprises"—unexpected increases in the rate of inflation. The second assumption was that economic decisionmakers' reaction to monetary surprises often involves temporary money illusion, which is a temporary failure to recognize that there has been an increase in the price level. The basic idea here is that although monetary surprises increase the prices of all goods and services, economic decisionmakers usually notice the effects of these increases on particular prices in which they have a special interest—their wages or the prices of the goods they produce—well before they notice their effects on the overall price level. Until they discover that the overall price level has increased, they mistakenly believe that the increases in the money (nominal) prices of the goods they care about represent increases in the real prices (relative prices) of those goods. This mistaken belief can lead households or firms to make decisions about saving, consumption, work effort, investment, production, and so forth that are quite different from the decisions they would have made otherwise. As a result, by creating monetary surprises monetary policy can influence the level of real economic activity.

One implication of Friedman's restatement of the quantity theory of money is that changes in monetary policy would have no real effects in the long run—that is, money would be long-run superneutral.

Here is a hypothetical sequence of events that illustrates how temporary money illusion can empower

5. *For this explanation to make sense some additional assumptions are required. See Espinosa and Russell (1997) for an explanation of these assumptions.*

monetary policy: Suppose the economy starts out in its long-run equilibrium at its normal inflation rate and its "natural" real rate of interest. Suppose further that the monetary authority begins to increase the money supply at a faster pace than previously. The most immediate consequence of this move will be a drop in nominal interest rates. Friedman explains, "Let the Fed set out to keep interest rates down. How will it try to do so? By buying securities. This raises their prices and lowers their yields . . . In the process, it also increases . . . the total quantity of money. The initial impact of increasing the quantity of money at a faster rate than it has been increasing is to make interest rates lower for a time than they would otherwise have been" (1968, 5–6).

Tobin saw real economic activity as being determined jointly by economic fundamentals and by monetary policy—even in the long run.

The next step in Friedman's chain of causation is that lower interest rates will stimulate spending, and this increase in spending will have a multiplier effect on the overall level of economic activity. Friedman writes, "The more rapid rate of monetary growth will stimulate spending . . . one man's spending is another man's income" (1968, 5–6). From this point, Friedman's analysis can be illustrated using the aggregate demand and aggregate supply (AD and AS) diagram that appears in many textbooks in introductory macroeconomics. The economy starts out in a long-run equilibrium at the intersection of the AD and AS curves. The intersection point represents the long-run equilibrium levels of real output and the price level. In the AD-AS framework, a change in a variable like the market interest rate leads to changes in the market environment that determined the location of the AD and AS curves and consequently produces a shift in at least one of these curves. In this case, the increase in spending that results from the decline in interest rates (which was caused by the increase in the money supply growth rate) is represented by a rightward shift in the AD curve. This increase in aggregate demand produces an increase in output and prices along the original aggregate supply curve.

According to Friedman, this change in the equilibrium will be strictly a short-run phenomenon. As soon as households and firms realize that lower interest rates and faster-rising wages and product prices are also associated with a more rapid rate of increase in the overall price level—as soon, that is, as they realize that real wages and prices have not changed—the households will reduce their supply of labor and the firms will cut back their production. On the diagram, this behavior is represented by a leftward shift in the aggregate supply schedule that exactly offsets the effects of the increase in aggregate demand. In the end, the economy will return to the original long-run natural level of economic activity but a higher rate of inflation. Friedman writes, "Rising income will raise the liquidity preference . . . and the demand for loans; it may also raise prices, which will reduce the real quantity of money. These three effects will reverse the initial downward pressure on interest rates in something less than a year. Together they will tend, after . . . a year or two to return [real] interest rates to the level they would otherwise have had" (1968, 5–6).

Friedman's theory of the short-run effects of monetary policy is sometimes described as the liquidity effect theory. In recent years this theory has been the basis for a great deal of recent research, both empirical and theoretical, about the short-run effects of monetary policy.

As the discussion has indicated, Friedman's liquidity effect theory is based on the belief that in the short run the decisions of firms and households are influenced by money illusion. In this theory, an increase in production and employment occurs not because there has been a change in economic fundamentals but because a more rapid rate of monetary growth has produced a higher rate of inflation. In Friedman's words, "The monetary authority can make the market rate less than the natural rate [of interest] only by inflation" (1968, 7).

The monetary surprises/money illusion hypothesis of Friedman and Phelps seemed to reconcile classical economic principles with the existence of Phillips-type relationships (a negative relationship between inflation and the real interest rate, a positive relationship between inflation and the level of real output, and so forth) created by monetary policy. Under this hypothesis the Phillips curve continued to represent a menu of choices involving trade-offs between real and nominal variables that were available to monetary policymakers—but only in the short run.

The Accelerationist Hypothesis. In tandem with this money illusion hypothesis, monetarists held firm to the classical premise that in the long run all real economic variables such as the real interest rate or the real unemployment rate have a natural level that is determined by economic fundamentals and is completely independent of the nature of monetary policy. In their view, temporary money illusion was the only mechanism by means of which monetary policy could affect real economic activity. It followed from these premises that continuous efforts by monetary policymakers to stimulate economic activity would translate mostly into an ever-increasing rate of inflation. While it might be possible for monetary policy to influence the level of interest rates (in particular) and real economic activity (in general) in the short run, once households and firms recognized that the rate of inflation had increased, the aggregate supply would shift back and the real effects of an increased inflation rate would disappear. Further reductions in interest rates and further stimulus to economic activity could be attained only via further increases in the rate of inflation. In Friedman's words, "Let the monetary authority keep the nominal market rate for a time below the natural rate by inflation. That in turn will raise the nominal natural rate itself, once anticipations of inflation become widespread, thus requiring still more rapid inflation to hold down the market rate" (1968, 7–8). The view underlying this "accelerationist" hypothesis is that while economic decisionmakers cannot be fooled permanently by a single increase in the inflation rate, they can be fooled persistently by accelerating inflation—that is, by a price level that increases over time at an increasing rate.

The Monetarists and the Keynesians in Perspective

The monetarists' persistent attacks on the Keynesians failed to convince the Keynesians that systematic efforts to use monetary policy to affect economic activity would fail. The monetarist argument that attempts to exploit the short-run inflation-unemployment trade-off would lead to accelerating inflation convinced Keynesians that balancing the competing economic goals of keeping inflation low and keeping real economic activity brisk would be harder than they had thought. However, the argument did not convince them that this balancing act was impossible.

To reiterate, during the 1960s Keynesian theorists came to regard the Phillips curve as a menu of options between inflation and unemployment from which policymakers could choose. They assumed that the Phillips curve was stable, which implied that monetary policy was powerful both in the short run and in the long run (that is, that money was not long-run superneutral). To Keynesians, the job of macroeconomic policymakers was to design demand-management policies that would strike the right balance between the competing problems of sustaining robust economic activity and controlling inflation.

Monetarists, on the other hand, believed the economy would be better off if the Federal Reserve supplied money according to a fixed, publicly announced formula and did not try to influence the level of real economic activity. Monetarists such as Friedman and Phelps disagreed with Keynesians regarding the effectiveness and usefulness of demand management. They viewed the "natural rate of unemployment" (the analog of Friedman's natural rate of interest: see above), together with the quantity theory of money, as solid enough arguments to assert beyond doubt the undesirability of activist monetary policy and the long-run superneutrality of money. Monetarists acknowledged the possibility that monetary policy might have short-run effects on employment, interest rates, and private spending, but they believed that these effects arose exclusively from the public's misperception of the impact of changes in the price level. According to the monetarists, the only way the monetary authority could have persistent real effects was by producing an ever-accelerating rate of inflation.

Friedman's liquidity effect theory is based on the belief that in the short run the decisions of firms and households are influenced by money illusion.

The debate between the monetarists and the Keynesians sometimes took the form of disagreements about the slope of the Phillips curve. These disagreements reflected differing views about the effectiveness of monetary policy in the short run versus the long run. During the 1970s, the Keynesians attempted to capitalize on the monetarists' tendency to frame the debate about monetary policy in terms of short- and long-run effects. Their strategy involved reinterpreting the Phillips curve in a way that reconciled the Keynesian and monetarist views of the timing of the inflation-unemployment relationship. This strategy forced the Keynesians to acknowledge that there were limits on the exploitability of the Philips curve.

A key element of the "compromise" offered by the Keynesians was the NAIRU, an acronym that stands for "non-accelerating inflation rate of unemployment" (see Espinosa and Russell 1997). In a diagram of the Phillips curve, the NAIRU is the unemployment rate at which the negatively sloping Phillips curve intersects Friedman's natural rate of unemployment. Monetarists believed that the existence of a natural rate implied that there was no useful trade-off between inflation and unemployment. Keynesians, however, interpreted the natural rate as a long-run constraint that policymakers have to face when trying to exploit an inflation-unemployment trade-off that remained both available and helpful in the short run. This revised Keynesian view of the trade-off was accepted by most policy-oriented economists and most economic policymakers. In the words of Tobin, the "consensus view accepted the notion of a nonaccelerating inflation rate of unemployment . . . as a policy constraint on policy" (1980, 24).

In retrospect it is clear that as much as the monetarists tended to dismiss Keynesian views, in many ways the two schools were not very far from each other—particularly in their analyses of the short-run consequences of monetary policy. These similarities become more evident when the monetarist-Keynesian debate is put in historical perspective. The years since the 1970s have witnessed the development of "neoclassical" macroeconomics—a new school of macroeconomic thought that is based on classical principles even more firmly than monetarism. One of the most influential branches of neoclassical macroeconomics is real business cycle theory. According to real business cycle pioneers such as Kydland and Prescott (1982) and Nelson and Plosser (1982), the cyclical pattern of recessions and expansions has little to do with monetary policy and can be explained almost entirely by "real shocks"—technological developments, changes in tax policy, and other unpredictable changes in economic fundamentals. Thus, the real business cycle theorists believe monetary policy has few or no effects even in the short run.

As economist Joseph Stiglitz points out, "Friedman is, in many ways, closer to the Keynesians than to the real business cycle theorists. He believes, for instance, that there are short run rigidities . . . such that any action by the monetary authority cannot immediately and costlessly be offset by changes in the price level" (1991, 48). Stated differently, the short-run predictions of the Keynesians and the monetarists differed in magnitude but not in direction. Both groups believed in a monetary policy transmission mechanism under which an increase in the money supply leads to an increase in economic activity accompanied by an increase in the general price level. The disagreement about magnitudes could, in principle, have been settled by the analysis of the empirical evidence (although in practice this was no easy task). But as long as the monetarists conceded that monetary policy had some short-run real effects it was impossible for them to make an unequivocal case against the exploitability of the Phillips curve.

In summary, the classical school saw the long-run level of economic activity as being determined independently of monetary policy. Metzler (1951) accepted much of the classical analysis but believed that there were situations in which monetary policy could have long-run real effects. The monetarists focused on money illusion as the only mechanism through which monetary policy could have real effects. In their view, economic fundamentals helped determine an individual's demand for money for transaction purposes. In the absence of surprises this money-demand relationship was fairly stable. It followed that in the long run, changes in the rate of money growth would produce proportional changes in the rate of inflation but would not affect real variables. Tobin (1965) sketched out a formal model in which changes in the rate of money growth could have long-run real effects. In his portfolio-based analysis, a permanent increase in the inflation rate led to more capital accumulation and a lower real rate of return on physical capital. Keynesians implicitly accepted the monetarist view of the role of money illusion in empowering monetary policy. They came to view Friedman's natural rates, which could be interpreted as long-run equilibrium values determined exclusively by fundamentals, as long-run constraints on policy strategies that remained effective in the short run. The short-run policy effect predictions of the Keynesians and the monetarists differed in regard to magnitude and persistence but not in regard to direction. Both schools agreed that in the short run a higher rate of money growth was associated with a higher rate of inflation, a lower real interest rate, and a spurt in economic activity. Keynesians did not themselves develop theories in which monetary policy was powerful in the short run but money was superneutral in the long run. Instead, they implicitly accepted the theoretical framework provided by their critics, the monetarists, although the two schools continued to disagree about some of the implications of this framework.

To this day, much of the economics profession continues to regard Keynesians' acceptance of the monetarists' position regarding long-run superneutrality as proof that there has been a rigorous scientific synthesis of the two theories. As discussed below, however, any synthesis of this sort is likely to be internally inconsistent.

The Neoclassical School

The arguments made by Friedman and Phelps against Keynesian theory were extended by economists such as Lucas (1972) and Sargent and Wallace (1976), who became the founders of the neoclassical school.[6] Lucas's 1972 article set the standards for neoclassical macroeconomics and, to a large extent, for all modern macroeconomics. The two pillars of his analysis were his assumption that economic decisionmakers had rational expectations and his use of a dynamic general equilibrium model. A dynamic general equilibrium model is a model that takes into account the intertemporal nature of many economic decisions and recognizes that economic variables interact with each other. Therefore, to determine the consequences of a postulated policy experiment one has to consider the relevant economic variables simultaneously and through time.

Lucas's article presented a very rigorous description of a situation in which (1) money is superneutral in the long run, and (2) the short-run real effects of monetary policy are bound to be rather limited, even in a scenario involving accelerating prices. A first step toward understanding Lucas's analysis is to recognize a key distinction between his assumptions and those of Friedman and Phelps. A simple way to describe this distinction is to say that the Friedman and Phelps analysis permitted persistent money illusion while Lucas's analysis ruled out persistent money illusion. Stated differently, the Friedman and Phelps analysis was based on the assumption that changes in prices or wages could cause households and firms to make "bad" economic decisions—decisions they would not have made if they had used available economic information more efficiently and had displayed more flexibility in reacting to the changes. Lucas, in contrast, assumes that the public processes economic information as efficiently as possible: in particular, individuals base their current decisions on the best possible forecasts of future events. His description of this decision-making process includes specific assumptions about how people form their economic expectations.

In Lucas's model there are two types of changes in prices: temporary changes in prices in particular industries, which are caused by short-run fluctuations in the demand for the goods produced by those industries, and changes in the price level, which are caused by changes in the growth rate of the money supply. There are also two types of changes in the growth rate of the money supply: systematic, permanent changes in the average (long-run) money growth rate and unsystematic, temporary changes in the current (short-run) money growth rate. The systematic changes result from deliberate changes in policy by the central bank; they produce a permanent increase in the average rate of inflation. The unsystematic changes result from errors in the implementation of the central bank's operating procedures. They do not reflect deliberate policy decisions, and they do not affect the long-run average money growth rate or inflation rates. They do, however, produce temporary changes in the current rate of inflation.

As has been indicated, Friedman and Phelps had assumed implicitly that economic decisionmakers have access to complete economic information but fail to use it efficiently. Lucas, on the other hand, assumes explicitly that decisionmakers use any information available to them in the most efficient way but do not always have access to complete information. The particular aspect of the economy that Lucas assumes decisionmakers do not have complete information about is the relationship between changes in the relative prices of the particular goods they produce and changes in the overall price level. This information gap is important because fully informed decisionmakers will react very differently to changes in the prices of their goods that represent changes in relative prices—that is, to situations in which the prices of their goods change but the general price level remains constant, or situations in which the general price level changes but the prices of their goods change by a larger or smaller proportion—than to changes in the prices of their goods that simply follow along with changes in the overall price level. More specifically, decisionmakers have no incentive to increase their work effort and production in response to increases in the overall price level for the same reason that one would not be any happier if a doubling of salary coincided with a doubling of the price of every good purchased. On the other hand, it makes sense for a person to increase effort and output if the relative price of the good produced has increased. Under Lucas's assumptions any such increases in effort and output will be temporary because the demand fluctuations that induce them are also temporary.

Now suppose that at a given moment in time, and in the absence of any changes in the economy's fundamentals, the overall inflation rate increases because of an unsystematic increase in the money supply. As the overall inflation rate increases, prices in every sector or industry increase. However, individuals are unable to tell, immediately, whether the price increases affecting their sector are relative or absolute changes. The reason is that people are assumed to have better information about prices of the goods and services in their industry than about the many different prices that figure in the overall price level. This lack of complete information about the overall level of prices leads people to assume that at least part of the increase in the price of their product has been caused by an increase in its relative price. As a result, they increase their work effort and production.

To Keynesians, the job of macroeconomic policymakers was to design demand-management policies that would strike the right balance between the competing problems of sustaining robust economic activity and controlling inflation.

The situation just described seems quite consistent with the Keynesian notion that there is a short-run tradeoff between rate of inflation and the level of economic activity. But does this trade-off indicate that monetary policy is powerful, in the sense that the central bank can use it to control the level of economic activity? Is this a model of the "tightrope walk" that aggregate demand managers are often described as having to perform? If the central bank in the model can use monetary policy actions to exert continuous and repeated influence over individual decisions concerning work effort and production, then the answer to these questions may be yes.

This situation turns out not to be possible, however. Suppose that the central bank announces a permanent change in the average growth rate of the money supply. Lucas's assumption that people have rational expectations implies that they understand the nature of the relationship between money growth and inflation. As a

6. For a detailed nontechnical description of Lucas's contribution see Espinosa and Russell (1997).

result, they will not increase their work effort or production in response to the resulting increase in the average rate at which prices change. Thus, permanent increases in the money growth rate have no effect on the level of output or employment, while temporary increases in the money growth rate will produce temporary increases in both output and employment.

Thus, in Lucas's model the rational expectations assumption implies that systematic changes in monetary policy should not have real effects. The rigorous nature of Lucas's analysis made his argument seem very convincing. It is important to note, however, that the argument also depends on Lucas's assumption, which is built into the structure of his model, that the effects of monetary policy on real economic activity are caused only by money illusion.

To this day, much of the economics profession continues to regard Keynesians' acceptance of the monetarists' position regarding long-run superneutrality as proof that there has been a rigorous scientific synthesis of the two theories.

Lucas's argument can be illustrated further by returning to the context of his model and exploring its implications in a somewhat less rigorous way. Suppose that the central bank in his model attempts to exploit the apparent trade-off between inflation and output by increasing the average money growth rate without making any announcement that it has done so. It is hoping that people will make inflation-forecasting mistakes because they will not recognize that any policy change has occurred. The increase in money growth will, of course, produce a permanent increase in the average inflation rate. Initially, people will mistake this systematic, policy-induced increase in the inflation rate for an unsystematic inflation rate increase caused either by a temporary demand disturbance or by an error in the execution of the original monetary policy rule. Since they will not be sure which of these two types of unsystematic increase has occurred, their work effort and production will rise (see above). Soon, however, people will start to recognize that there is a pattern to the unusually high rates of inflation they are observing. As a result, they will begin to think it less and less likely that the next above-average increase in the inflation rate was caused by a demand disturbance, and they will begin to cut back on their above-normal production and work effort. Ultimately, they will realize that the central bank has changed policy in a way that has caused the average inflation rate to increase. At this point, the increase in the average inflation rate will no longer have any effect on work effort and production.

The scenario just described suggests that systematic changes in monetary policy may have substantial short-run effects but no long-run effects, just as the monetarists argued and just as the Keynesians ultimately conceded. Suppose, however, that the central bank tries to repeat its short-run success by changing the average inflation rate from time to time in response, say, to other changes in the state of economy. In the real world people learn from past mistakes: as a result, each time the central bank engineers another systematic change in the inflation rate people will catch on to the policy change more quickly and the effects of the change will disappear more quickly. At some point, moreover, people will figure out which events motivate the central bank to change policy; they will then be able to detect policy changes as soon as they occur. At this point the policy changes will no longer have any effects, even in the short run.

These modified rational expectations assumptions about the way people obtain and use information seem consistent with one's economic intuition about the behavior of actual households and firms. In real-life economies, most people have a very good "micro" picture of the status of their firm or industry but a fairly fuzzy "macro" picture of the state of the economy at large. However, once they start getting surprised by unexpected price changes that make their decisions work out badly they become more interested in identifying the causes of changing prices. They start to use any information available to them to try to anticipate changes in the price level and distinguish them from changes in relative prices. As a result, future price level increases have less and less surprise effect. This sort of intelligently adaptive behavior is the real-life analogue of Lucas's formal assumption that economic decisionmakers have rational expectations.

Lucas's argument, and the closely related arguments of neoclassical economists such as Sargent and Wallace (1976), left Keynesians with only two intellectually legitimate choices. First, they could have tried to capture their intuition about the effects of monetary policy in a rational expectations general equilibrium model in which money was not long-run superneutral because monetary policy derived its real effects from some source other than monetary surprises. Many economists expected this approach. Sargent, for example, writes that "in the early 70's, I thought that Modigliani, Solow, and Tobin—our heroes in those days—were missing the boat by resisting the intrusion of rational expectations into macroeconomics, instead of commandeering it. Despite the appearances of its early incarnations like Lucas's 72 JET paper, the canons of rational expectations models . . . were evidently wide enough to include Lucas's brand of monetarism or, just as readily, accommodate the completion of Tobin's criticism of monetarism by fully bringing to bear the logic of Modigliani and Miller. Modigliani, Solow, and Tobin chose not to commandeer the movement, and left it to Kareken, Wallace, Chamley, Bryant and others to draw out many of the nonmonetarist implications then waiting to be exposed." (1996, 545). In retrospect it seems clear, as this quotation indicates, that an important reason Keynesians did not pursue this strategy was because they mistakenly believed that rational expectations implied long-run superneutrality of money.

Another alternative for Keynesians might have been to concede that monetary policy was long-run superneutral but to argue that frictions of various sorts might allow monetary policy to have real effects in the short run. Taylor's work on staggered contracts (1980), his work on slow adjustment of prices (1994), and the work of Ball and Mankiw (1995) concerning "menu costs" are illustrations of this line of research. This research has faced criticisms on two fronts. First, there is little empirical evidence to support this type of nominal rigidities (see, for example, Wynne 1995). Second, there has not been a clear explanation as to why these frictions could prevent people from changing their behavior so as to offset the effects of systematic changes in monetary policy. For example, what prevents individuals from relying on mechanisms such as indexing of nominal contracts to guard against the potential negative effects of nominal rigidities?

Most Keynesians chose to ignore the neoclassical critique and the potential problems with short-term frictions. They continued to claim that monetary policy had powerful short-run effects, while accepting the monetarist critique that it was powerless in the long run. For the most part, economists outside academia—business economists and economic policymakers—have adopted this "Keynesian consensus" view. To the extent that either group of economists has attempted to justify this view, they have done so by arguing that rational expectations is a sensible assumption only in the long run. In the short run, they argued, people could and often did misread the nature and effects of changes in monetary policy.

What is wrong with the Keynesian consensus? Lucas points out that, while it may seem reasonable on its face, it suffers from serious logical problems. Com-menting on Tobin's description of the Keynesian consensus, Lucas writes, "Here we have Model A, that makes a particular prediction. We have model B, that makes a strikingly different prediction concerning the same event. The event occurs, and Model B proves more accurate. A proponent of model A concludes: 'All right, I "accept" Model B too.' Consensus economics may be a wonderful thing, but there are laws of logic which must be obeyed . . . These models gave different predictions about the same event because their underlying assumptions are mutually inconsistent. If the Friedman-Phelps assumptions are now 'accepted,' which formerly accepted Keynesian assumptions are now viewed as discarded? Tobin does not say" (1981, 560–61). Lucas goes on to spell out the monetarist (model B)-Keynesian (model A) consensus, as viewed through the Keynesian glass. He writes, "Though I refer to Tobin as 'evading' a central issue, I do not think he sees it this way at all. He writes as though he is willing to concede the 'long-run' to Phelps and Friedman [the Monetarists], claiming only the 'short-run' for Keynesians. Where is the conflict?" (561). Lucas goes on to explain that the long run consists of a sequence of short runs. If a policymaker conducts short-run policy by choosing an annual money growth rate based on model A (the Keynesian model) every year, then he or she has implicitly used model A to pick the average rate of money growth for the long run. It is logically inconsistent to pretend that the long-run average money growth rate using model B (the monetarist model) can be a guide. Suppose, for example, that the central bank decides that in the long run the optimal growth rate of the money supply is 5 percent per year. However, it decides on the basis of short-run considerations that it would be a good idea to increase the money growth rate to 6 percent for the coming year. The same thing happens again in the following year, and in the year after, and so on. The end result is a departure from the optimal long-run money growth rate that may have adverse consequences for the economy. Thus, Lucas observes that "if we concede that Model A gives us an inaccurate view of the 'long-run,' then we have conceded that it leads us to bad short-run situations as well" (560–61).

Lucas's 1972 article set the standards for neoclassical macroeconomics and, to a large extent, for all modern macroeconomics.

Monetary Policy after Lucas. Starting in the late 1960s, Keynesian economic theory was the victim of a succession of setbacks, including the monetarist critique of Friedman and Phelps, the combination of high inflation and high unemployment that the United States experienced during the 1970s, and the neoclassical critique of Lucas (1972) and Sargent and Wallace (1976). As Keynesian theory lost ground in the academic community, so did belief in the power of monetary policy. In fact, much of the early work by neoclassical economists followed Lucas (1972) by constructing models that made debating points against the Keynesians by demonstrating that systematic changes in monetary policy would have no real effects, even in the short run. Unsystematic policy actions might have a short-lived influence on the level of economic activity, but any attempt to use systematic changes in policy to exploit this influence would be frustrated by changes in the expectations of the public.

The fact that the model described by Lucas (1972) became the "industry standard" in neoclassical theory has encouraged other neoclassical economists to focus on models that display long-run superneutrality of money. In recent years, the most popular vehicle for research on monetary policy by neoclassical economists has been the real business cycle model. In this model money is long-run superneutral, but temporary changes in monetary policy can generate small "liquidity effects" of the sort described in Friedman (1968). (See, for example, Lucas 1990; Christiano and Eichenbaum 1991, 1992; Fuerst 1992; Dow 1995).

Lucas's article presented a very rigorous description of a situation in which (1) money is superneutral in the long run, and (2) the short-run real effects of monetary policy are bound to be rather limited, even in a scenario involving accelerating prices.

During the mid-1970s economist Harry Johnson, reviewing what he labeled the Keynesian revolution and the monetarist counterrevolution, commented that "the monetarist counterrevolution has served a useful purpose, in challenging and disposing of a great deal of the intellectual nonsense that accumulates after a successful ideological revolution . . . If we are lucky, we shall be forced as a result of the counterrevolution to be both more conscious of monetary influences on the economy and more careful in our assessment of their importance" (1975, 106).

In fact, the monetarist counterrevolution had mixed effects on economists' views concerning the importance of monetary policy. On the one hand, monetarist arguments convinced many Keynesians that monetary policy had many of the same powers that they had attributed to fiscal policy. On the other hand the monetarists, as has been pointed out, completely dismissed the possibility that monetary policy might have long-run real effects. To the extent that Keynesians conceded this point they were also conceding that the importance of monetary policy was quite limited.

As shown above, the period of the monetarist counterrevolution was also a period when a few economists began to try to identify explicit mechanisms that would allow monetary policy to have long-run real effects. Metzler (1951) and Tobin (1965) developed theories that allowed the Keynesian, conventional wisdom to be extended to the long run. These theories allowed permanent increases in the money supply growth and inflation rates to be causally associated with permanently lower real interest rates and permanently higher levels of output.

Lucas's (1972) work suggested that macroeconomic theories of all sorts were in need of reevaluation. The theories of Metzler, Tobin, and the ones derived from Phillips's analysis were no exception. Lucas's interpretation of the Phillips curve analysis has been described above. The next section reviews subsequent research that tries to reformulate Metzler's and Tobin's theories using the neoclassical methodology Lucas introduced. This research indicates that departures from long-run superneutrality are possible because monetary policy does not necessarily derive all (or any) of its power from money illusion. Instead, changes in monetary policy may have lasting effects because it affects the supply of or the demand for credit.

Some Neoclassical Models That Deliver Long-Run Nonsuperneutralities

This section looks at the three challenges facing economists who want to develop neoclassical models that deliver results similar to those of Metzler, Phillips, and Tobin. The first challenge is simply to construct a plausible neoclassical model in which money is not long-run superneutral. The second challenge is to identify a mechanism under which the departures from superneutrality work in the "right direction," that is, a mechanism that allows increases in the money supply growth rate to be causally associated with lower real interest rates and higher levels of output. The third challenge is to find a mechanism that has some hope of generating departures from superneutrality that are large enough to have practical importance.

The Tobin Effect. An answer to the first challenge is to rely on the credit market as the ultimate source of the real effects of monetary policy. In this respect one could follow Tobin (1965). Tobin's analysis is based on the idea that the increase in the inflation rate that is induced by an increase in the money supply growth rate increases the supply of credit at any real interest rate. It does so because when the inflation rate rises money becomes a relatively unattractive asset, and the public wishes to cut back on its money balances and increase its holdings of bank accounts, bonds, stock, and other financial assets. Thus, there is a decrease in the demand for money and a matching increase in the supply of credit. The Tobin effect mechanism allows a permanent easing of monetary policy (a higher money growth rate) to lead to a higher inflation rate, a lower real interest rate (due to the increased supply of credit), and a higher level of output (due mostly to an increase in the capital stock).

Many economists believe that financial intermediation is one of the most important channels through which changes in monetary policy affect the economy (see for instance Bernanke and Gertler 1995). The central bank may be able to affect the composition of financial intermediaries' portfolios without relying on monetary surprises. Thus, permanent changes in monetary policy may affect financial intermediaries in a fundamental way and may have long-run real effects. It follows that a natural environment in which to analyze the Tobin effect would be one in which financial intermediaries were explicitly developed.

The starting point for assessing this possibility should be a realistic model of financial intermediation. Bencivenga and Smith (1991) were among the first economists to include financial intermediaries in a dynamic general equilibrium macroeconomic model. The Bencivenga-Smith intermediaries are similar to actual intermediaries in accepting deposits from, and lending to, a large number of individuals. They are also similar to actual intermediaries in making loans that are less liquid than the deposits they accept. As a result, they are forced to hold a liquid asset (money) on reserve to cover sudden deposit withdrawals.

In the Bencivenga-Smith model individuals could, in principle, manage their own asset portfolios (as in Tobin 1965). However, the financial intermediaries have an actuarial advantage over individuals in structuring a portfolio. Consequently, under most circumstances people will prefer to delegate this activity to financial intermediaries. Although Bencivenga and Smith's work contains the elements needed to pursue an analysis of the long-run effects of permanent changes in monetary policy, their analysis concentrates on the long-run implications of financial intermediaries for an economy's long-run performance. Based on the Bencivenga and Smith model, Espinosa and Yip (1998) study the growth-inflation implications of alternative fiscal and monetary policies in the presence of financial intermediaries. Espinosa and Yip can, thus, draw some qualitative lessons on the Tobin effect in a dynamic general equilibrium model that explicitly models financial intermediaries. Before listing their findings, it is useful to briefly review some recent empirical results on the relationship between inflation and growth.

Inflation and Growth. In part because money has been assumed to be long-run superneutral, there has not been much research on the long-run relationship between inflation and growth. Recently, however, interest in theoretical and empirical analysis of this relationship has revived. The empirical findings are not always in agreement. DeGregorio (1992) and Barro (1995) uncover a significant negative correlation between inflation and economic growth. On the other hand, Bullard and Keating (1995) and Bruno and Easterly (1998) do not find strong support for such an inverse relationship. Bullard and Keating find that the direction of the growth-inflation relationship depends crucially on the initial level of the inflation rate. In countries in which the rate of inflation starts out relatively low, a permanent increase in the inflation rate actually increases the long-run level of economic activity. Only for countries with relatively high initial inflation rates do Bullard and Keating find that permanent increases in the rate of inflation negatively affect long-run growth. These findings are partly confirmed by Bruno and Easterly, who are able to find an inverse relationship between inflation and growth only when the rate of inflation exceeds some critical value.

Clearly, these empirical studies do not settle whether monetary policy can have real effects that do not spring from monetary surprises and whether these effects are likely to be of the type described by Tobin, with higher inflation being associated with higher rates of growth. Espinosa and Yip (1998) address these questions in a model based on the model developed by Bencivenga and Smith (1991). Their analysis emphasizes the point (to be made very explicitly below) that fiscal and monetary policy are inevitably linked by the government budget constraint. In their model, monetary policy consists of changes in the growth rate of the money supply that are necessitated by changes in fiscal policy—specifically, by changes in the government budget deficit as a fraction of GDP.

Keynesians mistakenly believed that rational expectations implied long-run superneutrality of money.

Espinosa and Yip show that their model can produce the positive long-run relationship between inflation and growth that was predicted by Tobin. However, it is also possible for the model to produce situations in which lower rates of inflation result in higher rates of growth. The direction of the inflation-growth relationship depends on, among other things, the initial inflation rate, the degree of risk aversion of the average depositor, and the size of the government budget deficit. Thus, the Espinosa-Yip analysis provides a theoretical framework that helps reconcile the conflicting empirical findings about the direction of the long-run relationship between inflation and growth that were described in the preceding subsection.

Fiscal Policy and Open Market Operations. The Tobin effect has a potential drawback as a theory of the real effects of monetary policy (see, for example, Danthine, Donaldson, and Smith 1987). The shift in the credit supply curve produced by an increase in the inflation rate is essentially equal to the reduction in money demand that the increased inflation induces. Money demand is quite small (a small fraction of total output, or total assets, and so forth) and statistical evidence (for example, Hoffman and Raasche 1991) suggests that it is not very sensitive to changes in the inflation rate. As a result, the Tobin effect of moderate changes in the inflation rate on real interest rates and output is likely to be small.[7]

An alternative mechanism for linking monetary policy and the supply of credit has been developed by Sargent and Wallace (1981). This mechanism is based on the fact that changes in monetary policy affect the government's stream of revenues and thus necessitate changes in fiscal policy. To gain a better understanding of this mechanism, it is useful to review some basic elements of a government's budget constraint.

An important premise of the research described in this section is that both fiscal and monetary policy actions are constrained by the government's need to finance its expenditures. Consequently, these two types of government policy cannot be devised or executed independently from each other. The government of a country must decide on the level of government spending to finance domestically, how much of its domestic financing will rely on current taxes, and how much will take the form of newly issued debt. Stated differently, it is the government budget deficit that determines the need for new issues of government debt. Since government borrowing competes with private borrowing in the credit market, the amount of government borrowing is likely to influence the level of real interest rates.

Government policy concerning taxes, debt, and deficits is usually described as fiscal policy. The analysis just presented suggests that fiscal policy may affect real interest rates. However, monetary policy has a fiscal policy aspect to it because it may play a role in determining the size of the government budget deficit. To the extent that monetary policy has this effect, this analysis suggests that it will also have an impact on real interest rates. Thus, monetary policy may influence the real economy in ways that do not involve inflation surprises. If this influence can persist in the long run then money may not be long-run superneutral.

In practice, monetary policy is carried out via open market operations. Open market operations produce changes in the composition of the government's portfolio of liabilities—debt (bonds and bills) versus money.[8] Given the amount of government bonds currently outstanding, the government must decide what fraction of these bonds (if any) it will "monetize" by purchasing them with newly created currency. This decision, which determines the composition of the government's liability portfolio, also determines the amount of outstanding government debt in the credit markets and consequently has an impact on the market real rate of interest. More specifically, changes in the growth rate of the money supply affect the volume of government revenue from currency seigniorage (the "inflation tax"). Sargent and Wallace (1981) assume that the government's primary (net of interest) budget deficit is fixed by the tax and spending decisions of Congress and is not affected by changes in monetary policy. Consequently, the only way the government can offset the changes in its revenues that are caused by changes in monetary policy is to change the size of the national debt. Thus, this mechanism can be thought of as the neoclassical successor of Metzler (1951) (because of Metzler's emphasis on open market operations as the mechanism through which non-long-run superneutrality results could be attained).

The size of the national debt has substantial effects on the state of the government budget. On the one hand, the government has to pay interest on the debt. On the other hand, as the economy grows the government can allow the national debt to grow at the same rate without increasing the size of the debt relative to the economy.[9] The relationship between these two factors determines whether debt service is a financial burden for the government or whether the existence of the national debt actually increases the amount of government revenue.

To see why the latter situation is possible, suppose the government borrows just enough each year to keep the debt-GDP ratio constant. If the economy is growing, it will increase its borrowing each year by an amount that causes the real national debt to grow at the same rate as real GDP. Although the government will have to use some of the proceeds of this new borrowing to pay the interest on the current debt, if the real (also inflation-adjusted) interest rate on the debt is lower than the real GDP growth rate then the government will have funds left over to use for other purposes. In this case, the national debt actually provides the government with revenue on net. This source of revenue is sometimes referred to as bond seigniorage.[10] The difference between the real growth rate and the real interest rate is the net real amount that each real dollar of debt contributes to the government budget each year.

If the real interest rate on the government debt is higher than the output growth rate then the government's new borrowing will not be enough to cover the interest on the existing debt. As a result, some of this interest will have to be covered by funds from other sources. In this case the national debt is a financial burden for the government. (One can think of this as a case in which bond seigniorage revenue is negative.) The difference between the real interest rate and the real growth rate is the net real amount that each real dollar of debt costs the government each year.[11]

Once it is known whether the national debt is a source or a use of government funds one is in a position to determine how a change in the size of the national debt will affect the government's budget position. Other things being equal, an increase in government borrowing that increases the size of the national debt represents an increase in the quantity of credit demanded at each real rate of interest and will consequently produce an increase in the real interest rate. If the real interest rate is relatively high, so that the debt is a burden on the government budget, then the combination of a larger debt and a larger unit cost of financing the debt means that the debt will definitely become costlier to the government. Conversely, a smaller debt that will result in a lower real interest rate will reduce the government's costs. As a result, when the government cuts the money supply growth and inflation rates and loses money from currency seigniorage, it must compensate by cutting back on its borrowing and driving the real interest rate down. As a result, tighter monetary policy produces lower real interest rates and a higher level of output.

Suppose, on the other hand, that the real interest rate is relatively low, so that the national debt is a source of revenue for the government. In this case, a given change in the size of the debt (say, an increase) can either increase or decrease government revenue from bond seigniorage. An increase in the size of the debt tends to cause bond seigniorage revenue to increase: this is the "tax base effect" of the increase. On the other hand, an increase in the government debt drives the real interest rate closer to the output growth rate and reduces the real seigniorage revenue produced by each real dollar of debt. This is the "tax rate effect" of the increase. If the tax base effect is stronger than the tax rate effect then an increase in the size of the debt will increase the government's bond seigniorage revenue; otherwise, the amount of revenue will fall.

The tax rate effect tends to be largest when the government debt is large, because in this case any change in the real interest rate affects the revenue produced by a large volume of debt. Conversely, the tax base effect tends to be largest when the real interest rate is low because each dollar of debt generates a lot of revenue. A low real interest rate tends to be associated with a small volume of government debt, since when the real interest rate is low private credit demand is high and private debt crowds out government debt. Conversely, a high real interest rate tends to be associated with a large government debt. As a result, when the real interest rate is relatively low—well below the output growth rate—an increase in the size of the national debt tends to increase bond seigniorage revenue while when the real interest rate is higher an increase in the size of the debt tends to decrease the amount of revenue.

One can now put all the pieces of this story together to determine the possibilities for the long-run real effects of monetary policy. If the real interest rate is higher than the output growth rate, or lower than the output growth rate but not too much lower, then an increase in the size of the national debt decreases government bond seigniorage revenue and vice-versa. Thus, a decrease in the money growth and inflation rates that reduces government revenue from currency seigniorage will force the government to reduce the size of its debt and will drive the real interest rate down. This is the scenario described by Wallace (1984); it has the implication that monetary tightening will increase the level of real GDP in the long run. On the other hand, if the real interest rate is substantially below the output growth rate then a decrease in the money growth and inflation rates will allow the government to increase the size of its debt and will drive the real interest rate up. This is the scenario described by Espinosa and Russell (1998a, b). It is similar to the Tobin effect in having the Keynesian, or conventional, implication that a monetary tightening will reduce the level of real output.

The proposition that monetary policy does not have long-run real effects is far from unequivocally established.

Historically, the average real interest rate on U.S. government debt has been well below the average U.S. output growth rate. This situation makes Espinosa and Russell's Keynesian scenario seem plausible empirically. An additional reason why the scenario is appealing is that it weakens the link between the size of the money supply and the size of the shift in the credit supply curve that is induced by a change in monetary policy—the link that keeps the Tobin effect small. Although the fact that the money supply is small relative to GDP means that a change in the inflation rate will have a relatively small impact on government revenue (from currency seigniorage), if it takes a relatively large increase in the real interest rate to produce a substantial decrease in government revenue from bond

7. *Of course, if fiscal and monetary policy interactions led not only to long-run output level changes but to output growth changes, the Tobin effect could be of more significance.*
8. *In principle, of course, there exists the possibility that such a swap of liabilities results in no effects either real or nominal, either in the short or long term (a case made by Wallace 1984 and Sargent and Smith 1987 but not reviewed here), but under most circumstances it will.*
9. *The debt-GDP ratio cannot continue to grow forever. Otherwise, at some point the debt would get so large relative to households' income that it would be impossible for them to save enough to hold it.*
10. *This term seems to have been first used by Miller and Sargent (1984).*
11. *Thus, if the real interest rate is 2 percent higher than the real growth rate then each dollar of debt costs the government two cents each year, adjusted for inflation.*

seigniorage then the resulting change in the real interest rate and the level of output could still be large.

Why might it take a large change in the real interest rate to produce a substantial change in the revenue from bond seigniorage? When the real interest rate is low, the tax rate effect and the tax base effect tend to work against each other. As a result, the net change in the amount of revenue that is produced by a change in the real interest rate can be quite small. In fact, there is always a range of real interest rates over which the two effects offset each other almost perfectly. Over this range, the ratio of the change in the real interest rate to the change in the amount of revenue it produces will be extremely large.

The possibility that monetary policy has substantial long-run real effects deserves more attention from economists and policymakers.

The bottom line here is that, at least in principle, the Espinosa-Russell variant of the Sargent-Wallace "unpleasant arithmetic" can give us just what is needed: a theory that explains how a moderate but permanent increase in the money supply growth and inflation rates might result in a fairly large decrease in the real interest rate and a fairly large increase in the level of output.

Before concluding this section it is important to emphasize that the research just reviewed composes a relatively small part of the growing academic literature that studies the long-term effects of monetary policy in neoclassical models. Related work in this area includes Haslag (1998), Bhattacharya and others (1997), Schreft and Smith (1997), and Bullard and Russell (1998a, b). One implication of this line of research is that monetary economists may have spent too much time trying to forge direct links between changes in monetary policy and changes in the unemployment rate and the output growth rate. Instead, they perhaps should be devoting more effort to understanding the relationship between monetary policy and the economic fundamentals that drive saving and production decisions and also to exploring the relationship between monetary policy variables and "real" macroeconomic variables such as the government deficit, real interest rates, reserve requirements, and other variables that link the money market to the credit market.

Conclusion

This article has reviewed the history of the view that monetary policy has real effects in the short run but no such effects in the long run (so that money is long-run superneutral). This view grew out of a debate between the adherents of two influential schools of macroeconomic thought, the monetarists and the Keynesians. The final result of this conflict was a unilateral, Keynesian-produced synthesis that developed during the 1970s. Under this synthesis the Keynesians accepted the monetarists' view that money was superneutral in the long run but continued to disagree with them about the magnitude and desirability of the short-run effects of monetary policy on real interest rates, real GDP, unemployment, and other real variables.

The article has argued that the beliefs that monetary policy is powerful in the short run and that money is superneutral in the long run may not be mutually consistent. The basic problem with most theories that reconcile these beliefs is that they rely directly or indirectly on the assumption that economic decisionmakers are victims of money illusion. If money illusion is the reason monetary policy has real effects, however, then its short-run real effects will be small and policymakers will not be able to exploit them systematically to achieve their goals. This point has been demonstrated in seminal work by Lucas (1972).

In the years since the 1970s, academic macroeconomics has slowly but surely embraced the neoclassical methodology pioneered by Lucas, which employs dynamic general equilibrium models and assumes that decisionmakers have rational expectations. The results of Lucas's work and that of a number of other neoclassical economists has served to further strengthen the monetarist position concerning long-run superneutrality of money.

In recent years, empirical studies of the impact of monetary policy have concentrated on identifying its short-run effects. This focus has been motivated, at least in part, by the conviction that money is long-run superneutral. Many researchers seem to believe that there is overwhelming empirical evidence in favor of long-run superneutrality. In reality, however, the proposition that monetary policy does not have long-run real effects is far from unequivocally established: indeed, an exploration of the empirical literature on long-run superneutrality could easily be the subject of a separate article. For the purposes of this article, it may suffice to cite a remark by Robert King and Mark Watson, two prominent macroeconomists whose empirical research has produced evidence both for and against long-run superneutrality. King and Watson (1992) report that for the United States during the postwar period the data do not appear to be consistent with the hypothesis that, over the long run, money is superneutral or that nominal interest rates move one-for-one with inflation.

The fact that the empirical evidence on the long-run superneutrality of monetary policy is not as overwhelming as some analysts believe suggests that there may be a need to look at theories that explore potential sources of long-run real effects for monetary policy. As this article has explained, Lucas's path-breaking work was an attempt to conduct a rigorous analysis of the logical consequences of the monetarist assumption that the real effects of monetary policy result from monetary surprises—a fact that has led Tobin, a leading Keynesian, to refer to Lucas's model as the "Monetarist Mark II" model. Lucas did not attempt to argue that every reasonable combination of assumptions would produce superneutrality, and it is consequently a mistake—albeit a very common mistake, even in the academic community—to equate neoclassical economics with the proposition that money is superneutral.

To repeat, Lucas's renowned 1972 paper employed innovative methodology to explore the implications of a very particular set of assumptions. The methodology is logically separate from the assumptions and can be used to analyze the consequences of very different assumptions. In fact, it is possible that monetary policy influences real economic activity for reasons completely different from the ones Lucas identified. In a recent interview in *New Yorker* magazine, Lucas acknowledges that the real effects of monetary policy may not result from unexpected policy changes. He comments, "Monetary shocks just aren't that important. . . . There's no question, that's a retreat in my views" (Cassidy 1996, 55).

Abandoning the assumption that policy surprises are the main reason monetary policy can have real effects leaves two options. One is to accept the view of the real business cycle theorists that Federal Reserve policy actions are essentially irrelevant. A second option is to attempt to identify alternative channels through which the monetary authority could affect real economic activity. This article has reviewed a small part of the recent academic literature that explores the second option. The results reported in this literature indicate that monetary policy may be a great deal more powerful than most academic economists believe. They also suggest that the possibility that monetary policy has substantial long-run real effects deserves more attention from economists and policymakers.

REFERENCES

BALL, LAURENCE, AND N. GREGORY MANKIW. 1995. "Relative-Price Changes as Aggregate Supply Shocks." *Quarterly Journal of Economics* (February): 161–93.

BARRO, ROBERT J. 1995. "Inflation and Economic Growth." Bank of England *Quarterly Bulletin* (May): 166–76.

BENCIVENGA, VALERIE R., AND BRUCE D. SMITH. 1991. "Financial Intermediation and Endogenous Growth." *Review of Economic Studies* 58:195–209.

BERNANKE, BEN S., AND MARK GERTLER. 1995. "Inside the Black Box: The Credit Channel of Monetary Policy Transmission." NBER Working Paper No. 5146.

BHATTACHARYA, JOYDEEP, MARK G. GUZMAN, ELISABETH HUYBENS, AND BRUCE D. SMITH. 1997. "Monetary, Fiscal, and Reserve Requirement Policy in a Simple Monetary Growth Model." *International Economic Review* 38:321–50.

BRUNO, MICHAEL, AND WILLIAM EASTERLY. 1998. "Inflation Crises and Long-Run Growth." *Journal of Monetary Economics,* forthcoming.

BULLARD, JAMES, AND JOHN KEATING. 1995. "The Long-Run Relationship between Inflation and Output in Postwar Economies." *Journal of Monetary Economics* 36:477–96.

BULLARD, JAMES, AND STEVEN RUSSELL. 1998a. "How Costly Is Sustained Low Inflation for the U.S. Economy?" Federal Reserve Bank of St. Louis and IUPUI Working Paper.

———. 1998b. "An Empirically Plausible Model of Low Real Interest Rates and Unbacked Government Debt." *Journal of Monetary Economics,* forthcoming.

CASSIDY, JOHN. 1996. "The Decline of Economics." *The New Yorker,* December 2, 50–60.

CHRISTIANO, LAWRENCE, AND MARTIN S. EICHENBAUM. 1991. "Liquidity Effects, Monetary Policy, and the Business Cycle." Federal Reserve Bank of Minneapolis and Northwestern University Working Paper.

———. 1992. "Liquidity Effects and the Monetary Transmission Mechanism." *American Economic Review* 82:346–53.

DANTHINE, JEAN-PIERRE, JOHN B. DONALDSON, AND LANCE SMITH. 1987. "On the Superneutrality of Money in a Stochastic Dynamic Macroeconomic Model." *Journal of Monetary Economics* 20:475–99.

DEGREGORIO, JOSE. 1992. "The Effects of Inflation on Economic Growth." *European Economic Review* 36:417–24.

DOW, JAMES P., JR. 1995. "The Demand and Liquidity Effects of Monetary Shocks." *Journal of Monetary Economics* 36 (December): 91–115.

ESPINOSA-VEGA, MARCO, AND STEVEN RUSSELL. 1997. "History and Theory of the NAIRU: A Critical Review." Federal Reserve Bank of Atlanta *Economic Review* 82 (Second Quarter): 4–25.

———. 1998a. "Can Higher Inflation Reduce Real Interest Rates in the Long Run?" *Canadian Journal of Economics* 31:92–103.

———. 1998b. "The Long-Run Real Effects of Monetary Policy: Keynesian Predictions from a Neoclassical Model." Federal Reserve Bank of Atlanta Working Paper 98-6, April.

ESPINOSA-VEGA, MARCO, AND CHONG YIP. 1998. "Fiscal and Monetary Interactions in an Endogenous Growth Model with Financial Intermediaries." *International Economic Review,* forthcoming.

FISHER, IRVING. 1926. *The Purchasing Power of Money: Its Determination and Relation to Credit Interest and Crises.* New York: Macmillan Company.

FRIEDMAN, MILTON. 1956. "The Quantity Theory of Money—A Restatement." In *Studies in the Quantity Theory of Money,* edited by Milton Friedman. Chicago: University of Chicago Press.

———. 1968. "The Role of Monetary Policy." *American Economic Review* 68:1–17.

FUERST, TIMOTHY J. 1992. "Liquidity, Loanable Funds, and Real Activity." *Journal of Monetary Economics* 29:3–24.

GREENSPAN, ALAN. 1997. "Monetary Policy Testimony and Report to the Congress." Humphrey-Hawkins Testimony. Available on-line at <http://www.bog.frb.fed.us/boarddocs/HH/9707Test.htm> [August 12, 1998].

HASLAG, JOSEPH. 1998. "Monetary Policy, Banking and Growth." *Economic Inquiry* 36:489–500.

HOFFMAN, DENNIS L., AND ROBERT H. RAASCHE. 1991. "Long-Run Income and Interest Elasticities of Money Demand in the United States." *Review of Economics and Statistics* 73:665–74.

JOHNSON, HARRY G. 1975. *On Economics and Society.* Chicago: University of Chicago Press.

KEYNES, JOHN M. 1964. *The General Theory of Employment, Interest, and Money.* San Diego: Harcourt Brace and Company. Original edition, 1953.

KING, ROBERT, AND MARK W. WATSON. 1992. "Testing Long Run Neutrality." National Bureau of Economic Research Working Paper No. 4156.

KYDLAND, FINN, AND EDWARD PRESCOTT. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica* 50 (November): 1345–70.

LUCAS, ROBERT E. 1972. "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4:103–24.

———. 1981. "Tobin and Monetarism: A Review Article." *Journal of Economic Literature* 19:558–67.

———. 1990. "Liquidity and Interest Rates." *Journal of Economic Theory* 50 (April): 237–64.

Metzler, Lloyd A. 1951. "Wealth, Saving, and the Rate of Interest." *Journal of Political Economy* 59 (April): 93–116.

Meyer, Laurence H. 1997. "Monetary Policy Objectives and Strategy." *Business Economics* 32 (January): 17–20.

Miller, Preston J., and Thomas J. Sargent. 1984. "A Reply to Darby." Federal Reserve Bank of Minneapolis *Quarterly Review* (Fall): 21–26.

Nelson, Charles R., and Charles I. Plosser. 1982. "Trends and Random Walks in Macroeconomic Time Series." *Journal of Monetary Economics* 10:139–62.

Phelps, Edmund. 1967. "Phillips Curves, Expectations of Inflation, and Optimal Unemployment over Time." *Economica* 34 (August): 254–81.

Phillips, A.W. 1958. "The Relationship between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957." *Economica* 25 (November): 283–99.

Sargent, Thomas J. 1996. "Expectations and the Nonneutrality of Lucas." *Journal of Monetary Economics* 37:535–48.

Sargent, Thomas J., and Bruce Smith. 1987. "Irrelevance of Open Market Operations in Some Economies with Government Currency Being Dominated in Rate of Return." *American Economic Review* 77 (March): 78–92.

Sargent, Thomas J., and Neil Wallace. 1976. "Rational Expectations and the Theory of Economic Policy." *Journal of Monetary Economics* 2:169–83.

———. 1981. "Some Unpleasant Monetarist Arithmetic." Federal Reserve Bank of Minneapolis *Quarterly Review* (Fall): 1–17.

Schreft, Stacey L., and Bruce D. Smith. 1997. "Money, Banking, and Capital Formation." *Journal of Economic Theory* 73:157–82.

Stiglitz, Joseph E. 1991. "Alternative Approaches to Macroeconomics: Methodological Issues and the New Keynesian Economics." NBER Working Paper No. 3580.

Taylor, John B. 1980. "Aggregate Dynamics and Staggered Contracts." *Journal of Political Economy* 88, no.1:1–23.

———. 1994. "The Inflation/Output Variability Trade-off Revisited." In *Goals, Guidelines, and Constraints Facing Monetary Policymakers.* Conference Series No. 38. Boston: Federal Reserve Bank of Boston.

Tobin, James. 1965. "Money and Economic Growth." *Econometrica* 33:671–84.

———. 1980. "Stabilization Policy Ten Years After." *Brookings Papers on Economic Activity* no. 1:19–71.

Wallace, Neil. 1984. "Some of the Choices for Monetary Policy." Federal Reserve Bank of Minneapolis *Quarterly Review* (Winter): 15–24.

Wynne, Mark A. 1995. "Sticky Prices: What Is the Evidence?" Federal Reserve Bank of Dallas *Economic Review* (First Quarter): 1–12.

Credit Union Issues

ARUNA SRINIVASAN AND B. FRANK KING

Srinivasan is an assistant vice president in the Atlanta Fed's credit and risk management department. King is the associate director of research at the Atlanta Fed. *They thank Larry Wall, Gerald Dwyer, and Scott Frame for comments on previous versions of this article.*

CREDIT UNIONS AND THEIR LEGAL STATUS HAVE MADE THE FINANCIAL NEWS MORE THAN USUAL THIS YEAR. IN FEBRUARY THE U.S. SUPREME COURT PARTIALLY SETTLED A LONG-RUNNING CONTROVERSY ABOUT THE CONCEPT AND EXTENT OF COMMON BOND LIMITS ON THESE INSTITUTIONS' MEMBERSHIP. THE COURT INTERPRETED THE FEDERAL CREDIT UNION ACT AS LIMITING MEMBERSHIP IN A FEDERAL CREDIT UNION TO INDIVIDUALS SHARING A SINGLE COMMON BOND.

The ensuing debate about limits of credit union membership has extended, quite naturally, to credit union tax status (see, for example, Bickley 1997; McConnell 1998; Robinson 1998). Meanwhile, the U.S. House of Representatives and Senate have overwhelmingly passed and the President has signed a bill that would substantially annul the Supreme Court decision; grandfather past common bonds, membership, and membership eligibility; and establish principles for regulation and determining safety and soundness while leaving credit unions' favorable tax status intact and limiting their business lending (Anason and McConnell 1998; Anason 1998a, b).

The controversy swirling around credit unions is often depicted as a simple fight between a group of small mutual institutions with limited membership, limited (primarily consumer) product powers, and tax exemption and a group of generally larger, generally stockholder-owned institutions that are not tax exempt (for example, see McConnell 1998; National Association of Federal Credit Unions 1997; Robinson 1998; and Schaefer 1997). However, the issues and implications of solutions for the conflict are not so simple. Changes in credit union organization and taxation are likely to affect credit unions, their customers, and their competitors in several ways. These include impacts on ease of access to credit union services by consumers; credit unions' costs, risks, and methods of corporate decision making; their competitive position relative to other financial institutions; and the extent of operations allowed for tax-subsidized entities in providing consumer financial services.[1]

This article attempts to provide a basis for thinking about current credit union issues. It begins with a brief outline of credit unions' current place among American depository financial institutions. In order to explain the development of credit unions' special legal status around the beginning of this century, it outlines the origins of these features as attempts to solve a set of problems that plagued most depository financial institutions of the time. The problems included limited information about individual borrowers who could provide no security and costly procedures for collecting unsecured debt. The article describes how classic credit union characteristics—mutuality and common bond structure—developed to attack these listed problems and how more recent developments are generating pressures to relax common bond limits. The discussion considers the spillover of common bond issues into a debate on tax exemption for credit unions.[2] In conclusion, the article turns to some of the likely impacts of changes in credit unions' legal structure.

Credit Unions' Place in the Current American Financial System

As Table 1 shows, credit unions have played an increasingly important role in consumer banking over the last thirty-five years. From 1960 to 1985 their share of the consumer credit market almost doubled, increasing to 12.4 percent. Since 1985 they have lost less than a 1 percent share, while commercial banks were losing a 5 percent share.

As they do with commercial banks and thrifts, both state and federal governments charter credit unions. A federal insurance fund, the National Credit Union Share Insurance Fund (NCUSIF), insures individuals' shares of a majority of state and all federal institutions to a $100,000 per shareholder limit. The remaining state credit unions secure share insurance from various state and private funds. There were 6,957 federal and 4,396 state credit unions at the end of 1997. State laws govern state-chartered credit unions' common bond limits and powers, which vary from state to state; however, state credit unions and their regulators have been subject to legal attacks on common bond restrictions similar to those recently waged on federal credit unions.[3]

Table 2 shows that about 30 percent of all Americans are members of credit unions. At year-end 1997, there were more credit unions—over 11,000—than any other type of depository financial institution. Although that number represents a significant decline during the past decade, the number of Americans doing business at credit unions rose during that period.

Among depository financial institutions, credit unions are generally the smallest. Their median share value (equivalent to banks' total domestic deposits) totaled $5.1 million as of the end of 1997. This amount contrasts with a median total domestic deposits figure of $57.4 million for commercial banks.

Most credit unions offer a simple set of deposit and loan products to consumers. There are, however, some larger, more complex credit unions. For instance, the largest twenty have a median share value of $1.4 billion; they account for 12.5 percent of total credit union share accounts. The next 100 largest have a median share value of about $480 million; they account for 17.4 percent of total share accounts. It is these credit unions that have drawn much of their competitors' fire and received the most legislative attention in recent debates on common bond and taxation (see McConnell 1998, for example).

To the extent that they accept deposits (called shares) and make loans, credit unions resemble other depository institutions such as commercial banks. However, credit unions have several distinguishing legal characteristics. Each credit union member (holder of credit union shares) has one vote in selecting its board members and making other management and organization decisions. This voting structure (one member–one vote) differs from other mutual financial institutions such as mutual savings banks. The latter allocate voting rights in proportion to the size of a member's deposit. Credit unions derive their net worth by accumulating retained earnings. They do not issue capital stock. Most credit unions rely on unpaid, volunteer boards of directors elected by, and drawn from, each institution's membership, with the board setting policies for the credit union. In smaller credit unions most of the staff is composed of member-volunteers as well. Some credit unions also receive subsidies in the form of office space or member time from their sponsors or employers (GAO 1991).

In this country credit unions have, until recently, had rather strict limitations on their membership, generally based on an affinity or "common bond" among members.

Credit unions are not-for-profit institutions. They return earnings to their members as reduced fees, reduced interest rates on loans, or as higher "dividends on shares" (which is equivalent to interest on deposits). They may also reinvest the earnings in the credit union as "retained earnings." Important to recent debates on their status, credit unions are exempt from federal corporate income taxes.[4]

In this country credit unions have, until recently, had rather strict limitations on their membership, generally based on an affinity or "common bond" among members. For federal credit unions, this bond may be based on a common employer, association, or religious, social, or community organization.

Credit unions' competitors often assert that much of the growth in credit unions' share of consumer lending have been driven by the relaxation of membership requirements implemented in 1982 and later by federal credit unions' regulator, the National Credit Union Administration (NCUA), combined with continuing exemption from federal corporate income taxes.[5] In recent years, this concern has manifested itself primarily in litigation over the

1. *For more detailed discussions of credit unions and current policy issues related to them, see studies by the U.S. General Accounting Office (GAO) (1991) and the U.S. Department of the Treasury (1997).*
2. *Debate on the appropriate range of credit union products is also occurring. Typically, their limited ability to offer small commercial and farm loans is questioned (GAO 1991). However, few credit unions offer these loans, and they made up less than 1 percent of total assets of credit unions in the United States and its territories at the end of 1997.*
3. *For a summary of current state suits, see CUNA & Affiliates Legal Division (1998).*
4. *For a more detailed discussion of credit union characteristics, see GAO (1991) and Moysich (1990).*
5. *Specifically, NCUA reinterpreted section 109 of the Federal Credit Union Act to allow multiple common bonds in individual credit unions.*

TABLE 1 Composition of the U.S. Consumer Credit Market

	1960	1965	1970	1975	1980	1985	1990	1995
Banks and Bank Holding Companies	43.14	46.41	49.10	51.23	50.69	49.82	47.71	44.86
Thrifts	3.27	3.08	3.29	4.88	6.39	9.68	6.12	3.54
Credit Unions	6.37	7.49	9.73	12.41	12.41	12.44	11.29	11.65
Asset-Backed Securities Issuers	——	——	——	——	——	——	9.57	18.94
Finance Companies	26.31	25.36	24.03	19.94	22.19	22.26	17.04	13.48
Other[a]	20.92	17.66	13.85	11.54	8.33	5.80	8.27	7.52

Note: The market here includes the institutions reported by the source as holders of consumer debt. Figures are percentages.

[a] Includes nonfinancial corporate and nonfarm, noncorporate businesses

Source: Board of Governors of the Federal Reserve System, *Flow of Funds Accounts of the United States*, table L.222

TABLE 2 Characteristics of U.S. Credit Unions, 1987 and 1997

	December 31, 1987	December 31, 1997
Number of Credit Unions	15,049	11,353
Number of Members (millions)	53.2	72.1
Number of Potential Members (millions)	181.5	244.4
Median Share Value ($ million)	1.8	5.1
Median Share Value of Top 20 Firms ($ million)	529.7	1,444.4
Median Share Value of Next 100 Firms ($ million)	217.5	482.6
Median Share Value of Remaining Firms ($ million)	1.8	4.8
Market Share of Credit Unions among Depository Institutions (percentage)	4.97	8.37

Source: Board of Governors of the Federal Reserve System, private data base definition of the common bond and in proposals for eliminating consumers' access to credit unions' government subsidies from federal income tax exemption.

Origins of American Credit Unions and Their Special Features

Credit unions developed in response to a gap in the supply of consumer banking services in the United States at the turn of the century. At this time and into the post–World War II era, commercial banks concentrated primarily on providing services to businesses and affluent individuals or making secured loans to homebuyers and farmers. In the financial environment of the time, information upon which to base decisions on the creditworthiness of potential borrowers was difficult and costly to come by.

Historically, credit unions used the common bond requirement for membership to help determine the creditworthiness of individual borrowers and to provide peer pressure on borrowers to pay their debts. Credit unions based their lending decisions largely on the reputation of loan applicants in the relevant affinity group. Because credit union members were individually liable for the loans made to other members, strict membership criteria like the common bond helped limit the lending risks borne by members and encouraged their monitoring of borrowers. Founders of early credit unions often voluntarily imposed common bond restrictions to reduce default risk and reduce the costs of monitoring loans. This country's first credit union law, passed in Massachusetts in 1909, permitted organizers to specify in their charter "conditions of residence or occupation, which qualify for membership" (Moody and Fite 1971). Other states and eventually the federal government followed suit in making the common bond a key organizing principle for credit unions.

The Federal Credit Union Act in 1934 limited membership in a federal credit union to "groups having a common bond of occupation or association, or to groups within a well-defined neighborhood, community, or rural district" (GAO 1991). The act neither elaborated on this definition at the time nor stated the reason for the requirement. Some courts have inferred that the purpose of the 1934 common bond requirement was to facilitate safe and sound operations. Until this year Congress had not addressed this issue in subsequent amendments to the Federal Credit Union Act or other law.[6]

What Is Implied by Credit Unions' Differences from Banks?

Credit unions' special features are more than cosmetic. They result in important differences between credit unions and stockholder-owned financial institutions, like commercial banks, in goals, customer base, operations, and competitiveness.

The most fundamental difference between banks and credit unions lies in two aspects of ownership—common bond and mutuality. As discussed above, common bond restrictions promote members' knowledge of the creditworthiness of other members and allow exercise of moral suasion on debtors. Lack of these limits on commercial banks and other stock institutions allows broader ownership, but, arguably, it also makes credit analysis more costly.

The most fundamental difference between banks and credit unions lies in two aspects of ownership—common bond and mutuality.

Borrowers typically have more and better information about their own financial condition than anyone else does, including lenders. It is sometimes in their interest to withhold adverse information, knowing that revealing it to a lender could affect the amount and terms of lending. Lenders deal with this situation in a variety of ways. They obtain relevant information on potential borrowers from sources other than the borrowers, write contracts that provide protection against events about which they have little information, and monitor borrowers' financial condition to varying degrees. In the case of business loans, a requirement that the borrower maintain a deposit account provides a mechanism for monitoring on a continuing basis.

In the past, credit unions' common bond and mutuality organizational structure has addressed asymmetric information problems by requiring that these institutions lend only to members. The valuable information provided by records of size and pattern of balances in share accounts is often supplemented by the lending officer's personal knowledge of the borrower. In occupational credit unions, knowledge of an employer's condition can also be helpful. Prior to the general availability of on-line credit reports, the common bond and mutuality arrangement reduced costs of extending and monitoring credit to consumers whose financial statements and credit records had been difficult to acquire.

Credit unions' mutuality also has impacts on their corporate governance. The primary difference between stockholder and mutually owned institutions lies in who controls the firms and receives the earnings. A commercial bank's or stock thrift's stockholders vote for the firm's managers, distribute its profits, and are free to

6. The history of American credit unions is discussed more fully in Moody and Fite (1971).

sell their privileges. A mutual association, on the other hand, is owned by its depositors (called shareholders in the case of credit unions). In credit unions and mutual thrifts, each depositor has the right to vote for the managers of the firm.

Owners of a firm often employ managers to actively do that firm's business. These managers are their agents. What economists call agency problems refer to the difficulty that owners have in making sure that their agents—that is, managers—work in the owners' best interest. Managers, who may or may not also be stockholders, often have better information about the firm and different motivations from those of stockholders, who are often more widely dispersed geographically. This agency problem can result in improperly managed and inefficient operations with high management compensation.

A feature of credit unions' mutuality is the diversity of interests among their members/owners.

Manager/owner problems exist both in commercial banks and credit unions. Approaches to their solution are influenced by organizational structure. In mutuals like credit unions, officers and directors are often unpaid and, therefore, cannot inflate their salaries. In some credit unions other perks (such as office space) are constrained by sponsor contributions. These conditions go a long way toward mitigating the results of conflicts between owners (members) and managers. As members themselves, directors and unpaid officers have an incentive to monitor paid managers in the interest of all members. In stock firms like banks, creating incentives to resolve potential agency conflicts, including stock options for managers, can lead to higher costs. The threat of takeovers and stockholder and director revolts may also act as a check on wasteful expenditures.[7]

Some evidence indicates that credit unions' one depositor–one vote structure allows them to adapt successfully to change. A feature of credit unions' mutuality is the diversity of interests among their members/owners. Conflict amoung member groups can affect the manner in which a mutual, not-for-profit credit union is operated since the credit union cannot simultaneously maximize the dividend rate for savers and minimize loan rates for borrowers.

Some evidence on the results of member conflicts in mutual organizations comes from a study of German cooperative banks by Emmons and Mueller. Like credit unions in the United States, cooperative banks in Germany are mutual institutions that have been steadily increasing their market share relative to other types of financial institutions.[8] Emmons and Mueller focus on the diversity of interests between members of cooperative banks and highlight the dual role of members as borrowers and lenders. They develop a model showing that "a shift in the median (hence pivotal) member of the cooperative from predominantly a borrower orientation to a lender orientation causes the cooperative bank to shift its policy from underpricing credit towards the provision of competitively priced credit and deposit services" (1997, abstract). This result depends on cooperative financial institutions' one member–one vote, organization. Emmons and Mueller conclude that the democratic nature of cooperatives' ownership in fact creates opportunities for adaptation and survival. Together with a nationwide supporting infrastructure to capture scale and scope economies, the organization of German cooperative banks has allowed them to compete successfully with other, stockholder-held banking groups.

While the Emmons and Mueller model has not been directly tested on U.S. credit union data, some of the similarities between the structure and performance of U.S. credit unions and German cooperatives suggests that both gain ability to adapt from their one member–one vote characteristic.

Pressures on the Common Bond

In the past three decades, various changes in the environment in which financial institutions, particularly credit unions, operate have caused reconsideration of and changes in credit unions' common bond requirements. Developments in information technology have diluted the effect of restricting membership to a tight community. Increasing complexity and size have pushed more credit unions to seek professional managers. Extension of deposit insurance to credit union shares has lessened both member need and incentive credit monitoring. Dealing with credit union financial problems has made broader common bonds quite practical for their insurers and regulators.

Technological Change. Common bond requirements have become less important for the analysis of credit risks with the development of credit reporting services and other advances in collecting, transmitting, and analyzing credit information that have made it less costly to assess the likelihood of default on a particular loan. Both credit unions' competitors and credit unions themselves have adopted newer information technologies and greatly expanded the variety and availability of loans, both secured and unsecured.

Many credit unions—to meet customer demand and to compete with other depository institutions—also offer technology-based services such as ATMs and computer and electronic banking to take advantage of electronic account and transaction processing. The technology needed to provide such services involves substantial fixed costs. Adding more membership groups makes such investments more economical by allowing a credit union to spread its fixed costs over more members. A study of the productive efficiency of credit unions by Fried, Lovell, and Vanden Eeckaut (1993) concluded that credit unions can improve their performance by increasing their total membership as well as by increasing the number of accounts per member.

Credit-analysis advances provided by new technologies have not, however, eased another risk feature of the tight common bond. The more that a credit union's membership shares a common bond of employment or otherwise has similar exposure to plant closings or other economic risks, the less diversified is its exposure to credit risk. Diversifying the membership base makes the credit union more resilient in the face of problems experienced by any one local employer. This diversification can be accomplished by multiple common bonds.

Managerial Factors. Managerial factors may also create incentives for credit unions to grow by adding new membership groups. A credit union board of directors seeking to attract high-quality, professional managers may find it easier to do so if the credit union is large or has growth opportunities. Moreover, as nonprofit cooperatives, credit unions do not generally compensate their managers on the basis of profit or stock performance. Instead, management compensation often reflects a credit union's size and product offerings. Managers may therefore have an incentive to increase the credit union's size. Adding new membership groups is an obvious method of doing so (GAO 1991).

Share Insurance. Discipline to control risk taking by mutual depository institutions can be provided by creditors, depositors, owners, and managers. In the case of credit unions, if bankruptcy occurs creditors other than depositors are generally fully protected. Creditors have this protection because their position in the liquidation of a failed credit union is senior to that of depositors, whose shares are judged to represent equity, not debt. An important feature of the traditional common bond between members of a credit union was the willingness of some members to put their personal savings at risk by letting the credit union lend these funds to other members. This relationship between borrowers and savers originated to engender a higher sense of obligation than borrowers might otherwise feel toward ordinary creditors (GAO 1991).

The monitoring relationship between savers and borrowers has no doubt diminished since the introduction of share insurance in 1970. Credit union members still own their institutions. Since 1970, however, to the extent that their share accounts fall under $100,000, they are insured owners. This insurance dilutes the impact of the common bond in inducing shareholders to monitor borrowers and management.

Further dilution of the risk-management impact of the common bond may also have come from the increase in credit union size with the expansion of the common bond in the 1980s. Credit union membership has increased. The average credit union had more than 6,000 members as of year-end 1997.

Diversifying the membership base makes the credit union more resilient in the face of problems experienced by any one local employer.

Financial Difficulties. In the early 1980s, the technical and organizational pressures on the common bond, discussed above, combined with a practical need to reduce economic distortions associated with credit union financial troubles. These factors induced significant easing of common bond restrictions. In 1982, faced with major difficulties in the industry, the NCUA reinterpreted the National Credit Union Act to substantially ease its common bond policy. Through this change, credit unions were allowed to have more than one common bond group in the same organization. The NCUA adopted and later expanded this policy to allow merging of credit unions that had financial problems, to provide a diversity of membership that would help credit unions weather economic downturns, and to make credit unions' services more widely available (Burger and Dacin 1991; Murphy 1996).

This relaxation has also enabled credit unions to grow larger (Good 1996; Murphy 1996; Smale 1997). As of June 1996 more than half of the 7,244 federal credit unions had multiple group fields of membership. These credit unions had a total membership of 32.6 million and accounted for approximately 80 percent of total federal credit union shares (Smale 1997).[9]

Other Factors. Several other factors have encouraged credit unions to add new membership groups: Downsizing or closings at manufacturing firms, military bases, and other large employers have shrunk the membership base of many occupational credit unions. Worker mobility has made the membership base less stable than in the past, when many members had a long-standing relationship with their employers. In addition, restricting

7. See Jensen and Meckling (1976) and Fama and Jensen (1983) for a more complete discussion of agency problems.

8. Unlike U.S. credit unions, cooperative banks in Germany do not enjoy tax advantages.

9. Comparable information on state-chartered credit unions is not available.

credit unions to a single common bond has made credit union services unavailable to many segments of the population. Finally, since the minimum viable size of a credit union has been generally understood to be around 500, employees of small companies have faced barriers to forming successful credit unions (Evans and Shull 1998).

The Tax Exemption Issue

As credit unions have grown larger and developed more diverse membership, their long-standing exemption from federal income taxes has drawn more fire from competitors. One should not be surprised that tax exemption has become an issue that is attached to common bond extension.

Credit unions' special legal and tax status is often related to their role in providing an alternative source of financial services for less affluent individuals with few alternatives.

Credit unions' exemption from federal income tax dates back to the Revenue Act of 1916, which provided tax-exempt status to mutual thrift institutions and cooperatives. Because they were found to be "organized and operated for mutual purposes and without profit," the U.S. Attorney General ruled in 1917 that credit unions, which were all state-chartered then, were entitled to the exemption. According to Moody and Fite (1971) this ruling was relatively noncontroversial at the time. The first federal credit unions were chartered in 1934 and granted tax-exempt status in 1935 under a ruling by the Internal Revenue Service.

Since 1937 Congress has reconsidered the tax-exempt status of mutual financial institutions on several occasions. In 1951 it repealed the tax exemption for all mutual institutions except credit unions. This decision was based on the view that credit unions (unlike other mutual financial institutions) had remained true to their original purpose of providing cooperative financial services to members. Mutual savings banks, on the other hand, were deemed in a 1951 report by the Senate Finance Committee to be "in active competition with commercial banks . . . for the public savings, and . . . with many types of taxable institutions in the security and real estate markets" (cited in Burger and Lypny 1991, 16).

Competitor financial institutions as well as legislators attempting to balance the federal budget have challenged the tax-exempt status of credit unions since at least 1970. Commercial banks and thrifts have claimed that the easing of common bond limits and the expanded products and services that credit unions have been allowed to provide their members have eroded the distinction between banks and credit unions. This argument closely parallels arguments made a half-century ago, when Congress removed tax exemption from mutual savings and loan associations and savings banks (Moody and Fite 1971).

The credit union industry has evolved over the last sixty years in an environment that treats credit unions as nonprofit cooperatives. Without tax exemption the industry would probably have evolved differently. It is likely that credit unions would not have grown as fast, and some credit unions might not have formed. Moreover, credit union customers would have received some combination of lower deposit rates and higher borrowing rates.

Credit Unions' Public Purposes

Credit unions' special legal and tax status is often related to their role in providing an alternative source of financial services for less affluent individuals with few alternatives. Evidence on the current economic status of credit union membership and on available alternative sources of loans and deposit services may dilute the strength of these public purpose arguments.

The purpose of the Federal Credit Union Act as set forth in 1934 was "to make more credit available to people of small means" (GAO 1991). None of the common bond criteria in that law, however, address the economic status of members or potential members. While there are no statistically reliable data on the economic status of credit union members earlier in the century, it was accepted that members were generally not affluent (Moody and Fite 1971).

Expansions of the common bond requirement in the 1980s may have contributed to changes in membership characteristics. The little publicly available data on membership characteristics suggest that members are not all "of small means" but may still not be as well off as commercial banks' individual customers. Two recent published surveys give information. A Gallup Organization poll reported in the *American Banker* found that the average annual family income of credit union members was lower than the income of bank customers (Seiberg 1997). The average income for credit union members was also slightly below the 1996 level for the entire population. An earlier survey by the Secura Group for the American Bankers Association suggests that the typical credit union member is in his or her "early 40s, employed, with above-average income, better educated than a non-member and with access to financial services from a variety of sources" (reported in GAO 1991). This evidence is consistent with the results of an earlier survey in 1987 by CUNA & Affiliates, a credit union trade group (reported in GAO 1991).

Another major public purpose argument in the development of credit union laws hinged on the existence of few borrowing alternatives for consumers

(Moody and Fite 1971). While a paucity of alternatives may have characterized the beginning of this century and even the early postwar period, consumers now have a rather broad set of alternatives for credit for most purposes. Several types of suppliers exist for each type of consumer credit, and they make their services available in many markets.

Challenges in Court

Relaxing previously limited common bond restrictions has brought credit unions into more direct competition with other depository institutions such as banks. These institutions, banks in particular, have argued that credit unions' less restrictive common bond makes credit unions very similar to taxed financial institutions. They conclude that tax exemption amounts to a federal subsidy to credit unions and their members and gives credit unions unfair competitive advantages (Fettig 1996; Marshall 1996).

Banks have gone to court to limit the scope of credit unions whose charters define particularly large fields of membership on the grounds that potential members do not share the requisite common bond. Their suits have been filed in both federal and state courts. Until a recent U.S. Supreme Court decision, the federal suits had met with mixed results.[10] In the 1980s two courts found that the common bond contributed to the sound management of a credit union and thus to the safety and soundness of the industry as a whole. Yet neither court found much legislative guidance on limitations of the common bond. One of the courts inferred from a state statute that the common bond requirement had been imposed to promote the institution's financial stability. In a later case, a federal court dismissed the banking industry's challenge to a proposed charter for a multiple-bond credit union on the grounds that Congress had "purposefully sacrificed the competitive interest of banks" in favor of making credit more readily available to people of small means through the chartering of credit unions (GAO 1991).

Earlier this year, the U.S. Supreme Court ruled on a pivotal case involving the AT&T Family Credit Union (Supreme Court 1998). The institution had expanded from its original core group of employees of Western Electric Company in three North Carolina cities to 112,000 members in fifty states and more than 150 separate employer groups. The lower courts disagreed on the interpretation of the common bond language in the original statute. The district court ruled that the statutory language was ambiguous and deferred to the NCUA's interpretation of the law. However, the appeals court found the actual language of the statute to be clear in defining credit unions to include a single group with a common bond. Any subsequent groups wanting to join the credit union would have to share a common bond with the original group. According to the appeals court, Congress had used the common bond mechanism to "ensure both that those making the lending decisions would know more about applicants and that borrowers would be more reluctant to default. . . . [and, thereby to unite] credit union members in a cooperative venture" (U.S. Court of Appeals 1996).

The Supreme Court ruled that the NCUA's interpretation of the common bond language of section 109 of the Federal Credit Union Act was illegal. The case returned to lower court for a decision on whether common bond requirements should revert to their status as defined in 1982 or continue at their current status and, if so, whether credit unions with multiple common bonds should be allowed membership expansion in their existing groups. These questions became moot when President Clinton signed the legislation recently passed by the House and Senate.

Relaxing previously limited common bond restrictions has brought credit unions into more direct competition with other depository institutions such as banks.

This law maintains the concept of common bond but allows combining of groups with different common bonds in a single credit union. It does not change credit unions' tax exemption, but it limits credit unions' commercial loans of more than $50,000. The NCUA must still issue regulations based on the new law.

Conclusion

Recent and future actions on credit unions' common bond limits and federal tax status may well have implications for the efficiency, risk, and competitiveness of these institutions—and their competitor financial institutions. Clearly, credit union customers would be affected.

Allowing past multiple common bonds to stand and leaving open the way for others has positive implications for credit unions and their customers but negative implications for their competitors, their competitors' customers, and taxpayers. Individual credit unions and the industry will be better able to expand and to offer customers more products, taking advantage of scale economies, diversification, and tax exemption. Their growth

10. Currently thirteen states have suits on common bond in process (see CUNA & Affiliates Legal Division 1998).

and market share expansion will probably be greater. Countering these positive effects might be some small overall diminution in credit unions' ability to gather credit information and collect debts. This loss will be particularly true for small credit unions. Credit unions' gains will come at the expense of competitor financial institutions. Their individual customers would have the choice of moving to credit unions, and some likely would. Taxpayers, considered as a separate group in the abstract, would pay more subsidy for provision of consumer financial services.

If easing common bond restrictions allows larger credit unions and the industry grows, and if some credit unions approach their business loan limits, one might expect the movement to remove credit unions' tax- exempt status to become more active and credible. Credit unions would appear more like other financial institutions, such as mutual thrifts, that are taxed. Issues of whether they still primarily serve people of small means and whether they are one of a limited set of consumer financial alternatives are likely to receive a great deal of attention.

REFERENCES

ANASON, DEAN. 1998a. "Senate Passes Credit Union Bill; Big Loss for Banks." *American Banker,* July 29.

———. 1998b. "The Major Provisions of Controversial New Law." *American Banker,* August 10, available by subscription on-line at <http://www.americanbanker.com>.

ANASON, DEAN, AND BILL MCCONNELL. 1998. "Banking Panel Backs Broad Credit Union Membership." *American Banker,* March 27, available by subscription on-line at <http://www.americanbanker.com>.

BICKLEY, JAMES M. 1997. *Should Credit Unions Be Taxed?* Congressional Research Service Report No. 97-548E. May.

BURGER, ALBERT E., AND TINA DACIN. 1991. *Field of Membership: An Evolving Concept.* Madison, Wisc.: Filene Research Institute.

BURGER, ALBERT E., AND GREGORY M. LYPNY. 1991. *Taxation of Credit Unions.* Madison, Wisc.: Filene Research Institute.

CUNA & AFFILIATES LEGAL DIVISION. 1998. "Field of Membership Litigation Summary." Available on-line at <http://www.cuna.org/data/spec_reports/litsum.html> [June 22, 1998].

EMMONS, WILLIAM R., AND WILLI MUELLER. 1997. "Conflict of Interest between Borrowers and Lenders in Credit Cooperatives: The Case of German Cooperative Banks." Federal Reserve Bank of St. Louis Working Paper No. 97-009A, March.

EVANS, DAVID S., AND BERNARD SHULL. 1998. *Economic Role of Credit Unions in Consumer Banking Markets.* Paper prepared for Boeing Employees' Credit Union. January.

FAMA, EUGENE F., AND MICHAEL C. JENSEN. 1983. "Separation of Ownership and Control." *Journal of Law and Economics* 26, no. 2:301–25.

FETTIG, DAVID. 1996. "Banks Turn Up the Heat against Credit Unions." Federal Reserve Bank of Minneapolis *Fed Gazette* 8 (July): 1, 3–4.

FRIED, HAROLD O., C.A. KNOX LOVELL, AND P. VANDEN EECKAUT. 1993. "Evaluating the Performance of U.S. Credit Unions." *Journal of Banking and Finance* 17:251–65.

GOOD, BARBARA A. 1996. "The Credit Union Industry: An Overview." Federal Reserve Bank of Cleveland *Economic Commentary,* May 15.

JENSEN, MICHAEL C., AND WILLIAM H. MECKLING. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure." *Journal of Financial Economics* 3, no. 4:305–60.

MARSHALL, JEFFREY. 1996. "Credit Unions: True Threat or Mere Nuisance?" *U.S. Banker* 106 (November): 53–59.

MCCONNELL, WILLIAM T. 1998. "Viewpoints: Larger Credit Unions Ruining the Movement." *American Banker,* March 13, available by subscription on-line at <http://www.americanbanker.com>.

MOODY, J. CARROLL, AND GILBERT C. FITE. 1971. *The Credit Union Movement: Origins and Development, 1850–1970.* Lincoln, Nebr.: University of Nebraska Press.

MOYSICH, ALANE K. 1990. "An Overview of the U.S. Credit Union Industry." *FDIC Banking Review* 3 (Fall): 12–26.

MURPHY, MAUREEN M. 1996. *Multiple Group Credit Unions: Litigation and Potential Legislative Responses.* Congressional Research Service Report No. 96-997. December.

NATIONAL ASSOCIATION OF FEDERAL CREDIT UNIONS. 1997. "Why Congress Must Support the 'Credit Union Membership Access Act.'" Available on-line at <http://www.nafcunet.org/cuc/k4.htm> [June 22, 1998].

ROBINSON, KEN. 1998. "Viewpoints: Compromise by Banks Is Bait-and-Switch Tactic." *American Banker,* March 13, available by subscription on-line at <http://www.americanbanker.com>.

SCHAEFER, MARCUS. 1997. "Comment: Banks Could End Up Losing If They Win Credit Union War." *American Banker,* November 20, available by subscription on-line at <http://www.americanbanker.com>.

Seiberg, Jaret. 1997. “The Bank-Credit Union Battle May Hinge on Service to Poor.” *American Banker,* February 7, available by subscription on-line at <http://www.americanbanker.com>.

Smale, Pauline. 1997. *Multiple-Group Federal Credit Unions: An Update.* Congressional Research Service Report No. 97-267E. May.

Supreme Court of the United States. 1998. *National Credit Union Administration v. First National Bank & Trust Co. et al.* Decided February 25, 1998. Available on-line at <http://supct.law.cornell.edu/supt/html/96-843.ZS.html> [July 30, 1998].

U.S. Court of Appeals for the District of Columbia Circuit. 1996. *First National Bank and Trust Company et al. v. National Credit Union Administration and AT&T Family Federal Credit Union and Credit Union National Association.* Decided July 30, 1996. Available on-line at <http://laws.findlaw.com/DC/945295a.html> [July 24, 1998].

U.S. Department of the Treasury. 1997. *Credit Unions.* Report to Congress. December.

U.S. General Accounting Office. 1991. *Credit Unions: Reforms for Ensuring Future Soundness.* Report to Congress. July.

The Federal Government's Budget Surplus: Cause for Celebration?

GERALD P. DWYER JR.
AND R.W. HAFER

Dwyer is vice president in charge of the financial section of the Atlanta Fed's research department. Hafer is a professor in the Department of Economics, Southern Illinois University at Edwardsville, and a visiting scholar at the Atlanta Fed. *They thank Lucy Ackert, Frank King, Larry Wall, and Madeline Zavodny for helpful comments.*

PROJECTED SURPLUSES IN THE FEDERAL GOVERNMENT'S BUDGET HAVE GENERATED FANFARE SOMETIMES VERGING ON EUPHORIA. BECAUSE THE FEDERAL GOVERNMENT LAST HAD A SURPLUS IN 1969, A PROJECTED SURPLUS FOR FISCAL YEAR 1998 AND LATER YEARS IS BEING VIEWED AS SOMETHING OF A MILESTONE. UNLIKE POLICIES OF THE LAST THREE DECADES THAT HAVE SOUGHT TO LOWER THE DEFICIT, POLICY OPTIONS NOW MAY INCLUDE WAYS TO USE THE SURPLUS.

The budget surplus and projections of surpluses for a number of years have brought forth a variety of opinions and suggestions about what to do with them. Some have called for lowering taxes. Others have suggested that the federal government could now engage in greater spending. And some have called for retiring government debt. Herbert Stein, a former chairman of the president's Council of Economic Advisers, recently suggested that "We had an agreed-upon answer for what to do about deficits: Reduce them. . . . Now no one knows what the surplus constraint is. We are at sea" (1998).

Some of this rhetoric might be interpreted as hyperbole. Still policy discussions often seem to have focused on how to reduce the deficit to the exclusion of all else. This single-minded approach to the budget reflects the argument that federal government deficits absorb saving. As Benjamin Friedman put it, the claim is that deficits in the 1980s "consumed most of what individuals and businesses . . . saved during this period" (1988, 167). In this view deficits are bad because they reduce national saving, decrease funds available for net investment, and therefore retard economic growth. While the evidence has not been kind to this argument (Seater 1993), this reasoning suggests that a balanced budget or, even better, a surplus should be the goal of the federal government's fiscal policy.

The size of the deficit by itself does not provide much information about the federal government's activities. Federal government spending and taxation are more informative. Suppose that two economies both have balanced budgets. Conventional wisdom about deficits might suggest that the impact of the two governments on their economies is similar: both budgets are balanced. In one economy, though, government spending and taxes might be 90 percent of gross domestic product (GDP); in the other economy, government spending and taxes might equal 10 percent of GDP. The impact of the governments on the two economies is likely to be quite different even though both have balanced budgets.

The purpose of this article is to amplify on the importance of considering spending and taxes when analyzing the federal government's budget. The first section looks at the behavior of federal government budgets in the past, with special emphasis on trends in spending and taxation. The article then considers in more detail the prospects for future surpluses. A key element in the recent projections of surpluses is the role of trust funds, especially the Social Security trust fund, in federal budget accounting.

Perspective on the Surplus

Loosely speaking, when spending exceeds tax receipts, there is a deficit; when spending is less than tax receipts, there is a surplus. Although *deficit* is a more convenient term at times and *surplus* is more convenient at other times, the terms are mirror images. A negative deficit is a surplus and a negative surplus is a deficit.

Deficits characterize the federal budget for most of the last fifty years: the federal government's spending has generally exceeded its receipts. Chart 1 shows the federal government's unified budget surplus for fiscal years since 1950. The *unified budget* consolidates the spending and revenues of all federal government agencies and trust funds into an overall budget to reflect the government's transactions with the rest of the economy (Office of Management and Budget [OMB] 1998a, 323).[1] The fiscal year ends on June 30 through 1976 and on September 30 since then. Chart 1 reflects the change in fiscal year by not connecting the values for 1976 and 1977.[2] The surpluses in the chart are deflated by the GDP chain price index to put the figures in terms of 1992 dollars.[3] Many discussions of the deficit rely on current dollar measures of the deficit, a practice that is quite misleading for comparing deficits across time when there is substantial inflation. For instance, the federal government deficit was about $53.2 billion for fiscal year 1975 and about $107.4 billion for 1996, values that indicate a roughly doubled deficit. The level of prices as measured by the GDP chain price index, however, was 2.6 times higher in 1996 than in 1975. Hence, the larger deficit in terms of current dollars is really smaller in terms of the inflation-adjusted amount.

Changes in the economy, such as recessions and expansions, are one major reason the deficit changes, as Chart 1 shows. The shaded bars indicating recessions show how recessions are related to changes in the surplus. The surplus tends to decrease during recessions for two reasons. First, federal government tax receipts decrease during recessions, largely because income and related tax receipts fall. Second, recessions trigger automatic increases in federal government spending; for example, payments for unemployment compensation increase during recessions because more individuals are unemployed.

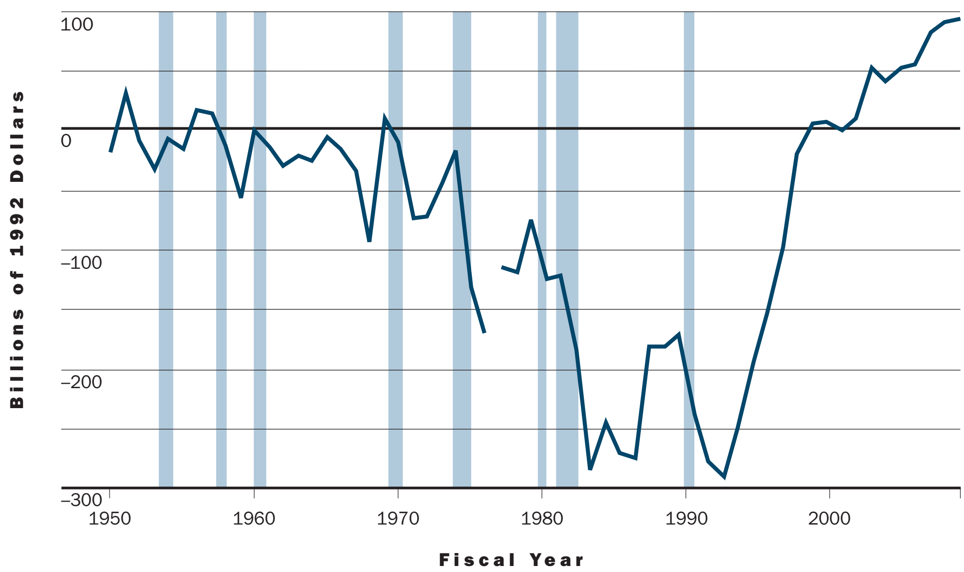
Chart 1 also includes projections of the budget surplus made by the Congressional Budget Office (CBO). The projections made in March 1998 indicate that federal government receipts will exceed outlays throughout the next decade. The CBO projects that by 2008 the budget surplus will reach $95 billion in 1992 dollars. As the chart makes clear, a decade of continued budget surpluses would be extraordinary compared with the past fifty years. Another ten years without a recession also would be extraordinary, which is one reason for being uncertain about this projection. Ten years of expansion would be longer than the previous record expansion of almost eight years from November 1982 to July 1990, and the implied expansion would be significantly longer than this, from April 1991 through September 2008. Recessions are hard to predict, however, and no evidence in March 1998 (or as of this writing) indicates much likelihood of one in the predictable future. Overall, based on current laws, this projection of a decade of budget surpluses is based on the best available information.

Does the current federal government surplus signal the onset of a new age in government fiscal policy? The answer to this question is not so obvious.

What are the trends in government spending and taxes that have produced the past deficits and projections of surpluses? Chart 2 shows federal government spending and revenue in 1992 dollars. The current dollar values are deflated by the GDP price index to make the dollar amounts more comparable across time. In addition, the vertical axis has a proportional rather than a linear scale. On this proportional scale any given distance on the vertical axis represents the same proportional dollar amount rather than the same dollar amount, as would be the case with a linear scale. As a result, the slope of the line connecting any two points in the chart is the growth rate. The chart indicates that the projected surpluses

1. *Evans (1997) provides a good introduction to the terminology and concepts of the budget.*
2. *The transition quarter in 1976 is excluded.*
3. *Additional adjustments, not necessary for this article, would improve the deficit as a measure of the change in indebtedness of the federal government (Dwyer 1982; Eisner 1986, chap. 2; Kotlikoff 1992; Penner 1982; Webb 1991).*

CHART 1 The Federal Government Unified Budget Surplus



Note: Shaded areas indicate recessions.
Sources: OMB (1998b, 23–24, table 1.3); CBO (1998a, table 2); recession dates from the National Bureau of Economic Research

result from a higher projected growth rate for revenue than for spending.

It would be easy to place too much reliance on the projected budget surpluses in Charts 1 and 2. As already mentioned, the projections rely on the absence of a recession. In addition, federal government laws affecting spending are not likely to stay the same. Even without these possibilities, ten years is a long time in terms of achieving any reliability in budget projections; the Congressional Budget Office calls the figures projections rather than forecasts to make the tentative nature of the numbers clear (CBO 1998a, chap. 1). Budget projections are subject to large changes. In January 1997 the CBO's forecast of the 1997 fiscal-year deficit was $100 billion too large (CBO 1998a, chap. 2). Fiscal year 1998 illustrates the difficulties again. In February 1998 the CBO forecast that the federal government would run a budget deficit of $5 billion in the fiscal year ending September 1998 (CBO 1998a, chap. 2). By May this deficit projection became a projected surplus of $43 billion to $63 billion, a change of $48 billion to $68 billion in the deficit projection for the fiscal year in progress (CBO 1998c). By July the projection was that the surplus would be "near the upper end of this range" (CBO 1998d).

While a ten-year period is long for reliable forecasts, it can be short for evaluating the long-run state of the budget. The projection of surpluses until 2008 shown in Charts 1 and 2 masks important concerns over longer periods. Most importantly, the trust funds for Social Security are projected to have a surplus of about $93 billion in fiscal year 1998 but are projected to begin running deficits by 2006 to 2018 (Social Security Administration 1998, 25). Partly because of the Social Security trust funds, the CBO also projects that the unified budget will swing from surpluses to persistent, increasing deficits by 2020 (CBO 1998b).

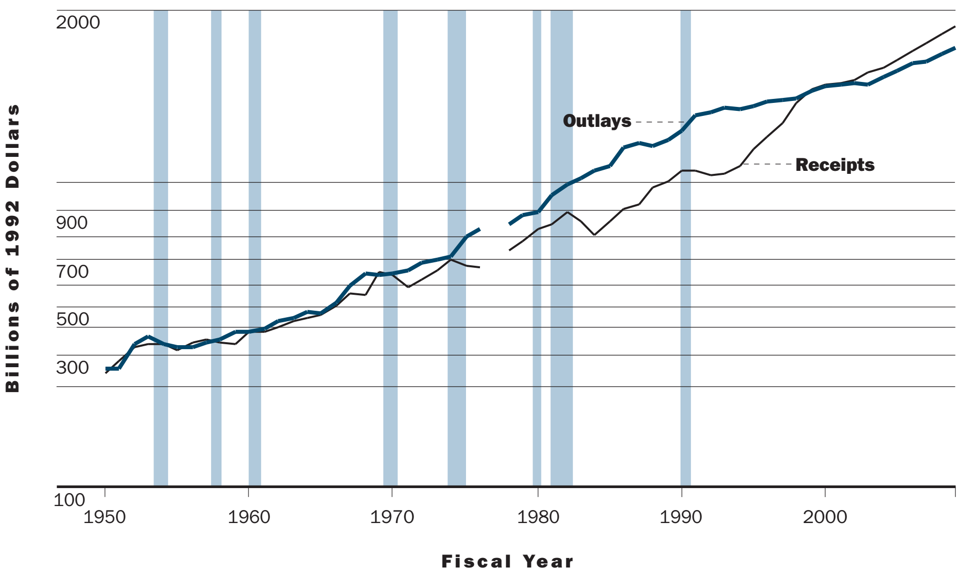
The next section of the article examines trends in federal government spending and revenues and taxes and discusses the implications of current policies for future spending and tax policies.

Federal Government Spending and Receipts

Spending. Federal government *spending* (also called *outlays*) is the sum of the amounts spent on goods and services and transfer payments plus net interest paid on outstanding federal debt. These parts of spending can have quite different effects on the economy.

Government Purchases. Government purchases are purchases by the government of newly produced goods and services; they represent withdrawals of resources from the economy that are used by the government for its activities rather than by private individuals for private purposes. Government purchases cover a wide range of goods, from airplanes and computers to pencils. They also include earnings received by government employees. When the government buys a computer, the value of the good is reflected in its price. Government employees' earnings also reflect the value of resources withdrawn from the economy if the employees' earnings are equal to their opportunity cost and that opportunity cost equals the value of goods or services the employees would have produced in private employment. Thus, government em-

CHART 2 Federal Government Spending and Taxes



Note: Shaded areas indicate recessions.
Sources: OMB (1998b, 23–24, table 1.3); CBO (1998a, table 2)

ployees' earnings is a reasonable measure of the value of forgone private goods and services.[4] Based on this supposition, government purchases equal the value of payments made by the government for goods and services currently produced. The government uses the goods and services purchased as well as government employees' services to provide services such as defense, education, and police protection that can provide benefits to the economy.[5]

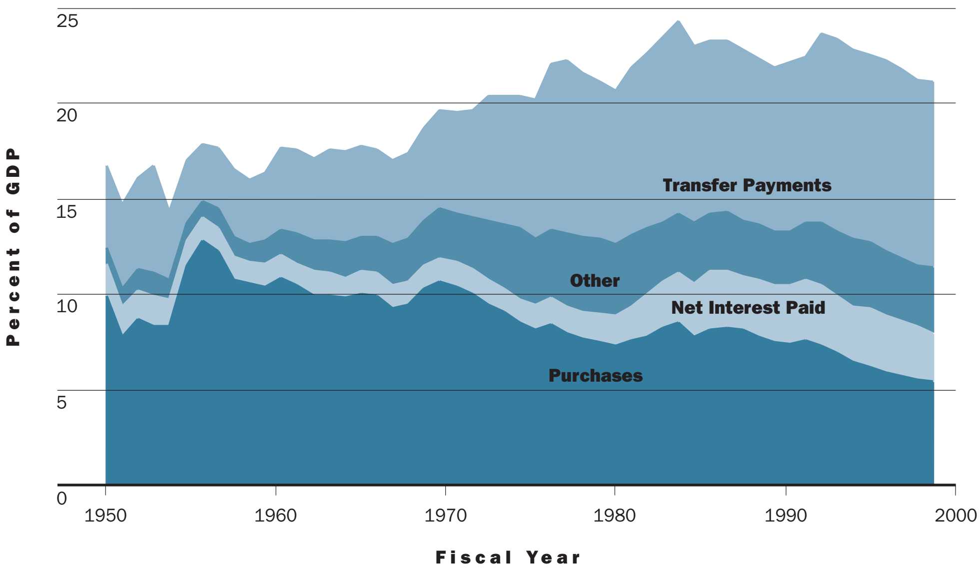
Transfer Payments. Transfer payments, unlike purchases, are not payments for goods currently produced or services currently rendered. Instead, transfer payments redirect income from one person to another via the government. Transfer payments include payments made to recipients in programs such as Aid to Families with Dependent Children, Medicaid, unemployment insurance, and Social Security. Perhaps less obviously, transfer payments also include pension payments to retired government employees, both civilian and military. These payments are, after all, for services rendered in the past, not services currently provided.

Net Interest Payments. Net interest paid is interest paid by the federal government less interest received. Net interest payments by the government, like transfer payments, also are not for goods and services that could have been used directly to produce other goods and services during the current period. Interest payments are made to the holders of government securities—those who lend funds by purchasing these securities. Unlike many other categories of spending, net interest payments are not under the federal government's immediate control, at least short of default. Net interest payments are a function of conditions that are largely determined by the federal government's past actions: the size of the outstanding debt issued in earlier years and the interest rate on that debt when issued.

Interest payments on outstanding federal government debt are sometimes viewed as a pernicious result of government deficits. Interest payments finance past spending that was financed by issuing debt rather than raising taxes. The payment of interest to service the current debt has led some observers (for example, Stein 1998) to suggest that current and prospective surpluses be used to retire outstanding federal government debt. From one point of view, it is always preferable that interest payments be lower. Current taxpayers would prefer to pay lower taxes, other factors being equal, and they could pay lower taxes if it were possible to have lower interest payments and keep everything else the same. In general, though, it is not possible to have everything else the same. For example, if the government had spent less in the past

4. *This supposition is not always a good one. When there was a military draft in the United States, the wages paid were not sufficient to induce people to join the military, so people were drafted.*
5. *Purchases are not the only way the federal government affects the allocation of resources in the economy. Legal requirements and regulations do not appear in the budget but also affect the economy.*

CHART 3 Federal Government Spending as a Percentage of GDP



Source: OMB (1998b, 262–64, table 14.1)

on highways, there would be fewer highways today and transportation costs would be higher. More generally, interest payments can be the result of past spending that provides current benefits, an arrangement that has the potential to make everyone better off. Interest on government debt is the price paid for postponing payment, just as for private interest payments, and interest payments similarly can reflect optimal or improvident behavior.

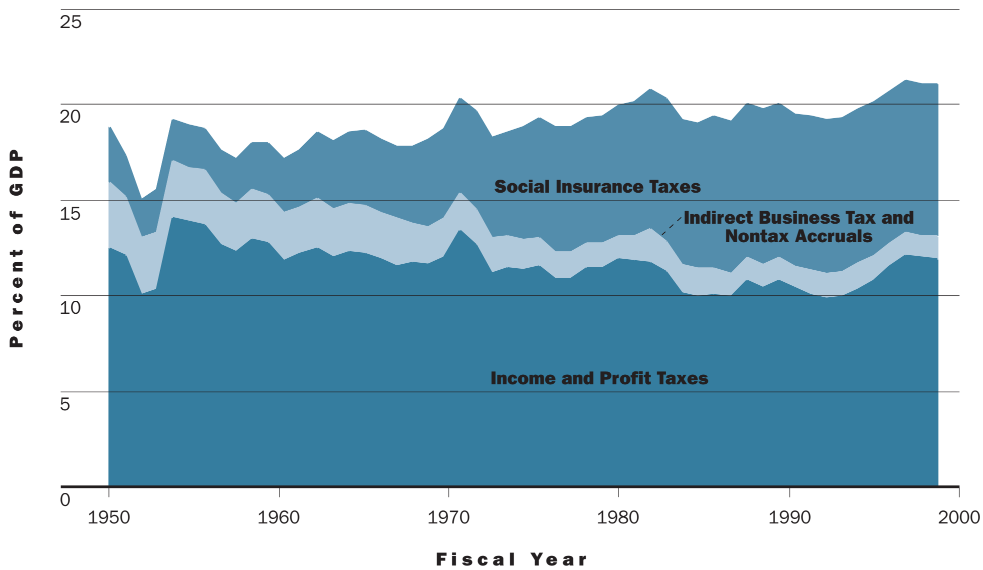
Chart 3 shows the trend of federal government spending relative to GDP since 1950. The chart breaks spending into four broad components: purchases, transfer payments, net interest payments, and other. The residual category denoted *other* includes grants to state and local governments and subsidies less surpluses of federal government enterprises.[6] The chart indicates a number of important developments during the past several decades. The most obvious development is the increase in federal government spending relative to GDP. In 1950, total spending was about 18 percent of GDP; in 1997 it was 22 percent of GDP.[7] Another development is a decline in the ratio of spending to GDP since the early 1980s. In 1983 spending was over 23 percent of GDP, higher than in any fiscal year since 1950.

Chart 3 also shows that the distribution of spending has changed over time. During the 1950s transfer payments accounted for a relatively small proportion of spending. For example, in 1951, transfer payments were 3.5 percent of GDP. Transfer payments increased in importance in the 1960s and 1970s, and by 1997 transfer payments accounted for about 9.9 percent of GDP. This growth of transfer payments understates the change in domestic transfer payments. Foreign aid, or transfers to foreigners, was 1 percentage point of GDP in 1951 and 0.2 percentage points in 1997. Domestic transfer payments almost quadrupled during that period, from 2.5 percent of GDP in 1951 to 9.7 percent in 1997. This shift in spending from purchases to domestic transfer payments indicates that the government is now withdrawing fewer resources from the economy than in the past and is redistributing more income.

Receipts. Federal government *receipts* primarily are taxes, which fund most of the government's spending. Descriptions of the sources of taxes often characterize them according to who writes the check for the tax. Common lists include individual income taxes, corporate income taxes, excise taxes, and social insurance payroll taxes. While such a division may be useful for some purposes, it is misleading for determining where the final tax burden lies. First, corporations do not pay taxes; people, whether they are shareholders, employees, or customers, do. Second, whether shareholders, employees, or customers ultimately bear the burden of a tax depends on the tax's effects on the prices paid and received and incomes of shareholders, employees, and customers. The burden of the tax is not necessarily, or even in general, borne by whoever has the legal liability for writing a check to the government.

Chart 4 shows total taxes as a percentage of GDP since 1950.[8] The federal government's tax receipts increase substantially, from 14 percent of GDP in 1950 to 20 percent in 1997. Chart 4 also indicates the composition of receipts. Income and profit taxes are the largest portion—about 66 percent of total tax receipts in 1950 and about 57 percent in 1997.

CHART 4 Federal Government Taxes as a Percentage of GDP



Source: OMB (1998b, 262–64, table 14.1)

Relative to GDP, social insurance taxes are noticeably higher in 1997 than in 1950. Sometimes called "contributions" instead of taxes because the individuals making the payments become entitled to certain benefits, these taxes generally are mandatory, not optional. In 1950 social insurance taxes were about 2 percent of GDP and 14 percent of federal government receipts; in 1997 they were about 8 percent of GDP and 39 percent of federal government receipts. The increase in this component is largely due to increases in receipts from Social Security and Medicare taxes.

This growth in Social Security tax receipts has come about primarily through increases in underlying tax rates. Measuring tax receipts relative to GDP removes the effect of general increases in income. The Social Security taxes paid are determined partly by the tax rate and the income subject to the tax, generally labor income (or earnings). In addition, there is a maximum level of earnings subject to the tax for Old Age, Survivors, and Disability Insurance (OASDI). There is no limit on earnings subject to the tax for Medicare (Hospitalization Insurance, or HI). The first payment in 1937 of Social Security taxes was quite small compared with payments in later years. When introduced, the total Social Security tax rate was 2 percent of earnings up to \$3,000. In 1998 the total Social Security tax rate includes an OASDI tax rate of 12.4 percent payable on earnings up to \$68,400 and an HI tax rate of 2.9 percent on all earnings.

Chart 5 shows the combined tax rate for OASDI and HI and the maximum income subject to the OASDI tax from 1937 through 1997. The tax rate includes both employees' and employers' contributions. Other than the legal distribution of tax liability, the only difference between the employers' and employees' share of the tax is whether the tax is included in employees' gross pay. If all Social Security taxes were paid by either employers or employees, the before-tax wage paid by employers, the after-tax wage received by employees, and the level of employment would be the same.[9]

The adjustment of earnings for inflation is important. From the inception of Social Security until 1951, the maximum amount of earnings subject to tax was \$3,000, which is less than one-twentieth of the maximum earnings of \$68,400 subject to OASDI tax in 1998. When adjusted for changes in the level of prices, however, the difference in taxable earnings is not even the same order of magnitude. The consumer price index was about 11.1

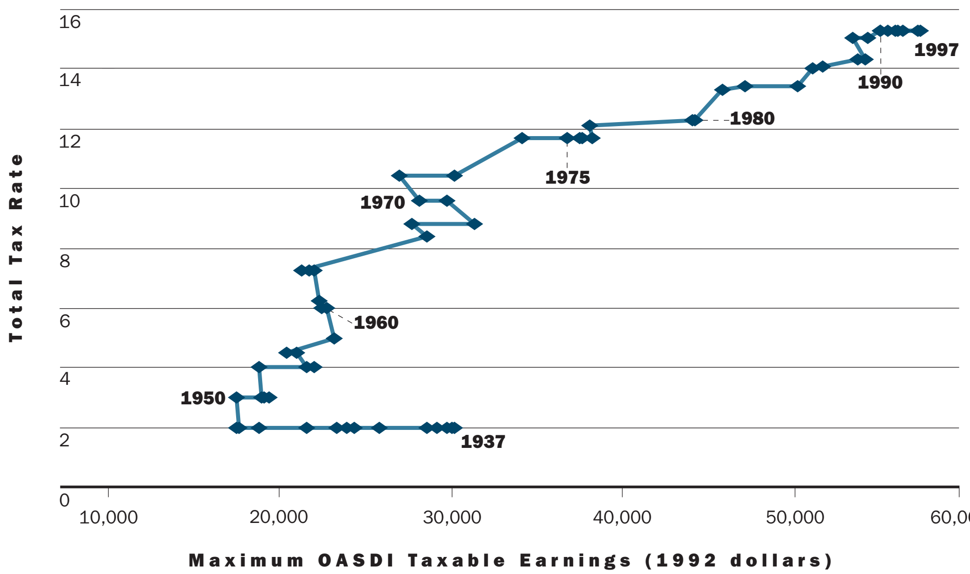
6. This category also includes wage disbursements less accruals, an item that was 0.1 percent of GDP in 1950 and 1996.

7. These figures include only the federal government; including all governments—federal, state, and local—magnifies the size of the increase.

8. These numbers do not include revenues from providing goods and services, such as selling electricity or admitting people into national parks. If state and local government receipts from taxes were included, government receipts would be 21 percent of GDP in 1950 and 30 percent in 1997 (OMB 1998a, 270, table 15.3).

9. This observation assumes that other tax rates would be adjusted to reflect the change in income subject to those taxes.

CHART 5 Social Security and Taxable Earnings



Source: Social Security Administration (1998, table II.B1)

times higher in 1997 than in 1937. Hence, the $3,000 subject to OASDI tax in 1937 would buy about the same amount as $33,600 in 1998. Still, with the maximum amount of earnings subject to Social Security tax in 1998 at over $68,000, the amount adjusted for inflation is twice what it was in 1937, not twenty times larger.

Chart 5 shows that the tax rate for Social Security has risen substantially since the program's inception. The rate was 15.3 percent in 1997, quite a bit higher than the 2 percent rate in 1937.

Such tax increases affect economic decisions and behavior. Consider a change in the Social Security tax rate on earnings. If other factors remain the same, an increase in the marginal tax rate on earnings lowers the private marginal return from activities that generate labor income, resulting in fewer hours worked and less output produced in the economy.

Social Security taxes are not unique in affecting private behavior. Virtually all, if not all, taxes and transfer payments change relative prices and consequently the decisions people make. Income taxes lower the marginal return from generating income. Transfer payments generally have implicit tax rates. Qualifying for a transfer payment generally depends on a person's income and assets. A higher income often reduces the size of the transfer payment, with the loss offsetting part of the increase in income, but the after-transfer change in income is less than the gross change in income.

Perhaps obviously, it does not follow that, because tax rates change people's behavior, there should be no taxes. Tax effects such as reductions in the quantity of labor supplied represent an *excess burden* of taxes—a cost of government actions that should be considered when making choices about the government's activities.

Fiscal Policy Now and in the Future

Is there a federal government surplus that somehow must be distributed? In one sense, the answer is yes. At a broad level, the government's unified budget is nothing more than a cash-flow identity that can be written

$$surplus = tax - pur - tr - int$$

or

$$deficit = pur + tr + int - tax,$$

where *tax* is government tax receipts, *pur* is government purchases, *tr* is transfer payments, and *int* is net interest payments. The deficit approximately equals the amount of additional debt issued by the Treasury. During 1998, it has become apparent that the federal government has an unexpected surplus, at least partly because of unexpectedly higher tax receipts. Something must happen with these higher receipts: they cannot disappear into a void. Either taxes will be decreased; government spending, interest payments, or transfer payments will be increased; or the federal government will reduce its outstanding debt.

Does the current federal government surplus signal the onset of a new age in government fiscal policy, as Herbert Stein suggested? The answer to this question is not just a matter of arithmetic and is not so obvious.

There are developments in the budget that throw cold water on euphoria about a surplus, whether due to lower federal government spending or higher taxes. In particular, the Social Security trust fund (OASDI plus HI)

TABLE 1 Projected Unified Budget Surpluses and Trust Fund Surpluses

	Surplus in Fiscal Year (billions of dollars)						
	Actual	Projected					
	1997	1998	1999	2000	2001	2002	2003
Unified Budget	–22	8	9	1	13	67	53
Trust Funds	126	149	171	173	177	201	202

Note: The unified budget projections are the CBO's projections issued in March 1998. The projections for the trust funds are from the budget proposed for fiscal year 1999 in January 1998 by the OMB (1998a) and therefore reflect any pertinent budget proposals.

Sources: Unified budget surplus projections from CBO (1998e, table 1); trust fund surpluses from OMB (1998a, 323).

has a projected surplus of about $93 billion in fiscal year 1998, which is greater than any estimate of the surplus in the unified budget. If Social Security were removed from the unified budget and no other changes were made, the federal government's budget would have a deficit on the order of $25 billion to $45 billion for fiscal year 1998. While small by recent standards, this deficit has rather different connotations than a budget surplus.

The federal government's unified budget includes current federal government activities and the receipts and expenditures of various dedicated funds established by legislation. Social Security's trust funds are the most prominent, but there are others that are nearly as large. Table 1 shows the overall surplus for the trust funds for fiscal years 1997 to 2003 along with corresponding estimates of the federal government's unified surplus. Overall, the trust funds are running surpluses and, consequently, accumulating nonmarketable federal government debt.[10] If the trust funds were not accumulating federal government debt and other taxes and spending were the same, deficits instead of surpluses would be projected for the federal government for these years.

The current Social Security surplus is not accidental: it is an expected result of changes in taxes in 1983 to cover future Social Security spending. The Social Security trust funds are accumulating nonmarketable Treasury securities in anticipation of increases in spending. When the projected increases in Social Security spending occur, the trust fund will exchange the securities with the Treasury for funds to pay for the increases. If federal government spending other than interest payments and taxes is unaffected, the Treasury then will issue debt to the public to acquire the funds to finance the higher spending. In sum, the current budget surplus reflects taxes paid now to finance expected increases in future spending.

Why might it be desirable to have the Social Security trust funds run surpluses now? A common justification is to tax people now who will receive benefits in the future. The fact that Social Security is designed partly to redistribute income limits the force of this argument.

An arguably more important reason to run surpluses is to smooth tax rates over time. If Social Security spending now and in the future remains unchanged, the only issue is when—not whether—taxes will be paid to finance the spending. Lowering taxes today implies raising taxes in the future, and, conversely, raising taxes today implies lowering taxes in the future. In short, pay now or pay later, but pay you will.

If Social Security spending were unchanged and Social Security tax rates were lower now, tax rates would have to be increased in the future from their current level. Lower tax rates now would lower the excess burden of the tax today at the cost of an increase in the excess burden in the future. Higher tax rates would increase the excess burden of the tax more in the future than it would decrease the burden today if, as is likely, the excess burden increases more at higher tax rates. Hence, higher tax rates today can be interpreted at least partly as an attempt to smooth the excess tax burden.

The CBO's projections of ten years of unified budget surpluses actually mask longer-term difficulties concerning the federal government's budget and Social Security. Even though Social Security is projected to have a surplus of about $93 billion in fiscal year 1998, Social Security

10. Nonmarketable securities are issued by the federal government but never sold to the public, to outside agencies, or in secondary markets. In effect, these securities represent funds borrowed and lent between different parts of the federal government. The Treasury borrows from trust funds when the latter run surpluses and uses these funds to replace borrowing from the public. Evans (1997, chap. 7) and the OMB (1998a, sec. 17) provide more detailed analyses.

spending is projected to exceed receipts by about 2012, and the trust fund is projected to be depleted by about 2032 (Social Security Administration 1998, 25). Obviously, projected depletion in precisely 2032 is subject to substantial uncertainty: any projection of what might happen decades from now based on current policies is fraught with peril.[11] Ten years of surpluses do not resolve these future problems (Social Security Administration 1998; CBO 1998b). As a recent CBO analysis states, "fundamental long-term budgetary problems will remain. Eventually the federal debt and deficit will start to rise as a result of pressures on the budget from Social Security, Medicare, Medicaid, and other programs that serve the elderly" (CBO 1998b, 1). Without a policy change, the CBO projects that the federal government's unified deficit would increase to 10 percent of GDP by the year 2040 (CBO 1998b, chap. 2). This situation hardly resembles financial well-being, and, indeed, changes in Social Security—whether cutting benefits, raising taxes, or privatizing Social Security—are quite likely.

Conclusion

Whether or not the projected federal government budget surplus for fiscal year 1998 is desirable, it is not a panacea. The projected budget surpluses in 1998 and succeeding years are based on projections of slower growth in federal government spending than in receipts. These surpluses have generated calls to decrease taxes, increase expenditures, or retire federal government debt. Actually, the surpluses can be interpreted as largely reflecting taxes paid now to finance expected increases in spending in the future, in particular on Social Security.

More generally, a budget surplus or deficit is not an adequate summary of how federal government spending and taxes affect the economy. A surplus or deficit is a result of choices concerning spending and taxation, choices that have substantial implications for the allocation of resources in the economy. Any analysis of fiscal policy that neglects spending, taxes, and tax rates is woefully deficient.

11. As Cogan (1998) points out, even assuming that the surplus will result in a reserve is inconsistent with the federal government's past behavior.

REFERENCES

Cogan, John F. 1998. "Social Security Surpluses Never Last." *Wall Street Journal,* July 31, sec. A.

Congressional Budget Office. 1998a. *The Economic and Budget Outlook: Fiscal Year 1999–2008.* Washington, D.C.: Government Printing Office. January.

———. 1998b. *Long-Term Budgetary Pressures and Policy Options.* Washington, D.C.: Government Printing Office. May.

———. 1998c. *Monthly Budget Review: Fiscal Year 1998, May 6, 1998.* Available on-line at <www.cbo.gov/showdoc.cfm?index=469&sequence=0&from=1> [July 13, 1998].

———. 1998d. *Monthly Budget Review: Fiscal Year 1998, July 9, 1998.* Available on-line at <www.cbo.gov/showdoc.cfm?index=648&sequence=0&from=7> [July 13, 1998].

———. 1998e. *Revised Baseline Budget Projections for Fiscal Years 1999–2008, March 3, 1998.* Available on-line at <http://www.cbo.gov/showdoc.cfm?index=356&from=4&sequence=0> [July 13, 1998].

Dwyer, Gerald P., Jr. 1982. "Is Inflation a Consequence of Government Deficits?" Federal Reserve Bank of Atlanta *Economic Review* 67 (August): 25–32.

Eisner, Robert. 1986. *How Real Is the Federal Deficit?* New York: Free Press.

Evans, Gary R. 1997. *Red Ink: The Budget, Deficit, and Debt of the U.S. Government.* San Diego: Academic Press.

Friedman, Benjamin M. 1988. *The Day of Reckoning: The Consequences of American Economic Policy under Reagan and After.* New York: Random House.

Kotlikoff, Laurence J. 1992. *Generational Accounting: Knowing Who Pays, and When, for What We Spend.* New York: Free Press.

Office of Management and Budget. 1998a. *Budget of the United States Government, Fiscal Year 1999, Analytical Perspectives.* Washington, D.C.: Government Printing Office.

———. 1998b. *Budget of the United States Government, Fiscal Year 1999, Historical Tables.* Washington, D.C.: Government Printing Office.

Penner, Rudolph G. 1982. "How Much Is Owed by the Federal Government?" In *Monetary Regimes and Protectionism,* edited by Karl Brunner and Allan H. Meltzer. Carnegie-Rochester Conference Series on Public Policy, vol. 16. Amsterdam: North-Holland Publishing Company.

Seater, John J. 1993. "Ricardian Equivalence." *Journal of Economic Literature* 31 (March): 142–90.

Social Security Administration. 1998. *1998 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds.* Washington, D.C.: Government Printing Office.

Stein, Herbert. 1998. "At Sea with Surpluses." *Wall Street Journal,* May 19, sec. A.

Webb, Roy H. 1991. "The Stealth Budget: Unfunded Liabilities of the Federal Government." Federal Reserve Bank of Richmond *Economic Review* 77 (May/June): 23–33.