Can the Fed be a Payment System Innovator?

Jeffrey M. Lacker and John A. Weinberg

We live in a time of rapid technological change, in which the arrival of new ways of conducting business has become a commonplace occurrence. One segment of the economy where these changes are having a particularly significant effect is the payment system, the web of banks and other institutions through which payments for goods and services are cleared and settled. New mechanisms such as smart cards and internet-based electronic money have captured the imagination of many payment system observers and participants. While earlier predictions of the death of paper money have proven premature, the unprecedented pace of technological advance in the last decade has given new hope to the prophets of the electronic age.

The Federal Reserve (the Fed) plays a prominent role in the payment system, both as a provider of payment services and as a regulator. The public interest in an economically efficient payment system has been at the core of Fed payment system policy since the Fed's founding in 1914. With new electronic payment mechanisms apparently within grasp, there has been renewed attention to the role of the Fed in the innovative process. A committee headed by Federal Reserve Board Vice Chair Alice Rivlin recently completed a study of the Fed's role in the payment system, which gave special attention to how active a role the Fed should play in guiding payment system innovation.[1]

Within the Federal Reserve System, electronic check presentment (ECP) is seen as a potentially promising step in the evolution toward electronic payments. With ECP, consumers and businesses continue to make payments with paper checks, but banks and clearinghouses that clear and settle payments use

Jeffrey M. Lacker is a vice president and economist in the Research Department. John A. Weinberg is a research officer and economist in the Research Department. This article originally appeared in this Bank's *1997 Annual Report*. The authors wish to thank Ed Green, Tom Humphrey, Ned Prescott, Marsha Shuler, John Walter, Marvin Goodfriend, Walter Varvel, and Al Broaddus for helpful comments and discussions. The views expressed do not necessarily reflect those of the Federal Reserve System. The authors remain solely responsible for the contents of this article.

[1] Federal Reserve System (1998).

electronic information "captured" from the checks shortly after they are first deposited in the banking system. (See Appendix.) While some ECP services are now available, many important aspects of full-scale implementation are still under discussion. The Fed's role in developing and promoting ECP is clearly aimed at the public interest objective of enhancing payment system efficiency. In what follows, we ask whether the Fed can be a payment system innovator while remaining loyal to its fundamental public interest objective. In particular, how can we ensure that the Fed's payment system leadership contributes to economic efficiency?

Our approach to this policy question is founded on the notion that the payment system is a communications industry. Such industries involve substantial common costs—costs that cannot be uniquely attributed to any one user. This cost characteristic has important implications for industry behavior. The critical issue in such industries is how common costs are allocated across users.

Markets for communication services (including payment services) tend to be heavily regulated and, in some instances, served by government-owned enterprises, such as the U.S. Postal Service. Concerns about "universal access" often motivate government intervention. Here, universal access is usually interpreted as a concern about the cost of services to a particular class of users: residential phone customers, rural postal patrons, or small and remote depository institutions. Access has been provided through price regulation, as in telecommunications, and by direct government provision, as in the U.S. Postal Service.

We show that government involvement in other communications industries offers lessons for the role of the Federal Reserve in the payment system. In both the telecommunications and postal services industries, legal barriers to competition historically have helped sustain the provision of universal access. Barriers to competition allow the shifting of common costs to be pushed to the point where some users are subsidized, in a sense that we will make precise later. Such subsidization is inconsistent with economic efficiency, and would be impossible without barriers to competition. We point out that the Federal Reserve Banks still benefit from some barriers to competition—privileged treatment under current check presentment regulations—that would allow them to subsidize should they choose to do so. Federal Reserve policy explicitly seeks to prevent subsidization, and there is no direct evidence that the Fed currently subsidizes any segment of the check collection market. At the same time, however, it is clear that available analytical methods for determining the absence of subsidies are imperfect.

Barriers to competition impede technological progress by distorting adoption choices. The contrasting experiences of the telecommunications and postal services industries illustrate this fundamental conflict. In telecommunications, the removal of barriers and a retreat from access have been accompanied by rapid technological innovation. The U.S. Postal Service has retained barriers to competition and in the view of some observers has been relatively slow to adopt innovations. Barriers provide the opportunity for cross-subsidies that distort the innovative process. Against this background, we argue that the Fed should act to preclude subsidization by removing remaining barriers to competition. As we emphasize, however, this step may require some retreat from the goal of providing universal access.

1. THE PAYMENT SYSTEM AS A COMMUNICATIONS INDUSTRY

In the U.S. economy, roughly 220 million market transactions are made without cash daily, with a total dollar value of $1.6 trillion.[2] These transactions all involve credit. The seller receives a financial instrument representing a claim on either the buyer or a third party. For example, a check is the liability of the check writer and his or her bank. A credit card sale results in a claim—a "sales slip"—that entitles the merchant to good funds. Similarly, a debit card transaction gives the merchant a claim to good funds.

The clearing and settling of credit instruments used as means of payment intrinsically requires communication. A vast web of bookkeeping systems records the assets and liabilities of various economic entities—bank accounts, loan balances, investment funds, and the like. Noncash payment instruments are fundamentally bookkeeping instructions to debit an account of the buyer and credit an account of the seller. Those instructions must be communicated to the relevant bookkeeping systems in order to carry out the necessary accounting entries.

The payment system bundles together communication and financial services. Arrangements governing the use of payment instruments specify the allocation of risks associated with payment failures. For example, the merchant accepting a check bears the risk that the check writer may fail to cover it, but the merchant does not bear the risk of a fraudulent credit card purchase. While these risk-sharing arrangements are an important feature of the evolution of the payment system, they do not make the fundamental function of payment arrangements inherently different from other communication services.

Every new development in communication technology brings with it a new possibility for sending payment instructions. Improvements in freight transportation increase the speed and reliability with which checks can be delivered to a buyer's bank. Improvements in computer and telecommunications technologies facilitate the sending of payment instructions in electronic form directly to and from banks. Optimism about the transition to electronic payment instruments is based on the assessment that the technologies

[2] Bank for International Settlements (1996).

underlying electronic communications systems are improving rapidly, while physical transportation technologies are improving only slowly at the present time.[3]

2. SOME NOTEWORTHY CHARACTERISTICS OF COMMUNICATIONS INDUSTRIES

Economists have long noted that communications industries share certain distinct characteristics that have, in turn, heavily influenced industry behavior. The most salient of these is the prevalence of common costs. The allocation of these costs among diverse users is fundamental to the operation of communications industries. Governments tend to intervene in such industries to allocate these common costs in such a way as to promote access.

In what follows we employ a few technical terms that are necessary for a clear understanding of the economics of communications industries. While these terms are defined as they are introduced, they also appear in a glossary at the end of the article.

Common Costs

Every communication benefits two parties, the sender and the receiver. How should the costs of a message (a phone call, a letter, an e-mail) be divided between the two beneficiaries? The answer is not entirely obvious. While providers of communication services often collect fees from the sender, services are provided jointly to both parties. The costs of providing these services cannot be uniquely attributed to either of the beneficiaries. We call such costs *common costs*. Common costs also extend beyond the level of the individual message. A large part of the infrastructure costs of a communication system, such as phone lines and information processing resources, are common to all users of the system.

The significance of common costs distinguishes communications from many other industries. For most other goods and services, a large part of the costs of an individual's consumption can be uniquely attributed to that individual. The time that a dentist, a barber, or a mechanic spends serving a customer is a cost of serving that customer exclusively. Costs that can be unambiguously associated with the provision of goods or services to a particular individual are *attributable costs*.

For some costs, specifying whether they are common or attributable is not so simple. Costs that arise from a single message from a sender to a receiver (a single phone call, for instance) are attributable to that pair of users but are common between them. Similarly, the transportation costs of a single shipment

[3] Federal Reserve System (1998).

of mail between two points are attributable to the group of people sending or receiving letters on that shipment but are common among the members of that group. In communications industries, there are very few costs that are attributable to individual users, but there are many costs that are attributable to specific groups of individuals. There are also substantial costs that are common to entire communication systems.

Common costs are often *fixed costs*; they do not vary with the amount of goods or services produced. An industry that has large fixed costs and relatively small variable costs will exhibit economies of scale (declining average costs) over some range of output levels. When there are costs that are common to the production of multiple goods, then production is said to exhibit economies of scope. Economies of scale and scope are important characteristics of many communications markets because many costs are common among all users of the network.

Another notable feature of communications markets is in the nature of demand for such services. The economic value to an individual of having access to a communication system depends on the individual's own demands for connection to others and on the extent of the network of individuals connected by the system. A consumer will be willing to pay more for a communication service that allows communication with a larger set of correspondents. This relationship between an individual's valuation of a communication service and the extent of the network is referred to by economists as a *network effect*. Note, however, that a network effect is a consequence of both the interdependence of demand for communication and the existence of common costs. The idea that an individual "belongs to a network" is only meaningful if there are common costs associated with linking people together.

The presence of common costs and network effects makes it difficult to unambiguously specify the cost of serving a particular individual or group. On the one hand, one can ask, "Given that services are already being provided to others, what would it cost to extend service to this particular group?" The answer to this question yields the *incremental cost* of serving a group of users.[4] This definition excludes costs that are common to the delivery of service to this group and to others. On the other hand, one might ask, "What would it cost to provide services to this group if no one else were being served?" The answer to this question yields the group's *stand-alone cost*, which includes all common costs. Clearly, when common costs are substantial, incremental cost is much smaller than stand-alone cost.

[4] It is important to distinguish between incremental and marginal costs. Marginal cost is the added cost of the last unit of a good produced. Incremental cost is all of the additional costs that arise from extending a particular set of services to a particular set of users. This may include costs that are fixed with regard to the quantity of services provided, such as the costs of connecting a group of users to an existing network.

When there are no network effects, a group's incremental cost is simply the attributable cost of extending service to that group. When there are network effects, the addition of the new group also has the effect of creating benefits for other users. These benefits work to reduce the net cost of adding a new group of users. Hence, we need a more general definition of incremental cost:

> The *incremental cost* of extending service to a new group of users is the cost of adding that group to the network, minus the benefits for others created by that group's participation.

Some commentators have interpreted governmental concern for universal access in communications industries as a necessary response to network effects. Many believe reducing prices to some users may enhance efficiency by compensating them for the network benefits they bring to other participants. If the total benefits of an added participant, both to himself and to others, however, are greater than the costs of adding that participant, then a privately operated network will have an incentive to compensate the added participant. This would be the case even in the absence of government intervention. Network effects do not, by themselves, induce market failures.[5]

Allocating Common Costs

In pricing a communication service, a provider must decide who should bear the common costs. There are many possibilities. One could recover all such costs from one small group of buyers, or try to spread the burden evenly among all buyers. We can evaluate alternative cost allocations according to two criteria. First, are they consistent with efficient use of the service? Second, could they arise under competitive market conditions?

While there are many dimensions to the efficiency of a communication services market, one essential consideration is that the allocation of costs must provide customers with the right incentives to participate in the network. An individual's participation is economically efficient if the resulting benefits exceed the additional costs incurred. If the prospective customer is charged less than incremental cost, his or her participation could be inefficient, creating benefits smaller than the costs incurred. Hence, a minimal requirement for a cost allocation to be consistent with efficient use of the service is that no individual or group of users should pay less than its incremental cost.

Like prices that are too low, prices that are too high can also interfere with efficient use of the service. In particular, suppose some prices were greater than stand-alone cost. There might then be users willing to pay this cost, but not willing to pay the higher cost imposed by the seller's prices. Such users would

[5] For alternative discussions of network effects as a source of market failure, see Economides (1996) and Weinberg (1997).

be inefficiently excluded by prices that exceed stand-alone cost. Efficiency also requires prices below stand-alone costs.

There is a natural tendency for market forces to produce prices that respect the bounds of incremental and stand-alone costs. If there are no barriers to the entry of new competitors, then the threat of such entry will serve to discipline the pricing and cost allocation practices of incumbent suppliers. Suppose, for example, that a group of customers is collectively paying more than its stand-alone costs. This market segment would be particularly vulnerable to entry by an alternative provider. The threat of such entry will limit the ability of the incumbent to charge more than stand-alone cost.

The threat of competition, which prevents any individual or group from bearing too large a share of common costs, also prevents anyone from bearing too small a share. If a provider is to at least break even on the sale of services and tries to charge some group less than their full incremental cost, then the provider must recover from other users all of the common costs plus the deficit created by undercharging the favored group. Consequently, some set of buyers must pay more than their stand-alone cost. With potential competition, however, this allocation of costs is not sustainable. Potential competition therefore places both an upper and a lower bound on how much a customer or group of customers can bear. The lower bound is the incremental cost of serving those users, while the upper bound is stand-alone cost or the incremental cost of adding those users to a competing network. Note that these bounds are the same as those that guard against inefficient use of a service. In short, competitive pressures prevent inefficiency.

The evaluation of cost allocations on efficiency grounds is complicated by the fact that incremental cost can be difficult to measure. The categorization of costs as attributable and common is not always straightforward. Even more difficult, however, is the identification and measurement of the benefits that one individual's or group's participation brings to others. On the other hand, it is relatively easy to determine whether there are significant barriers to competition. If one can guard against such barriers, then market forces will tend to produce cost allocations that respect the bounds of incremental and stand-alone cost.

Government Intervention

In the United States and other countries, communications industries have typically been the object of substantial government intervention. Government agencies or government-owned firms have typically provided postal services and, in many countries, telecommunication services. In other cases, such as telecommunications in the United States, provision of services by private enterprises has been subject to substantial price and product regulation.

The structural characteristics of communications industries drive government intervention. There are, however, two distinct views about how these characteristics motivate intervention. These industries are conducive to relatively concentrated markets, which could give sellers the ability to exercise monopoly (or near monopoly) power over prices. One common view is that government intervention in communications industries is motivated by a desire to limit anticompetitive behavior in markets that have natural monopoly characteristics.

An alternative view states that government intervention is motivated by a desire to place the cost allocation problem inherent in the pricing of communication services under political control. In communications industries, government intervention has tended to tilt the allocation of common costs away from those buyers with high attributable costs. This group of buyers often represents individuals in remote, rural locations. For instance, postal rates are independent of the location to which mail is sent, although delivery costs are clearly higher in rural areas. Also, when there are scale economies associated with service to individual buyers, the per-unit attributable costs of serving large commercial and industrial users will be less than those of serving small residential users.

When government intervenes to allocate service costs away from some users and toward others, it might appear that the latter are subsidizing the former. Intuitively, we might say that an individual buyer or a market segment is subsidized if it is paying less than its share of production costs. As emphasized earlier, however, common costs make it difficult to unambiguously define the share of total costs borne by an individual or group. Subsidization is less ambiguously defined with reference to incremental costs. That is, an individual or group is *subsidized* if it pays less than its incremental cost. If the provider must cover all costs while subsidizing a set of buyers, payments received from other buyers must be covering more than 100 percent of the common costs. In this case (and only in this case) we say that some buyers *cross-subsidize* others. As previously noted, competition or potential competition will limit a seller's ability to engage in pricing that results in such subsidies.

Government intervention that respects the bounds of incremental and stand-alone costs can be consistent with the efficient provision of services. The history of public sector intervention in communications markets suggests that sometimes the beneficial treatment of groups has gone further, resulting in prices that are below incremental cost. First-class mail service to many hard-to-reach endpoints, for instance, is widely believed to be subsidized. This sort of cross-subsidization, however, is only possible if there are limits on competition. Prices in a market segment in which the seller enjoys a legally protected monopoly are not constrained to be below stand-alone cost. The seller might then be able to raise enough revenues in the protected segment to cover any losses incurred in selling to a subsidized segment.

When cost allocations are subject to political control, through either the regulation of private providers or the public provision of services, allocation choices are often justified in terms of access. Governments have tended to view themselves as guarantors of widespread access to communication systems. This interest in access has sometimes been motivated by the view that the universality of a communication network is an inherently worthy goal. In other instances, the motivation arises from the concern for the consequences of market outcomes for certain high-cost segments of users—rural postal customers, for example. In either case, interest in access may result in cost allocations in which some users subsidize others.

3. A LOOK AT OTHER COMMUNICATIONS INDUSTRIES

Communications industries, we have argued, are characterized by common costs—costs that cannot be uniquely attributed to any particular user or set of users. Government intervention in such industries is often aimed at altering the allocation of common costs across users. In the name of universal access, such intervention often reduces the portion of common costs borne by some users. Legal barriers to competition aid in cost shifting, but distort the decisions of potential competitors.

The twentieth-century experience of two prominent communications industries—telecommunications and postal services—offers valuable insights. In both there are significant common costs and a tendency toward few competitors. Both were subject to significant government intervention that shifted the incidence of common costs and raised barriers to competition, although in recent decades these barriers have come under pressure. Public policy has responded very differently in each industry with divergent results, particularly with regard to technological innovation. The history of these two industries offers revealing lessons for the Federal Reserve's role as a payment system innovator.

Telecommunications

For many decades, the telecommunications industry adhered to the model of protected, regulated monopoly.[6] The prevailing industry structure had its beginnings in the 1920s, when AT&T was allowed to amass a virtual monopoly in phone services and operate free from the threat of competition. In exchange, AT&T made large sunk investments in infrastructure to extend the national network and subjected itself to rate-of-return regulation that sought to keep charges to any buyers from being "too high." This deal was supported by AT&T's argument that telephone service was a natural monopoly and that it (AT&T) could provide universal access at lower cost than could a fragmented industry.

[6] For further resources on the history of the U.S. telecommunications industry, see Brock (1986) or Bornholz and Evans (1983).

For basic local telephone services, buyer-specific fixed costs are significant and variable costs are low. Hence, attributable costs per call tend to fall with the number of calls over a wide range. Large industrial and commercial users' average attributable costs are likely to be lower than those for small business and residential users. The public interest in widespread access has typically promoted price structures that mute these cost differences by shifting common costs away from small users and toward large business users. In addition, cost allocations tended to favor local service at the expense of long distance.

Through a series of moves by market participants and regulators, the structure of the telecommunications industry has evolved from one of an integrated, regulated monopolist to one of more open competition. The Consent Decree of 1982, which settled a Justice Department antitrust case against AT&T, brought competition to long distance markets, while the regional Bell companies retained monopoly positions in local telephony.

Regulated pricing of local service continued to attempt to shift common costs away from high-cost residential and rural users in particular. Such an allocation required higher recovery from large commercial users and contributed to commercial customers' interest in alternatives to the regional Bells' local service, particularly with the proliferation of fax and data services. The long-standing status of local service providers as protected, regulated monopolies was increasingly unsustainable in the changing technological environment. The 1996 Telecommunications Act opened all markets to competition, and explicitly recognized that doing so would put pressure on the industry's ability to provide inexpensive access to such high-cost users as rural hospitals.

The dismantling of barriers to competition in telecommunications has been accompanied by rapid adoption of new technologies. While the by-pass services that hastened the arrival of competition were made possible by technological progress, competition itself has accelerated technological change by encouraging innovation. In the process, the telecommunications industry and its regulators have retreated from the goal of providing access through subsidized cost allocations.

Postal Services

The U.S. government has been involved in postal services since the founding of the nation and has long made universal access the central goal of federal postal policy. In the nineteenth century, the flow of information arising from a universally accessible and affordable postal service was seen as an important factor in the growth of a nation. The U.S. Postal Service's legal monopoly status has been seen as essential to the goal of universal access. With its protected position, the Postal Service can deliver first-class mail to all locations in the United States at a single price. Without this protection, competitors would "skim the cream" by taking low-cost local business, thereby raising the costs of serving the remaining markets. This view suggests cross-subsidized pricing, since prices that are free of subsidies would be immune to cream-skimming.

The Postal Service's legal monopoly on first-class mail appears to have affected other markets in which the monopoly does not apply. In parcel post and package delivery, for instance, private firms are allowed to compete directly with the U.S. Postal Service, although they are significantly constrained in their ability to do so.[7] Critics have claimed that the Postal Service uses funds earned in the protected first-class market to offer new services priced below incremental cost in the more competitive package delivery business. A private, profit-seeking provider might not have an incentive to engage in such pricing of new products; unprofitable entry into a new market is not a compelling goal by itself. As a public entity, however, the Postal Service's motivations are less well defined. While a public entity is charged with serving the public interest and may generally seek to do so, it is hard to prevent at least some decisions from being motivated by other goals. Entry into a new market, for instance, may enhance the overall size and influence of the organization.

Without the discipline of potential competition, the U.S. Postal Service's incentives to maintain and enhance the cost efficiency of its operations are muted. Some observers have noted the difficulties the Postal Service has experienced in the automation of mail processing.[8] At the same time, potential competitors incentives to develop innovative products and processes may well be blunted by the Postal Service's ability to subsidize its prices in competitive market segments.

In short, the postal services and telecommunications industries in the United States have followed divergent paths. While the telecommunications industry has placed increasing reliance on markets to provide pricing discipline and incentives to innovate, the U.S. Postal Service has retained a protected monopoly structure that may distort competition and can stifle technological progress. And while in telecommunications the pace of technological innovation has been quite brisk, with the U.S. Postal Service the pace has been relatively slow.

4. FEDERAL RESERVE CHECK CLEARING

Check collection and other payment services share many features with network communications industries like telecommunications and postal services. From the earliest years Reserve Banks have enjoyed legal privileges that have aided the Fed's entry into check collection and have made the shifting of common costs in the pursuit of universal access at least possible. Some competitive advantages remain today, most notably the "six-hour monopoly," which we

[7] Sidak and Spulber (1996) give a detailed account on the restrictions facing private carriers.

[8] Sidak and Spulber (1996).

discuss below. These privileges make it theoretically possible for the Fed to subsidize some check-clearing services, in the specific sense that term was defined above. If the Fed were engaged in subsidization, by our definition, the Fed's presence could detract from economic efficiency. Moreover, as demonstrated by the contrasting cases of telecommunications and postal services, the capacity to subsidize would not bode well for the Fed's ability to innovate in the public interest. The critical question regarding Fed participation in check collection, then, is whether under barriers to competition some check collection services are in fact subsidized. If so, then the Fed's participation would not only detract from economic efficiency but could also distort the innovative process.

The Six-Hour Monopoly

The Federal Reserve Banks enjoy certain legal privileges in the check collection business. The most important is the Reserve Banks' right to present checks to a paying bank until 2:00 p.m. and receive payment the same day; private-sector banks must present by 8:00 a.m. in order to insist on same-day funds. In practice, private-sector banks can and often do present after 8:00 a.m., but only after negotiating a voluntary agreement with the paying bank, presumably offering the paying bank compensation in the form of reciprocity or presentment fees. The Reserve Banks need not obtain prior permission. Thus, the Reserve Banks enjoy a six-hour monopoly on free par presentment for same-day funds. Other advantages also exist but they appear to be of minor significance.[9]

The six-hour monopoly originated shortly after the founding of the Federal Reserve System. The Federal Reserve Act of 1913 authorized the Reserve Banks to offer check collection services to their member banks. An amendment enacted on June 21, 1917, extended this authorization to allow the Reserve Banks to clear checks for all banks. The amendment also prohibited charging presentment fees against Reserve Banks, but this provision only applied to banks that voluntarily joined the Fed's collection system.[10] The prohibition

[9] The Reserve Banks voluntarily refrain from presenting between noon and 2:00 p.m. in most markets. The six-hour monopoly is not the only legal presentment privilege enjoyed by the Reserve Banks. For example, private-sector banks do not have as much flexibility as Reserve Banks in choosing where to present checks to paying banks. In addition, the paying bank controls the intraday timing of payment to a private-sector presenting bank, while the Reserve Banks have the right to debit the paying bank's account within a specified time period. Because the other legal privileges appear to be of minor significance relative to the six-hour presentment monopoly, we will focus on the latter, although what we have to say will apply equally well to these other privileges. See Board of Governors (1998) and General Accounting Office (1989) for more details.

[10] The amendment provided that any bank could make "reasonable charges, to be determined by the Federal Reserve Board, but in no case to exceed 10 cents per $100," but that "no such charges shall be made against the Federal Reserve Banks." An opinion of the U.S. Attorney General established that this latter provision applied only to banks that voluntarily joined the Fed's clearing system. Note that a state-chartered bank did not have to become a member of the Federal Reserve System in order to participate in the Fed's check collection plan.

codified and expanded a stipulation the Federal Reserve Board had imposed earlier by regulatory fiat on member banks.[11] Banks retained the right to charge presentment fees to any other banks presenting by mail, however. Only the Reserve Banks could mail checks to participating banks and demand immediate par settlement.

The Fed's par presentment privilege was by all accounts essential in the subsequent growth of the Reserve Bank check collection system. The ability to present at par to member banks gave the Reserve Banks a cost advantage over competitors. This advantage gave nonmember banks an incentive to join the Fed's collection system to obtain access to low-cost presentment at member banks. The Reserve Banks required that banks joining the system also agree to *accept* presentment at par. The upshot was that the more banks that joined the Fed collection system, the greater the value of joining.[12]

From its founding in 1913, the Federal Reserve was eager to increase participation in the Reserve Banks' check collection system. For members of the Federal Reserve System, access to the system was a benefit that offset, in part, the cost of stricter Fed reserve requirements, while nonmembers gained the ability to present to participating banks at par. Despite these benefits, the Fed never completely monopolized interbank check collection. For some nonmember banks the income from presentment fees was apparently worth more than the net value of lower-cost clearing services available from the Reserve Banks, so these "nonpar banks" continued to charge presentment fees, a practice that persisted for decades.[13]

The Monetary Control Act (MCA) of 1980 dramatically changed the nature of the Fed's check collection service. The MCA required Reserve Banks to charge fees for their payment services which must, over the long run, cover the direct and indirect costs of providing the services, including imputed costs

[11] The first Reserve Bank check-clearing arrangement, the so-called "voluntary plan" adopted in 1915, required that member banks joining the plan accept checks at par from the Reserve Banks. The "compulsory plan" adopted in May 1916 also included the same requirement but had the Reserve Banks covering the expense of shipping notes or lawful money from the bank to the Reserve Bank in payment. Such expenses were obviously not the only paying bank costs attributable to check collection. Note that because nonmembers had to agree voluntarily to join the Fed clearing plan, the amendment gave the Reserve Banks no real advantage over private banks, since both needed to offer inducements to obtain par presentment rights. The amendment's effect was to codify the Reserve Bank's right to present to member banks at par, by mail, without prior permission. For discussions of the Fed's entry into check clearing see the classic account of Warren Spahr (1926), or more recently, Ed Stevens (1996, 1998) and Alton Gilbert (1998).

[12] Note that the effect of the size of the Fed check collection system on the value of joining did not necessarily reflect a network effect. Federal Reserve policy deliberately tied the service of collecting a bank's outgoing checks to that bank's willingness to pay par on its incoming checks. There was no technological link between the number of banks sending checks to the Fed and the number of banks to which the Fed could send checks.

[13] See Jessup (1967) and Stevens (1998).

that would be incurred if the services were provided by a private firm.[14] The MCA also imposed uniform reserve requirements on all depository institutions and granted nonmembers access to Reserve Bank payment services. Prior to the MCA, free check clearing was one of the benefits of membership. Access to Fed services was now divorced from membership and was explicitly priced.

By forcing the Reserve Banks to charge prices that cover actual and imputed costs, the MCA went a long way toward leveling the competitive playing field. The Fed retained presentment privileges nonetheless. Private collecting banks had no practical means of obtaining same-day funds.[15] In response to public concerns about the remaining asymmetry, the Board sought public comment in 1988 on a proposal to extend Reserve Bank presentment rights to private-sector banks, allowing them to present until 2:00 p.m. for settlement the same day. Corporations objected to the proposal, however, because it would hamper their ability to manage their accounts within the day.[16] The compromise that was finally adopted, effective January 1994, established the current regime in which all banks have the right to same-day settlement for checks presented by 8:00 a.m. The Reserve Banks retained the privilege of presenting until 2:00 p.m. for same-day funds.[17]

The six-hour monopoly could give the Reserve Banks an advantage over competitors in some market segments. It means that the Reserve Banks can collect a given set of checks on better terms than a private provider: for example, by offering a later deposit deadline or better availability (less check float). A private-sector competitor would have to incur additional costs to clear the same checks with the same availability. In some markets, particularly for small and remote depository institutions where transportation time can be

[14] The Federal Reserve's cost recovery requirement includes a "private sector adjustment factor" that consists of the taxes, fees, and return on capital applicable to a comparable private-sector provider.

[15] The rights of private collecting banks were governed by provisions of the Uniform Commercial Code. For a description, see General Accounting Office (1989), p. 28.

[16] In arrangements called "controlled disbursements," banks notify their corporate customers early in the day of the value of the corporation's checks presented that day, allowing the customers to fund their accounts by selling money market securities. Later presentment makes such arrangements more difficult because money markets become progressively less liquid in the afternoon. These costly efforts effectively skirt the prohibition on interest on corporate demand deposits and are wasteful from society's point of view. Note that corporate objections to extending private presentment time to 2:00 p.m. are not directly relevant to the question of whether private and Reserve Bank presentment times should be equalized; presumably they would also object if asked whether the Reserve Banks should be able to present at 2:00 p.m. The objections might suggest that, without interest on corporate checking accounts, equalization should take place at a time earlier in the day rather than later. See Board of Governors (1991), p. 4747, for discussion of public comments on the 1988 proposal.

[17] The Board of Governors has recently requested public comment on the effect of the January 1994 same-day settlement rule. In addition, the Board is considering reducing or eliminating legal disparities between Reserve Banks and private-sector collecting banks in the check collection process, including the six-hour monopoly (Board of Governors 1998).

significant, this advantage has given the Reserve Banks a dominant market share. Indeed, in some locations only the Fed presents checks. In more geographically concentrated markets—large cities, for example—the six-hour monopoly provides little or no competitive advantage and the market share of the Reserve Banks is correspondingly low.

The Allocation of Common Costs

How do the Fed's check collection activities affect the allocation of the common costs? Since implementation of the MCA in the early 1980s, the Reserve Banks price structure has determined the allocation of common costs. Early on, Reserve Bank pricing under the MCA was relatively uniform, although prices varied according to the destination of the check. At first, prices at various Fed offices depended only on whether the item was bound for a city or a remote location. More recently, the price structure has become increasingly complex with finer geographical differentiation.

The increasing complexity of the Reserve Banks' pricing has been a response to competitive pressures. Initially, alternatives to Fed check clearing were not well established. As private-sector clearing has grown over time, increased price differentiation has lowered margins in market segments in which alternative providers can compete effectively with the Fed. Maintaining full cost recovery then requires higher margins in market segments where customers have relatively few viable alternatives. Such markets are generally those in which the Fed's six-hour monopoly supports a dominant market share—presentment to remote banks. Accordingly, common costs have shifted away from market segments in which the six-hour monopoly yields no significant competitive advantage for the Fed—presentment to city endpoints.

The six-hour monopoly could allow the Fed to set prices below incremental costs so that subsidization results. We previously noted that in industries which have substantial common costs (like communications), competitive pressures constrain the way those costs can be allocated across market segments; market discipline generally prevents subsidization. Governmental barriers to competition can loosen the constraints of competitive pressure, however, because they allow over-recovery of costs in protected market segments in order to fund prices below incremental costs in other market segments. The six-hour monopoly is exactly this type of barrier to competition. By raising the costs of competitors, this advantage could allow the Reserve Banks to charge more than stand-alone cost in the protected market segment (checks drawn on remote banks) in order to price below incremental cost in contested market segments (checks drawn on city banks). While these prices could further the goal of universal access, they would be detrimental to economic efficiency, since some users would face prices below incremental social cost.

Reserve Bank price setting is constrained by a specific methodology designed to prevent cross-subsidies. Per-item fees must be above "floor cost," which is defined essentially as average (attributable) variable cost. The individual check is not the only relevant increment, however. There are often significant costs that are attributable to a group of checks but not specifically attributable to individual checks. For example, local transportation costs are attributable to the collection of checks drawn on a particular group of banks, though not to an individual customer or item. The total floor cost for a group of checks is an underestimate of incremental cost if it excludes costs that are attributable to that group of checks but not to any individual item.[18] It is also possible that floor costs overstate incremental costs, since network effects, if they exist, reduce the true incremental cost of serving a market segment.

We need to entertain two alternative hypotheses, therefore, about the Fed's allocation of common costs. One hypothesis is that the Reserve Banks generally do not set fees below incremental costs or above the stand-alone costs. The other is that in some market segments the Reserve Banks set some fees below incremental costs and thus set fees above stand-alone costs elsewhere.[19] These two hypotheses have different implications, as we will see, for how we approach questions about the Fed's role in payment system innovation.

Access

As noted earlier, the Federal Reserve lists payment system accessibility as an important policy goal.[20] The usual articulation of this goal speaks of the Fed providing payment services to all depository institutions, particularly "smaller institutions in remote locations that other providers might choose not to serve."[21] Since there is undoubtedly *some* price at which alternative providers would choose to serve a given location, access to the payment system must be interpreted in terms of the cost of payment system services to small and remote banks. Enhancing access to the payment system must mean lowering the cost to small and remote banks.

Does the Fed lower the costs of check clearing for small and remote banks? We have argued that the Fed's presence tends to shift common costs toward checks drawn on remote banks. Hence, cost allocation among *banks* is determined by whether checks *drawn on* remote banks make up a smaller portion of the checks *collected by* remote banks than they do of checks collected by

[18] Critics who have charged the Fed with unfairly subsidizing check collection have focused on whether the Fed's cost accounting methodology understates the overall cost of Fed check collection. This question is separate from the question we discuss: cross-subsidization within Fed check processing. The Board of Governors requires that the Reserve Banks annually recover the full cost of check collection services from check collection fees.

[19] Our reasonable hypothesis is that the Reserve Banks recover the full costs of check collection in the aggregate.

[20] The Monetary Control Act states that prices for Federal Reserve services "shall give due regard to competitive factors and the provision of an adequate level of such services nationwide."

[21] Board of Governors (1990), p. 295.

city banks. If so, the Fed's presence tends to favor small and remote banks. Although to our knowledge no formal data is available, anecdotal evidence suggests that the difference, if there is one, is not large. The shift of common costs toward checks drawn on remote banks does not appear to alter appreciably the relative burden imposed on small and remote banks. There are, however, other dimensions of pricing along which the Federal Reserve may still be able to pursue a goal of moderating costs for small and remote banks, although direct quantitative evidence is unavailable.[22]

While we lack direct evidence on the extent to which the Fed shifts common costs away from small and remote banks, some indirect evidence is available. Last year Federal Reserve Board Vice Chair Alice Rivlin headed a committee that examined the role of the Federal Reserve in the payment system.[23] As part of its work, the committee held a series of public forums. Many participants at these forums expressed the widely shared belief that the Fed's exit from check clearing would raise the cost of check collection to small and remote banks. Thus according to many people intimately involved in the check collection industry, the Fed's cost allocation does have the effect of enhancing universal access. A reasonable working hypothesis is that the Fed's presence does shift at least some common costs away from small and remote banks.[24]

5. THE FED AS A PAYMENT SYSTEM INNOVATOR: ELECTRONIC CHECK PRESENTMENT

We have argued that the Federal Reserve's involvement in the check collection industry closely parallels government involvement in the telecommunications and postal services industries. Under this view, the Fed promotes universal access by shifting common costs in the presence of legal barriers to competition. Rapid technological change is currently creating new opportunities for innovation in payment services. As a major provider of payment services, the Federal Reserve must determine its appropriate role in pursuing and promoting innovations.

Our reading of the history of communications industries strongly suggests that barriers to competition are fundamentally incompatible with the efficient

[22] For instance, Reserve Banks' prices depend on the amount of sorting done by depositing banks prior to depositing checks with the Fed. Small, remote banks are more likely to make unsorted deposits than are large, city banks. The Fed could pursue its interest in access by setting lower price-cost margins for unsorted than for sorted deposits, thereby lowering the cost of check collection for small, remote banks.

[23] Federal Reserve System (1998).

[24] The shift of common costs away from small and remote banks might be independent of the six-hour monopoly. Some participants in the Rivlin Committee Forums believe that the Federal Reserve accepts a lower rate of return than would be required by commercial providers or that the Fed does not account for the full costs of providing service.

adoption of new technologies. Barriers weaken the effectiveness of an organization's innovative efforts, and they create opportunities for subsidies that can distort the choices users make with respect to new technologies. For both reasons, truly good innovations may fail to reach the market, while unworthy ones may actually take hold. Without barriers to competition, cross-subsidization would not be sustainable, and so we can have confidence that the innovative process is genuinely beneficial.

How does one resolve the conflict between cross-subsidization and innovation? One approach is to measure incremental costs rigorously in order to prevent subsidization. This approach, in essence, is the Federal Reserve's current practice. Earlier, however, we pointed out that the need to gauge incremental costs and network effects across a wide assortment of user subgroups is likely to make comprehensive measures of incremental costs difficult to obtain. Accounting data alone are not likely to convince a skeptic of the absence of cross-subsidies.

An alternative approach to the conflict between cross-subsidization and innovation as it pertains to Reserve Banks is to remove the conditions that might lead to cross-subsidization. In the absence of special legal privileges, competitive pressures will preclude cross-subsidization, as we defined it earlier. Removing the remaining barriers to competition would clearly demonstrate the Fed's commitment to efficient innovation.

These principles apply to the Fed's current efforts to implement ECP. As with any innovation, the near-term prospects of ECP are uncertain. A recent study by Joanna Stavins (1997), an economist at the Federal Reserve Bank of Boston, attempts to quantify the overall costs and benefits to society of a transition to ECP. One advantage would come from replacing the resource cost of transporting and processing paper checks with the lower cost of sending electronic messages. On the other hand, some people prefer to get their checks back. Further, under a variety of state laws, certain check writers are either entitled or required to receive their canceled checks. While the estimates reported by Stavins favor ECP, the results are sensitive to reasonable alternative assumptions, particularly with regard to the intrinsic value of canceled checks to consumers. As with other recently proposed payment innovations, such as stored-value ("smart") cards, it is probably too early to tell whether ECP will make society better off or not.

Ideally, innovations would succeed in the marketplace if and only if they were truly beneficial to society. Accordingly, the Fed should introduce ECP in such a way that we can be assured it will succeed if and only if it improves payment system efficiency. In the absence of impediments to competition, a new product or service generally will be profitable if its value to customers, as measured by willingness to pay, exceeds the cost at which providers are willing to supply it.[25] The usual presumption is that innovation in competitive settings yields outcomes that are beneficial to society as a whole. A necessary condition is that prices are not inefficient, that is, they do not embody cross-subsidies. Barriers to competition allow inefficient pricing. One way to ensure that the Fed's implementation of payment system innovations contributes to payment system efficiency, therefore, is to remove artificial barriers to competition like the six-hour monopoly.

Removing barriers to competition would help avoid some of the potential pitfalls that face a public entity participating in a commercial enterprise. The Reserve Banks' special legal status as public institutions, as opposed to private, profit-seeking businesses, could inhibit their pursuit of improvements in products and processes. The structure of Federal Reserve decisionmaking could result in unnecessarily high costs of research and development. It is often difficult for large organizations, particularly public institutions, to respond nimbly to new technological opportunities. The difficulties experienced by the U.S. Postal Service in implementing automation illustrate the challenge of innovating at large, public-sector institutions.

An even more worrisome possibility is that an organization that is not fully subject to market discipline could make wasteful investments designed to hold on to market share. Many observers expect electronic payment instruments, such as debit cards, credit cards, or smart cards, increasingly to displace checks. In this context, ECP could be viewed as an attempt to stem the expected decline in check use. By reducing the cost of paper checks, ECP could slow the transition to fully electronic payment instruments that are even more beneficial. As long as barriers shield the Fed from competitive pressures, there is the potential for the Fed's pursuit of payment system innovations to conflict with payment system efficiency.

Yet there are good reasons for the Fed to pursue ECP research and development. The Fed, the largest processor of paper checks in the economy, maintains a substantial capital stock dedicated to that activity. The Fed would need to integrate ECP investments into its current check collection infrastructure. As a result, the Fed is likely to have a comparative advantage in evaluating the technical characteristics of ECP investments. In addition, the Reserve Banks have strong incentives to pursue innovations that, if successful, would enhance the value of their existing check infrastructure. To the extent that the Fed's decisionmaking mimics that of a private business, the interdependence of paper and electronic check collection gives the Fed appropriate incentives regarding ECP research and development.

[25] We mean profitability in the sense that the expected present discounted value of net cash flows from the introduction of an innovation are positive. The Board of Governors imposes a tighter constraint on Reserve Banks; net cash flows must be positive each year in each priced service line (check collection, automated clearing house, and so on).

Implementing ECP

What does all this mean for the implementation of ECP? Because it is uncertain whether ECP will actually contribute to economic efficiency, the Fed should do everything possible to ensure that ECP flourishes only if genuinely warranted. If ECP truly is to enhance economic efficiency, it ought to be possible to offer it in a competitive market at prices that cover costs and attract users voluntarily. Any implicit cross-subsidy could distort outcomes by driving some prices below costs, so that users find ECP attractive even if social costs exceeded benefits. Similarly, a legal privilege that dampens competitive pressures could artificially tilt users through nonprice incentives toward an ECP service offered by the Fed.

Because the paying bank has the right to insist on presentment of the paper check, a key issue for ECP is inducing the paying bank to accept electronic presentment. Stavins' (1997) estimates indicate that while paying banks realize significant cost savings from ECP, check writers incur increased costs and lose the benefits of receiving canceled checks. Although her estimates suggest a small net gain to paying banks and their customers, there will undoubtedly be some instances in which ECP would raise the net cost to paying banks and their customers. If the total benefits of ECP exceed the total cost for payors and payees combined, then it ought to be possible for paying banks and their customers to be compensated by other participants. Such compensation could take the form of fees for checks presented electronically, or alternatively, charges to paying banks that wish to receive paper checks.

Net revenues from ECP services should cover the full incremental cost of ECP if it is to be implemented without subsidization. In the absence of barriers to competition, the Fed could not systematically violate this bound. Theoretically, the six-hour monopoly gives the Fed the capability to subsidize ECP; paying banks could be induced to adopt ECP before it is efficient to do so. Eliminating barriers to competition like the six-hour monopoly would help ensure that ECP will succeed if and only if it is truly beneficial to society.

One frequent suggestion by ECP advocates is that the Federal Reserve alter its check presentment regulations so that paying banks are required to accept presentment in electronic form as well as paper. Paying banks could no longer insist on presentment of the paper check. This change is consistent with a competitive market approach as long as paying banks who prefer to receive paper presentment are able to compensate collecting banks. If the paying bank's willingness to pay to receive paper exceeds the cost to collecting banks of presenting paper, then the paying bank ought to be able to stay with paper. Otherwise, the paying bank will receive presentment electronically.[26]

[26] ECP was implemented quite rapidly in Switzerland under just such a scheme. Paying banks must pay a substantially higher fee to receive paper checks.

Simply mandating participation by paying banks would short-circuit the competitive discipline imposed by the need to enlist voluntary cooperation. Then an ECP plan that marginally lowers the costs of collecting banks as a whole might succeed, even though it imposes greater additional costs on paying banks and their customers. Such a scheme would not be in society's interests, and yet it might be adopted if acceptance by paying banks of electronic presentment were mandated with no opt-out provision.

What about Access?

We have interpreted access in terms of the costs of check collection to small and remote banks. Fed participation in the check collection system is intended, in part, to make the cost to these banks lower than it otherwise would be. This interpretation is consistent with two alternative hypotheses. First, the Fed's priced services could be free of cross-subsidies, and therefore efficient, even though its allocation of common costs might favor small and remote banks. Second, the Fed's pricing could involve cross-subsidies. In order to maintain prices below incremental costs, the Fed would need to rely on market privileges such as the six-hour monopoly.

If the six-hour monopoly is essential to achieving the Fed's access goals, then its continued presence could distort the implementation of ECP or other innovations in check clearing. If current pricing involves cross-subsidies, then the revenue from customers paying more than their stand-alone costs could be used to push ECP prices below incremental cost. If the Fed's current pricing does not involve cross-subsidization, then the six-hour monopoly is not essential to the status quo price structure. In this last case it should be possible for the Fed to implement ECP efficiently without sacrificing universal access.

Which of these two hypotheses is correct? As we noted above, available data cannot discriminate between the two, and the Fed's floor-cost methodology may not guarantee the absence of subsidies. Moreover, it will always be difficult to objectively verify the absence of cross-subsidies. As long as cross-subsidies are possible, there will be those who question the Fed's actions, particularly with regard to new product offerings. How can the public be confident that the Fed's innovative efforts in the payment system enhance efficiency? The simplest and most transparent measure would be to eliminate artificial barriers to competition like the six-hour monopoly.

6. CONCLUSION

We have drawn lessons for Federal Reserve payment system policy from the history of other communications industries. Government intervention in these industries has been driven largely by the desire to allocate common costs in order to enhance access for some users. We have argued that Federal Reserve Banks' provision of check collection services fits the same pattern.

Providing access conflicts with technological progress when access is supported by subsidized prices and barriers to competition. In the telecommunications industry rapid innovation was stimulated by deregulation that required a retreat from universal access. The U.S. Postal Service provides a contrasting example in which protected markets were maintained but at an apparent cost in foregone efficiency-enhancing innovation. The lesson for the Federal Reserve seems clear: a pursuit of access that makes use of cross-subsidization interferes with the efficient implementation of payment system innovations. Subsidies erode market discipline and distort choices among competing technologies.

Let us emphasize that it is not at all clear that the Fed currently subsidizes any segment of the check collection market. Federal Reserve policy explicitly seeks to prevent subsidization and promote payment system efficiency. With its efforts to promote ECP, the Fed seeks to establish itself as a leader in payment system innovation. The Fed is well suited to understand, evaluate, and help implement new technologies in this area. For the Fed to be an effective leader, however, the public must be confident that its choices are in the public interest. Eliminating any remaining competitive advantages would deny the Fed the capacity to subsidize and thus would enhance the credibility of the Federal Reserve's commitment to payment system efficiency.

GLOSSARY OF COST-RELATED TERMS

Common costs: Costs that cannot be attributed to a particular individual or group. Note that there can be costs that are attributable to a group but common among the members of the group.

Attributable costs: Costs that arise directly from the provision of services to a particular individual, group, or market segment.

Fixed costs: Costs that do not vary with the quantity of a service produced. Fixed costs can be common among all users or attributable to a subset of users.

Network effects: The benefits that one group's participation creates for other users of a communication service.

Incremental costs: The additional cost of extending a given amount of a service to a particular individual, group, or market segment, given that others are already being served. Incremental costs are attributable costs (fixed and variable) less any network benefits created for others by extending service to the particular individual or group.

Stand-alone costs: The cost of providing a free-standing service to an individual or group, in isolation from other users. Stand-alone costs include the value of the network benefits that the group loses by not sharing joint services with other users.

Subsidization: When the payments received from a group are less than the incremental cost of providing service to that group.

Cross-subsidization: When the deficit created by subsidizing one group is made up for by charging another group more than its stand-alone cost.

APPENDIX

Electronic Check Presentment

While many payment system innovations take the form of new payment instruments, electronic check presentment (ECP) is simply a means of bringing modern information technology to bear on the clearing and settlement of a very old payment instrument. The standard method of clearing and settlement begins when the person or firm receiving a check deposits the check in his or her bank. If the check is drawn on a different account at the same bank, the check stays there and is paid with a bookkeeping transfer. Otherwise, the check is physically transported to the bank on which it is drawn (the paying bank). After physical presentment of the check takes place, funds are sent from the paying bank to the collecting bank. Often this process is intermediated by other banks (correspondents), Federal Reserve Banks, or by private contractors (courier services, for example). A check that is not honored for some reason—because of insufficient funds in the check writers' account, for example—is returned to the bank at which it was initially deposited.

With electronic check presentment, consumers and businesses still conduct transactions using paper checks. At some point in the process of clearing the check, the relevant payment information is transferred into electronic form and then sent on to the paying bank. The check itself may or may not continue on its path to the paying bank. If the check is not sent to the paying bank, it is called check truncation. Although truncation is not a necessary part of ECP, many proponents believe that ECP can make its greatest contribution to payment system efficiency in combination with truncation. Indeed, to the extent that there are savings associated with substituting the flow of electronic information for a paper flow, it would seem to make sense to have paper items truncated as early as possible in the clearing process. On the other hand, the occasional need to inspect the physical check suggests that it might be economical for truncation to occur at a more central point in the process in order to concentrate the storage of paper items.

All Reserve Bank offices currently offer paying banks the option of receiving electronic check presentment. Slightly less than 14 percent of the checks processed by the Fed in 1997 were presented electronically. For about another 9 percent, information was sent electronically to the paying bank, although actual presentment was made with paper checks. Reserve Bank representatives are actively involved in several collaborative efforts with industry representatives aimed at finding ways of increasing the use of ECP.

REFERENCES

Bank for International Settlements. *Statistics on Payment Systems in the Group of Ten Countries*. Basle, Switzerland: Bank for International Settlements, 1996.

Board of Governors of the Federal Reserve System. "Collection of Checks and Other Items by Federal Reserve Banks and Availability of Funds and Collection of Checks" (Docket No. R–1009), *Federal Register,* vol. 63 (March 16, 1998), pp. 12700–06.

__________. "Proposed Rule" (Docket No. R–0723), *Federal Register,* vol. 56 (February 6, 1991), pp. 4743–57.

__________. "The Federal Reserve in the Payments System," *Federal Reserve Bulletin*, vol. 76 (May 1990), pp. 293–98.

Bornholz, Robert, and David S. Evans. "The Early History of Competition in the Telephone Industry," in David S. Evans, ed., *Breaking Up Bell*. New York: North-Holland, 1983.

Brock, Gerald W. "The Regulatory Change in Telecommunications: The Dissolution of AT&T," in Leonard W. Weiss and Michael W. Klass, eds., *Regulatory Reform: What Actually Happened*. Boston: Little, Brown and Company, 1986.

Committee on the Federal Reserve in the Payments Mechanism. *The Federal Reserve in the Payments Mechanism*. Washington: Board of Governors of the Federal Reserve System, January 1998.

Economides, Nicholas. "The Economics of Networks," *International Journal of Industrial Organization,* vol. 14 (October 1996), pp. 673–99.

Gilbert, R. Alton. "Did the Fed's Founding Improve the Efficiency of the United States Payments System?" Manuscript. Federal Reserve Bank of St. Louis, January 1998.

Jessup, Paul F. *The Theory and Practice of Nonpar Banking*. Evanston, Ill.: Northwestern University Press, 1967.

Sidak, Gregory J., and Daniel F. Spulber. *Protecting Competition from the Postal Monopoly*. Washington: AEI Press, 1996.

Spahr, Walter Earl. *The Clearing and Collection of Checks*. New York: Bankers Publishing Company, 1926.

Stavins, Joanna. “A Comparison of Social Costs and Benefits of Paper Check Presentment and ECP with Truncation,” Federal Reserve Bank of Boston *New England Economic Review* (July/August 1997), pp. 27– 44.

Stevens, Ed. “Non Par Banking: Competition and Monopoly in Markets for Payments Services.” Manuscript. Federal Reserve Bank of Cleveland, Financial Services Research Group, January 14, 1998.

__________. “The Founders’ Intentions: Sources of the Payments Services Franchise of the Federal Reserve Banks,” Working Paper No. 03–96. Cleveland: Federal Reserve Bank of Cleveland, Financial Services Research Group, December 1996.

U.S. General Accounting Office. *Check Collection: Competitive Fairness is an Elusive Goal*. Washington: U.S. General Accounting Office, 1989.

Weinberg, John A. “The Organization of Private Payment Networks,” Federal Reserve Bank of Richmond *Economic Quarterly*, vol. 83 (Spring 1997), pp. 25–43.

The Bond Rate and Actual Future Inflation

Yash P. Mehra

It is widely believed that bond yields contain useful information about expected inflation. Many have empirically investigated this issue by examining whether the slope of the term structure has any predictive content in forecasting future inflation. That research, however, has produced disparate results. In a series of papers, Mishkin (1990a, 1990b, 1991) and Jorian and Mishkin (1991) report evidence that indicates the slope has predictive content at long horizons but not at short horizons.[1] In contrast, Engsted (1995) investigates whether the spread between the long-term interest rate and the one-period inflation rate predicts future one-period inflation. While this spread *does* help predict future inflation for a number of countries, it does not for the United States.[2]

In this article, I provide new evidence on the predictive content of the bond rate for future inflation using cointegration and error-correction modeling. The empirical work here corrects for two possible shortcomings of the previous research that may account for the disparate results described above. First, I relax the assumption made in previous studies that the ex ante real interest rate is constant. If the ex ante real rate is variable, then short-run movements

The views expressed are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System. The author thanks Alex Wolman, Pierre-Daniel Sarte, and Wenli Li for many helpful comments.

[1] In this research the horizon forecasts of inflation match that of the slope of the term structure. Hence, the result that the slope has predictive content at long horizons but not at short horizons should be interpreted to mean long-term bonds help predict inflation at long horizons, but short-term bonds do not help predict inflation at short horizons. In this article, by contrast, there is no such matching. In fact, I explore the predictive content of the long bond rate at short and long inflation horizons.

[2] Though this spread does Granger-cause the U.S. inflation rate, the sum of coefficients that appear on lagged values of the spread in the inflation equation is small in magnitude. Engsted, however, does not test whether the sum of these coefficients is different from zero.

in the bond rate do not necessarily reflect movements in its expected inflation component. In that environment the predictive content of the bond rate for future inflation should be investigated controlling for the influences of variables that capture movements in the real rate of interest. The empirical results indicate that inferences regarding the predictive content of the bond rate for future inflation are sensitive to such conditioning. Second, the empirical work in this article examines whether the predictive content has changed over time, in particular between pre- and post-1979 periods. Recent research reported in McCallum (1994) and Rudebusch (1995) indicates that the term structure's ability to predict future economic variables may be influenced by the way the Fed conducts its monetary policy.[3] Most economists would agree that since 1979, the Fed has made repeated attempts to bring down the trend rate of inflation and contain inflationary expectations. In that environment an increase in the current bond rate, even if it correctly signals an increase in long-term expected inflation, may not necessarily translate into higher actual future inflation.

The results here focus on the behavior of the nominal yield on ten-year U.S. Treasury bonds over the period 1959Q1 to 1996Q4. The economic variables that appear in the cointegration and error-correction modeling are the bond rate, the actual inflation rate, the nominal federal funds rate, and the output gap. The last two variables control for variations in the real component of the bond rate that are due to funds rate policy actions and the state of the economy. The test results indicate that the bond rate is cointegrated with the actual inflation rate during the full sample period, implying that the bond rate and the inflation rate move together in the long run. Permanent movements in the inflation rate are associated with permanent movements in the bond rate. The estimated error-correction model, however, indicates that a change has occurred in the way these two variables have adjusted in the short run. In the pre-1979 period, when the bond rate rose above the current inflation rate, actual future inflation accelerated. In the post-1979 period, however, the rise in the bond rate was reversed, and actual future inflation did not accelerate. Thus the bond rate signaled an acceleration in future inflation in the period before 1979 but not thereafter.

The results indicate that the above-noted change in the predictive content of the spread for future inflation may be due to change in Fed policy since 1979. In the post-1979 period, future inflation is inversely related to the current stance of Fed policy measured by the real funds rate, indicating Fed policy was geared towards reducing inflation. No such effect is found prior to 1979. Together these results are consistent with the hypothesis that after 1979, Fed policy prevented any increase in inflationary expectations (evidenced by

[3] In the context of rational expectations hypothesis tests, McCallum (1994) shows how the reduced-form regression coefficients depend upon the Fed's policy rule when the Fed smooths interest rates and responds to movements in the long-short spread.

the rise in the bond rate spread) that would have become embodied in higher actual future rates of inflation. As markets understand and believe in such Fed behavior, increases in inflationary expectations will be less common. The bond rate will then increasingly reflect phenomena other than expected inflation, thereby undermining its usefulness as a precursor of actual future inflation.

1. THE MODEL AND THE METHOD

The Fisher Relation, the Bond Rate, and Future Inflation

In order to motivate the empirical work, I discuss what the Fisher relation implies about the predictive content of the bond rate for future inflation. The Fisher relation for the *m*-period bond rate is

$$BR_t^{(m)} = rr_t^{(m)} + \dot{p}_t^{e(m)}, \tag{1}$$

where $BR^{(m)}$ is the *m*-period bond rate, $\dot{p}^{e(m)}$ is the *m*-period expected inflation rate, and $rr^{(m)}$ is the *m*-period expected real rate of interest. The Fisher relation (1) relates the bond rate to expectations of inflation and the real rate over the maturity (*m*) of the bond.

If the expected real interest rate is constant and if expectations of inflation are rational, then the Fisher relation above can be expressed as in (2) or (3):

$$BR_t^{(m)} = rr + \dot{p}_{t+m} - \epsilon_{t+m} \tag{2}$$

or

$$BR_t^{(m)} - \dot{p}_t = rr + (\dot{p}_{t+m} - \dot{p}_t) - \epsilon_{t+m}, \tag{3}$$

where rr is the constant real rate, $\dot{p}_{t+m}$ is the *m*-period future inflation rate, $\dot{p}_t$ is the one-period current inflation rate, and ϵ_{t+m} is the *m*-period future forecast error that is uncorrelated with past information. Equation (2) indicates that the bond rate contains information about the (*m*-period) future inflation rate, and equation (3) similarly shows that the spread between the bond rate and the current inflation rate has information about a change in the future inflation rate.

Testing the Predictive Content of the Bond Rate for Future Inflation

Previous Studies

Equations (2) and (3) above form the basis of empirical work in most previous studies of the predictive content of the bond rate for future inflation. Previous researchers have investigated the term structure's ability to predict future inflation by running regressions that are of the form

$$(\dot{p}_{t+m} - \dot{p}_{t+n}) = a + b\ (BR_t^{(m)} - BR_t^{(n)}) + \epsilon_{1t} \tag{4}$$

and

$$(\dot{p}_{t+m} - \dot{p}_t) = c + d\ (BR_t^{(m)} - \dot{p}_t) + \epsilon_{2t}, \tag{5}$$

where $BR^{(n)}$ is the n-period bond rate, $\dot{p}_{t+n}$ is the n-period future inflation rate, and other variables are as defined before. As can be seen, these regressions are merely rearranged versions of Fisher relations (2) and (3).[4] In (4) the spread between the m-period and n-period nominal interest rates is used to predict the difference between the m-period and n-period inflation rates, and in (5) the spread between the m-period bond rate and the (one-period) inflation rate is used to predict change in future inflation. Regressions like (4) appear in Mishkin (1990a, 1990b, 1991) and those like (5) appear in Engsted (1995). If $b \neq 0$ in (4) or $d \neq 0$ in (5), then that result indicates that the slope of the term structure does help predict future inflation.

But, as noted before, equations (2) and (3) (or regressions [4] and [5]) embody the assumption that the expected real interest rate is constant. This is a questionable assumption. Plosser and Rouwenhorst (1994) in fact present evidence that indicates that the long end of the term structure does seem to contain information about the real economic activity and therefore about the real rate of interest. If the expected real rate is not constant, then the disturbance term in these regressions ([4] or [5]) may be correlated with the spread.[5] In that case ordinary least squares may provide inconsistent estimates of parameters b and d, biasing inferences concerning the predictive content of the term structure for future inflation. Hence, in order to examine robustness to change in the assumption about the real rate, the predictive content should be investigated conditioning on variables that may control for potential short-term movements in the real rate.

Another issue not investigated fully in previous research is that slope parameters in (4) and (5) are likely to be influenced by the way the Fed conducts its monetary policy (McCallum 1994). For example, if the Fed has in place a disinflationary policy, then higher actual inflation may not follow a current increase in the bond rate spread (as in [5]). This could happen if current widening in the bond rate spread causes the Fed to raise the funds rate, leading to slower real growth and lower actual inflation in the future. In this scenario a current increase in the bond rate spread still reflects expectations of rising future

[4] That is, we get (2) and (3) if we impose the restrictions $a = 0$, $b = 1$, $c = -rr$, and $d = 1$.

[5] To explain it further, assume the real rate c in (5) is not constant but in fact moves systematically with certain economic factors as follows:

$$c_t = c_0 + c_1 Z_t + u_t,$$

where Z is a set containing determinants of the real rate. If we replace c in (5) by c_t as above, then the disturbance term in (5) contains terms like $c_1 Z_t + u_t$. If the spread variable in (5) is correlated with the determinants Z_t, then the spread will be correlated with the disturbance term.

inflation. However, the ensuing Fed behavior prevents those expectations that would have been embodied in higher actual inflation. Therefore, in regressions like (5), the estimate of the slope parameter (d) may be small in periods during which the Fed has been vigilant. Those considerations suggest that parameters that measure the predictive content of the term structure for future inflation may not be stable during the sample period.

Cointegration and Error-Correction Modeling

The empirical work here examines the predictive content of the bond rate using cointegration and error-correction modeling. This empirical procedure, as I illustrate below, yields regressions that are similar in spirit to those employed in some previous research but differ in that it includes additional economic variables that control for potential movements in the real rate of interest.

As indicated before, the Fisher relation (1) for interest rates relates the bond rate to expectations of future inflation and the real interest rate. If one assumes that those expectations can be proxied by distributed lags on current and past values of actual inflation and other fundamental economic determinants, then the Fisher relation implies the following regression (6):

$$BR_t = a + \sum_{s=0}^{k} b_s \dot{p}_{t-s} + \sum_{s=0}^{k} c_s X_{t-s} + U_t, \tag{6}$$

where $\dot{p}_t$ is the actual inflation rate, X_t is the vector containing other economic determinants of the real rate, and U_t is the disturbance term. The presence of the disturbance term in (6) reflects the assumption that distributed lags on actual values of economic determinants may be good proxies for their anticipated values in the long run but not necessarily in the short run.[6]

If levels of the empirical measures of these economic determinants, including the bond rate, are unit root nonstationary, then the bond rate may be cointegrated with these variables as in Engle and Granger (1987). Under those assumptions, regression (6) can be reformulated as in (7):

$$BR_t = d_0 + d_1 \dot{p}_t + d_2 X_t + e_t. \tag{7}$$

Equation (7) is the cointegrating regression. The coefficients that appear on $\dot{p}_t$ and X_t in (7) then measure the long-run responses of the bond rate to inflation and its other real rate determinants. I investigate the question whether the bond rate incorporates expectations of future inflation by testing whether the bond rate is cointegrated with the actual inflation rate. My analysis thus views the positive relationship between the bond rate and actual inflation as a long-run phenomenon.

[6] The only assumption I make about the random disturbance term in (2) is that it has a zero mean.

The cointegrating bond rate regression defines the long-run equilibrium value of the bond rate. Should the bond rate rise above its long-run equilibrium value, then either the bond rate should fall, the economic determinants including inflation should adjust in the direction needed to correct the disequilibrium, or both. I examine such short-run dynamic adjustments by building a vector error-correction model that consists of short-run inflation and bond rate equations. The behavior of the error-correction variable, defined below, then provides information about ways the bond rate and inflation adjust in the short run. Therefore, if the error-correction term is positive and statistically significant in the short-run inflation equation, then that evidence can be interpreted to mean that the bond rate signals future inflation.[7]

To illustrate, assume that the bond rate depends only on the inflation rate in the long run and that the expected real rate is mean stationary. The cointegrating regression is then defined by the relation

$$BR_t = a + b\, \dot{p}_t + U_t, \tag{8}$$

where U_t is the short-term error. This variable, defined as the error-correction variable, measures the extent to which the bond rate differs from its long-run equilibrium value in the short run. The presence of cointegration implies the following error-correction model in ΔBR and $\Delta\dot{p}$:

$$\Delta BR_t = c_0 + \sum_{s=1}^{k} c_{1s}\Delta BR_{t-s} + \sum_{s=1}^{k} c_{2s}\Delta\dot{p}_{t-s} + \lambda_1 U_{t-1} + \epsilon_{1t} \tag{9a}$$

and

$$\Delta\dot{p}_t = d_0 + \sum_{s=1}^{k} d_{1s}\Delta BR_{t-s} + \sum_{s=1}^{k} d_{2s}\Delta\dot{p}_{t-s} + \lambda_2 U_{t-1} + \epsilon_{2t}, \tag{9b}$$

where U_{t-1} is the lagged value of the error-correction variable from (8) and where all other variables are as defined above. The presence of cointegration between BR_t and $\dot{p}_t$ implies that in (9) either $\lambda_1 \neq 0$, $\lambda_2 \neq 0$, or both. Thus, if λ_2 is positive and statistically significant, then a rise in the spread ($U_t = BR_t - a - b\,\dot{p}_t$) signals higher actual future inflation. Since the real interest rate is assumed to be mean stationary, not constant, the error-correction equations should be estimated including other (stationary) short-run determinants of the real interest rate.[8]

[7] Miller (1991) has used this methodology to investigate short-run monetary dynamics.

[8] It is worth pointing out that Engsted (1995) uses an equation like (9b) to investigate whether the spread between the bond rate and the actual inflation rate (U_{t-1} in [8] here) helps predict future inflation. He, however, derives this implication of the Fisher hypothesis under the assumptions that expectations of inflation are rational and forward-looking and that the expected real interest rate is constant. To explain it further, consider the following version of the Fisher

Data and Definition of Economic Determinants in the Multivariable Analysis

The empirical work here examines the dynamic interactions between the bond rate and the inflation rate within a framework that allows for movements in the real component of the bond rate. The descriptive analysis of monetary policy in Goodfriend (1993) and the error-correction model of the bond rate estimated in Mehra (1994) indicate that the real component of the bond rate is significantly influenced by monetary policy actions and the state of the economy. Therefore, the economic variables that enter the analysis here are the bond rate, the actual inflation rate, the nominal federal funds rate, and the output gap that measures the state of the economy.

The empirical work uses quarterly data that spans the period 1959Q1 to 1996Q4. The estimation period, however, begins in 1961Q2, the earlier observations being used for lags. In addition, the sample is broken in 1979Q3, and results are provided for subperiods 1961Q2 to 1979Q3 and 1979Q4 to 1996Q4. The bond rate is the nominal yield on ten-year U.S. Treasury bonds (*BR*). Inflation as measured by the behavior of the consumer price index (excluding food and energy) is the actual, annualized quarterly inflation rate ($\dot{p}$). The measure of monetary policy used is the nominal federal funds rate (*NFR*), and the output gap (*gap*) is the natural lag of real GDP minus the natural log of potential GDP; the latter is generated using the Hodrick-Prescott filter (Hodrick and Prescott 1997).[9] The interest rate data are for the last month of the quarter.

Tests for Unit Roots and Cointegration

Cointegration and error-correction modeling involves four steps. First, determine the stationarity properties of the empirical measures of economic

hypothesis (1) for the long-term bond rate:

$$BR_{(t)} = rr + (1 - b) \sum_{j=1}^{\infty} b^{j}\ E_t\ \dot{p}_{t+j}, \tag{a}$$

where rr is the constant real rate and $b = \bar{e}^{\,i} \approx (1 + rr)$ is the discount factor (Engsted 1995). That is, the long bond rate is given as the constant real rate plus a weighted average of expected future one-period inflation rates ($E_t\ \dot{p}_{t+j},\ j \geq 1$). If BR_t and $\dot{p}_t$ are nonstationary and expectations are rational, then the above equation can be reformulated as

$$BR_t - b\ \dot{p}_t \equiv\ S_t = rr + \sum_{j=1}^{\infty} b^{j}\ E_t \Delta \dot{p}_{t+j}. \tag{b}$$

Equation (b) implies that BR_t and $b\ \dot{p}_t$ are cointegrated and that the spread $S_t = BR_t - b\ \dot{p}_t$ is an optimal predictor of future changes in inflation. Engsted (1995) examines the second implication by estimating a VAR in S and $\Delta \dot{p}$ and then testing whether S Granger-causes $\Delta \dot{p}$.

[9] I have examined the sensitivity of results to some changes in specification. For example, alternatively defining output gap relative to a linear trend produces qualitatively similar results.

determinants suggested above. Second, test for the presence of cointegrating relationships in the system. Third, estimate the cointegrating regression and calculate the residuals. Fourth, construct the short-run error-correction equations.

In order to determine whether the variables have unit roots or are mean stationary, I perform both unit root and mean stationarity tests. The unit root test used is the augmented Dickey-Fuller test, and the test for mean stationarity is the one advocated by Kwiatkowski, Phillips, Schmidt, and Shin (1992). Thus a variable X_t is considered unit root nonstationary if the hypothesis that X_t has a unit root is not rejected by the augmented Dickey-Fuller test and the hypothesis that it is mean stationary is rejected by the mean stationarity test.

The test used for cointegration is the one proposed in Johansen and Juselius (1990), and the cointegrating relations are identified by imposing restrictions as in Johansen and Juselius (1994). Also, the cointegrating relations are estimated using an alternative estimation methodology, Stock and Watson's (1993) dynamic OLS procedure.

2. EMPIRICAL RESULTS

Unit Root and Mean Stationarity Test Results

As indicated before, the economic variables that enter the analysis are the bond rate (*BR*), the inflation rate ($\dot{p}$), the nominal funds rate (*NFR*), and the output gap (*gap*). The output gap variable by construction is stationary. Table 1 reports test results for determining whether other variables have a unit root or are mean stationary. As can be seen, the t-statistic ($t_{\hat{\rho}}$) that tests the null hypothesis that a particular variable has a unit root is small for *BR*, $\dot{p}$, and *NFR*. On the other hand, the test statistic ($\hat{n}_u$) that tests the null hypothesis that a particular variable is mean stationary is large for all these variables. These results indicate that *BR*, $\dot{p}$, and *NFR* have a unit root and are therefore nonstationary in levels.

Cointegration Test Results

Table 2 presents test statistics for determining the number of cointegrating relations in the system (*BR*, $\dot{p}$, *NFR*, *gap*). Trace and maximum eigenvalue statistics presented in the table indicate that there are three cointegrating relations in the system.[10] This result holds in both the sample periods 1961Q2 to 1996Q4 and 1961Q2 to 1979Q3.

[10] The lag length parameter (k) for the VAR model was chosen using the likelihood ratio test described in Sims (1980). In particular, the VAR model initially was estimated with k set equal to a maximum number of eight quarters. This unrestricted model was then tested against a restricted model, where k is reduced by one, using the likelihood ratio test. The lag length finally selected in performing the Johansen-Juselius procedure is the one that results in the rejection of the restricted model.

Table 1 Tests for Unit Roots and Mean Stationarity

Series		Panel A Test for Unit Roots				Panel B Test for Mean Stationarity
X	ρ	$t_{\hat{\rho}}$	k	$x^2(2)$	$x^2(4)$	$\hat{n}_u$
BR	0.96	−1.7	5	1.6	1.3	0.80*
$\dot{p}$	0.89	−2.4	2	2.1	1.8	0.39**
NFR	0.89	−2.8	5	1.1	0.40	0.46*

*Significant at the 5 percent level.

**Significant at the 10 percent level.

Notes: BR is the bond rate; $\dot{p}$ is the annualized quarterly inflation rate measured by the behavior of the consumer price index excluding food and energy; and NFR is the nominal federal funds rate. The sample period studied is 1961Q2 to 1996Q4. ρ and t-statistics ($t_{\hat{\rho}}$) for $\rho = 1$ in panel A above are from the augmented Dickey-Fuller regressions of the form

$$X_t = a_0 + \rho\, X_{t-1} + \sum_{s-1}^{k} a_s\, \Delta X_{t-s},$$

where X is the pertinent series. The series has a unit root if $\rho = 1$. The 5 percent critical value is 2.9. The lag length k is chosen using the procedure given in Hall (1990), with maximum lag set at eight quarters. $x^2(2)$ and $x^2(4)$ are Chi-squared statistics that test for the presence of second-order and fourth-order serial correlation in the residual of the augmented Dickey-Fuller regression, respectively. The test statistics $\hat{n}_u$ in panel B is the statistic that tests the null hypothesis that the pertinent series is mean stationary. The 5 percent critical value for $\hat{n}_u$ given in Kwiatkowski et al. (1992) is 0.463 (10 percent critical value is 0.347).

Table 3 presents estimates of the cointegrating relations found in the system. I first test the hypothesis that the three-dimensional cointegration space contains cointegrating relations that are of the form (10) through (12):

$$BR_t = a_0 + a_1\, \dot{p}_t + u_{1t};\ a_1 = 1, \tag{10}$$

$$NFR_t = b_0 + b_1\, \dot{p}_t + u_{2t};\ b_1 = 1, \tag{11}$$

and

$$gap_t = c_0 + u_{3t}. \tag{12}$$

Equation (10) can be interpreted as the Fisher relation for the bond rate and equation (11) as the Fed reaction function. Equation (12) simply states that the output gap variable is stationary. As shown in Johansen and Juselius (1994), these cointegrating relations can be identified imposing restrictions on long-run parameters in the cointegrating space.

In the full sample period, the hypotheses that cointegrating relations are of the form (10) through (12) and that $a_1 = b_1 = 1$ are consistent with data (the x_1^2 statistic that tests those restrictions is small; see Table 3, panel A). However, in the subsample 1961Q2 to 1979Q3, the restrictions that $a_1 = b_1 = 1$ are

Table 2 Cointegration Test Results

Panel A: 1961Q2 to 1996Q4

System	**Trace** **H0**		**Maximum Eigenvalue** **H0 vs H1**	***k***
$(BR,\ \dot{p},\ NFR,\ gap)$	$r = 0$	8.9*	$r = 0$ vs $r \leq 1$: 28.6*	8
	$r \leq 1$	40.3*	$r = 1$ vs $r \leq 2$: 23.9*	
	$r \leq 2$	16.3*	$r = 2$ vs $r \leq 3$: 11.7*	
	$r \leq 3$	4.6	$r = 3$ vs $r \leq 4$: 4.6	

Panel B: 1961Q2 to 1979Q3

System	**Trace** **H0**		**Maximum Eigenvalue** **H0 vs H1**	***k***
$(BR,\ \dot{p},\ NFR,\ gap)$	$r = 0$	66.2*	$r = 0$ vs $r \leq 1$: 33.4*	5
	$r \leq 1$	32.8*	$r = 1$ vs $r \leq 2$: 19.8*	
	$r \leq 2$	13.0*	$r = 2$ vs $r \leq 3$: 10.6*	
	$r \leq 3$	2.5	$r = 3$ vs $r \leq 4$: 2.5	

*Significant at the 10 percent level.

Notes: Trace tests the null hypothesis that the number of cointegrating vectors (r) is less than or equal to a chosen value, and maximum eigenvalue tests the null hypothesis that the number of cointegrating vectors is r, given the alternative of $r + 1$ vectors. The VAR lag length (k) was chosen using the likelihood ratio test in Sims (1980).

rejected by data, and the cointegrating relations are thus estimated without such restrictions.[11] As can be seen, estimates indicate that the bond rate is cointegrated with the inflation rate, but the bond rate does not adjust one-for-one with inflation. Therefore, inflation is the main source of the stochastic trend in the bond rate.

The estimation procedure in Johansen and Juselius (1990, 1994) is a system estimation method. In order to check the robustness of estimates, I also present estimates of the cointegrating relations (10) and (11) using a single equation estimation method. Panel B in Table 3 presents results using the dynamic OLS procedure given in Stock and Watson (1993). As shown in the table, this procedure yields estimates that are remarkably close to those reported above.

Results on the Error-Correction Coefficient in the Error-Correction Model

The cointegration test results described in the previous section are consistent with the presence of cointegrating relations that are of the form

[11] In estimating error-correction equations for the pre-1979 period, I consider cointegrating regressions with $a = b = 1$. This restriction implies that the bond rate does adjust one-for-one with inflation. The basic results do not change if instead this restriction is not imposed (see footnote 15).

Table 3 Estimates of Restricted Cointegrating Vectors

	Panel A: Johansen-Juselius Procedure	
	Sample Period 1961Q2 to 1996Q4	**Sample Period 1961Q2 to 1979Q3**
A1	$BR_t = 3.1 + \dot{p}_t + U_{1t}$	$BR_t = 3.2 + 0.67\ \dot{p}_t + U_{1t}$
A2	$NFR_t = 2.3 + \dot{p}_t + U_{2t}$	$NFR_t = 2.7 + 0.66\ \dot{p}_t + U_{2t}$
	$x_1^2(3) = 0.92\ (0.82)$	$x_2^2(1) = 0.01\ (0.91)$
	Panel B: Dynamic OLS	
	1961Q2 to 1996Q4	**1961Q2 to 1979Q3**
A1	$BR_t = 2.9 + 1.0\ \dot{p}_t + U_{1t}$	$BR_t = 3.2 + 0.66\ \dot{p}_t + U_{1t}$
A2	$NFR_t = 2.2 + 1.0\ \dot{p}_t + U_{2t}$	$NFR_t = 2.5 + 0.67\ \dot{p}_t + U_{2t}$

Notes: Panel A above reports two of the three cointegrating vectors that lie in the cointegration space spanned by the four-variable VAR ($BR, \dot{p}, NFR, gap$). The cointegrating vectors A1 and A2 are the Fisher relation for the bond rate and the funds rate. $x_1^2(3)$ and $x_2^2(1)$ are Chi-squared statistics (degrees of freedom in parentheses) that test the null hypothesis that the identifying restrictions imposed are consistent with data (Johansen and Juselius 1994).

Panel B above reports the same cointegrating vectors estimated using the dynamic OLS procedure (eight leads and lags are used).

$$BR_t = a_0 + a_1\ \dot{p}_t + U_{1t} \tag{13}$$

and

$$NFR_t = b_0 + b_1\ \dot{p}_t + U_{2t}, \tag{14}$$

where U_1 and U_2 are stationary disturbance terms. I now examine the behavior of the error-correction term $U_{1t} = BR_t - a_0 - a_1\dot{p}_t$ in short-run equations of the form

$$\Delta BR = b_0 + \sum_{s=1}^{k1} b_{1s}\ \Delta BR_{t-s} + \sum_{s=1}^{k2} b_{2s}\Delta\dot{p}_{t-s} + \sum_{s=1}^{k3} b_{3s}\Delta NFR_{t-s} + \sum_{s=1}^{k4} b_{4s}\ gap_{t-s} + \lambda_1\ U_{1t-1} + \delta_1\ U_{2t-1} \tag{15a}$$

and

$$\Delta\dot{p}_t = c_0 + \sum_{s=1}^{k1} c_{1s}\ \Delta BR_{t-s} + \sum_{s=1}^{k2} c_{2s}\Delta\dot{p}_{t-s} + \sum_{s=1}^{k3} c_{3s}\Delta NFR_{t-s} + \sum_{s=1}^{k4} c_{4s}\ gap_{t-s} + \lambda_2\ U_{1t-1} + \delta_2\ U_{2t-1}, \tag{15b}$$

where all variables are as defined before. The short-run equations include first differences of the bond rate, inflation, and the funds rate and level of the output gap, even though the last two variables do not enter the long-run bond equation (13). These variables capture the short-run impacts of monetary policy and the state of the economy on the bond rate and other variables. As indicated before, the parameters of interest are λ_1, λ_2 and the sums of coefficients that appear on the bond rate in equation (15b). The expected signs of the error-correction term U_{1t-1} are positive for $\Delta\dot{p}$ and negative for ΔBR.

Following Campbell and Perron (1991), the lag lengths used in the error-correction model are chosen using the procedure given in Hall (1990). This procedure starts with some upper bound on lags, chosen a priori for each variable (eight quarters here) and then drops all lags beyond the lag with a significant coefficient. I do present tests of the hypothesis that excluded lags are not significant, however.

Table 4 reports the error-correction coefficients (t-values in parentheses) when the long-run bond equation is (13). In addition, it also reports the sums of coefficients that appear on (first differences of) the bond rate in the inflation equation. Parentheses that follow contain t-statistics for the sum of coefficients, whereas brackets contain Chi-squared statistics for exclusion restrictions. Panel A reports results for the full sample 1961Q2 to 1996Q4 and panels B and C for the subperiods 1961Q2 to 1979Q3 and 1979Q4 to 1996Q4.[12] In full sample regressions the error-correction coefficient is negative and statistically significant in the bond equation (ΔBR), but in inflation equations ($\Delta\dot{p}$), it is generally small and not statistically different from zero.[13] Furthermore, individual coefficients that appear on two lagged values of the bond rate in the inflation equation are 0.50 and -0.33. These coefficients are individually significant, but their sum is not statistically different from zero, indicating that ultimately, increases in the bond rate have not been associated with accelerations in actual inflation.[14] Together, these results indicate that the short-run positive deviations of the bond rate from its long-run equilibrium values were corrected mainly through reversals in the bond rate. Actual inflation did not accelerate.

The results for the first subperiod 1961Q2 to 1979Q3 reported in panel B of Table 4 are, however, strikingly different. As can be seen, the error-correction coefficient is negative and significant in the bond rate equation but is positive and significant in the inflation equation. These results suggest that

[12] Inflation equations include dummies for President Nixon's price and wage controls.

[13] The error-correction coefficients are in fact negative in the inflation equation that includes other determinants of the real rate. In the inflation equation that includes only lagged values of inflation, the coefficient that appears on the error-correction term is positive, small in magnitude, and not statistically different from zero. The latter result is similar in spirit to the one in Engsted (1995).

[14] This result, of course, means that the bond rate Granger-causes inflation.

Table 4 Granger-Causality Results from Error-Correction Equations: General to Specific, Using Hall Approach

Panel A: Cointegrating Regressions, 1961Q2 to 1996Q4

$\boldsymbol{BR_t = 2.9 + \dot{p}_t + U_{1t}; \; NFR_t = 2.2 + \dot{p}_t + U_{2t}}$

Equation	U_{1t-1}	$\sum_{s=1}^{k1} \Delta BR_{t-s}$	(k_1, k_2, k_3, k_4)	$x^2(sl)$
ΔBR_t	-0.20 (3.5)		(7,7,8,1)	9.5 (0.39)
$\Delta \dot{p}_t$	-0.13 (1.3)	0.17 (0.6) [10.2]*	(2,8,8,8)	5.3 (0.51)

Panel B: Cointegrating Regressions, 1961Q2 to 1979Q3

$\boldsymbol{BR_t = 1.7 + \dot{p}_t + U_{1t}; \; NFR_t = 1.0 + \dot{p}_t + U_{2t}}$

Equation	U_{1t-1}	$\sum_{s=1}^{k1} \Delta BR_{t-s}$	(k_1, k_2, k_3, k_4)	$x^2(sl)$
ΔBR_t	-0.24 (3.5)		(8,7,6,1)	8.9 (0.54)
$\Delta \dot{p}_t$	0.32 (3.2)		(0,0,0,0)	38.4 (0.24)

Panel C: Cointegrating Regressions, 1979Q4 to 1996Q4

$\boldsymbol{BR_t = 4.2 + \dot{p}_t + U_{1t}; \; NFR_t = 2.5 + \dot{p}_t + U_{2t}}$

Equation	U_{1t-1}	$\sum_{s=1}^{k1} \Delta BR_{t-s}$	(k_1, k_2, k_3, k_4)	$x^2(sl)$
ΔBR_t	-0.39 (2.9)		(7,0,6,8)	14.5 (0.21)
$\Delta \dot{p}_t$	-0.01 (0.4)		(0,6,8,8)	11.4 (0.33)

Notes: The coefficients reported are from error-correction regressions that include the bond rate (*BR*), the inflation rate ($\dot{p}$), the nominal federal funds rate (*NFR*), and the output gap (*gap*) (see equation [15] of the text). In addition, the model has two error-correction variables (U_{1t} and U_{2t}). (k_1, k_2, k_3, k_4) refers to lag lengths that are chosen for $BR, \dot{p}, NFR,$ and *gap*. Parentheses contain t-statistics for the error-correction variable (U_{1t-1}) or for the sum of coefficients that appear on the bond rate $\left(\sum_{s=1}^{k1} \Delta BR_{t-s}\right)$. For the latter, brackets contain the Chi-squared statistic for the null hypothesis that every coefficient in this sum is zero. $x^2(sl)$ tests the null hypothesis that remaining lags are not significant (significance levels follow in parentheses).

positive deviations of the bond rate from its long-run equilibrium value were eliminated partly through declines in the bond rate and partly through increases in actual inflation. Consequently, in the pre-1979 period, actual inflation did accelerate when the spread between the bond rate and the one-period inflation rate rose.[15]

[15] I get similar results if cointegrating regressions (13) and (14) are estimated without restrictions $b_1 = a_1 = 1$. In particular, over the sample period 1961Q2 to 1979Q3, the error-correction variable U_{1t-1} has a positive coefficient in the inflation equation, indicating that actual inflation did accelerate following an increase in the bond rate spread.

In the aforementioned result, the fact that the spread signaled an increase in inflation in the pre-1979 period but not in the full sample period implies that the spread must have lost its predictive content in the post-1979 period. This implication is consistent with the subperiod results reported in panel C of Table 4: the error-correction term is no longer significant in the inflation equation.

Comparison with Previous Studies

The full sample results discussed in the previous section indicate that the spread between the bond rate and the one-period inflation rate does not help predict one-quarter-ahead changes in the rate of inflation. Since inflation is a unit root process, the results above also imply that the spread has no predictive content for long-horizon forecasts of future inflation. The latter implication is in contrast with the finding in Mishkin (1990a, 1990b, 1991) that at long horizons the long end of the slope of the term structure does help predict future inflation.

As indicated before, an important assumption implicit in the regressions used by Mishkin is that the ex ante real rate of interest is constant. This assumption may not be valid. Therefore, the predictive content of the spread for future inflation should also be investigated conditioning on variables that capture changes in the short-run determinants of the real rate.

In order to illustrate whether results are sensitive to such conditioning, I also investigate the predictive content of the spread between the bond rate and the (one-period) inflation rate for future inflation by estimating regressions of the form

$$(\ln[P_{t+m}/P_t]/m) - \ln(P_t/P_{t-1}) = a_0 + \lambda_c\ U_{1t} + V_{1t}, \tag{16}$$

$$\begin{aligned}(\ln[P_{t+m}/P_t]/m) - \ln(P_t/P_{t-1}) = a_0 + \lambda_d\ U_{1t} + \sum_{s=1}^{k1} a_{1s}\ \Delta\dot{p}_{t-s} \\ + \sum_{s=1}^{k2} a_{2s}\ \Delta NFR_{t-s} + \sum_{s=1}^{k3} a_{3s}\ \Delta BR_{t-s} + \sum_{s=1}^{k4} a_{4s}\ gap_{t-s} + V_{2t},\end{aligned} \tag{17}$$

and

$$\begin{aligned}(\ln[P_{t+m}/P_t]/m - \ln(P_t/P_{t-1}) = a_0 + \lambda_e\ U_{1t} + \delta U_{2t} + \sum_{s=1}^{k1} a_{1s}\ \Delta\dot{p}_{t-s} \\ + \sum_{s=1}^{k2} a_{2s}\ \Delta NFR_{t-s} + \sum_{s=1}^{k3} a_{3s}\ \Delta BR_{t-s} + \sum_{s=1}^{k4} a_{4s}\ gap_{t-s} + V_{2t},\end{aligned} \tag{18}$$

where

$$U_{1t} = BR_t - a_0 - a_1\ \dot{p}_t,$$

$$U_{2t} = NFR - b_0 - b_1\ \dot{p}_t,$$

and where m is the number of quarters, and other variables are as defined.[16] U_1 measures the spread between the bond rate and the (one-period) inflation rate and U_2 the spread between the nominal funds rate and the inflation rate. Regression (16) examines the predictive content of the spread for long-horizon forecasts of future inflation without controlling for variations in the spread due to real growth, monetary policy actions, and inflation. Regressions (17) and (18), however, control for such variations. Regression (18) is similar to regression (17) except in that it also includes the current stance of short-run monetary policy measured by the funds rate spread (U_{2t}). The regressions are estimated over the full sample period as well as over subperiods 1961Q2 to 1979Q3 and 1979Q4 to 1996Q4 and for horizons up to four years in the future. In addition, I consider the subperiod 1983Q1 to 1996Q4, during which inflation has remained relatively low.

In Tables 5 and 6, I present estimates of the coefficient (t-values in parentheses) that appears on the bond rate spread variable (λ_c in [16]), λ_d in [17], and λ_e in [18]).[17,18] I also report the coefficient on the funds rate spread variable (δ in [18]). In Table 5 the results are for the full sample period and the first subperiod and in Table 6 for two post-1979 subperiods. If we focus on pre-1979 regression estimates, we will see that they indicate that the bond rate spread does help predict future inflation (see t-values on λ_c, λ_d, and λ_e in Table 5, panel B). This result holds at all forecast horizons and is not sensitive to the inclusion of other variables in regressions. Furthermore, the funds rate spread variable that controls for policy-induced movements in the real component of the bond rate is never significant in those regressions, indicating that at the time the current stance of monetary policy had no predictive content for future inflation. Therefore, the widened bond rate spread was followed by higher actual future inflation during this subperiod.

[16] These regressions differ from those reported in Mishkin (1990a, 1990b, 1991). Mishkin uses zero-coupon bond data, derived from coupon-bearing bonds that have actually been traded. So, he is able to match the horizon of the inflation forecast with that of the term spread. The empirical work here instead uses yield-to-maturity data on coupon bonds and the inflation forecast horizon does not match with that of the term spread. These differences, however, do not reduce the importance of examining the potential role of additional variables that may provide information about movements in the real rate of interest.

[17] The t-values were corrected for the presence of moving-average serial correlation generated due to overlap in forecast horizon. The degree of correction in the moving-average serial correlation was determined by examining the autocorrelation function of the residuals. This procedure generated the order of serial correlation correction close to the value given by $(m-1)$, where m is the number of quarters in the forecast horizon. Furthermore, the use of realized multi-period inflation in these regressions led to the loss of observations at the end of the sample, so that the effective sample sizes are 1961Q2 to 1996Q4-m and 1961Q2 to 1979Q3-m.

[18] All regressions are estimated including four lagged values of other information variables. Furthermore, those lagged values are always statistically significant as a group in regressions (17) and (18).

Table 5 Long-Horizon Inflation Equations

Panel A: 1961Q2 to 1996Q4

Cointegrating Regressions: $BR_t = 2.9 + \dot{p}_t + U_{1t}$; $NFR_t = 2.2 + \dot{p}_t + U_{2t}$

Horizons in	Equation (C)	Equation (D[c])	Equation (E)	
Quarters (m)	λ_c (t-value)	λ_d (t-value)	λ_e (t-value)	δ (t-value)
4	0.16 (1.5)	0.09 (1.0)	0.02 (0.2)	0.07 (0.6)[a]
8	0.20 (1.5)	0.04 (0.4)	0.14 (0.9)	−0.12 (1.0)[a]
12	0.24 (1.6)	0.01 (0.1)	0.21 (1.2)	−0.26 (1.4)[a]
16	0.25 (1.7)	−0.07 (0.6)	0.22 (1.2)	−0.33 (1.7)[a]

Panel B: 1961Q2 to 1979Q3

Cointegrating Regressions: $BR_t = 1.7 + \dot{p}_t + U_{1t}$; $NFR_t = 1.0 + \dot{p}_t + U_{2t}$

Horizons in	Equation (C)	Equation (D[c])	Equation (E)	
Quarters (m)	λ_c (t-value)	λ_d (t-value)	λ_e (t-value)	δ (t-value)
4	0.56 (5.4)	0.61 (7.4)	0.62 (3.3)	−0.00 (0.0)[b]
8	0.81 (7.9)	0.80 (6.5)	1.0 (4.4)	−0.25 (0.8)[b]
12	0.96 (7.9)	0.89 (12.6)	0.94 (3.4)	−0.10 (0.2)[b]
16	0.99 (9.1)	0.99 (13.2)	0.77 (2.8)	0.33 (0.9)[b]

[a] The restriction $\lambda_e + \delta = 0$ is consistent with data.

[b] The restriction $\lambda_e + \delta = 0$ is not consistent with data.

[c] Additional variables included in equations (D) and (E) are always statistically significant as a group.

Notes: The coefficients reported are from regressions of the form

$$p(t,m) = f_0 + \lambda_c \; U_{1t}, \tag{C}$$

$$p(t,m) = g_0 + \lambda_d \; U_{1t} + \sum_{s=1}^{k1} g_{1s} \; \Delta BR_{t-s} \tag{D}$$

$$+ \sum_{s=1}^{k2} g_{2s} \; \Delta\dot{p}_{t-s} + \sum_{s=1}^{k3} g_{3s} \; \Delta NFR_{t-s} + \sum_{s=1}^{k4} g_4 \; gap_{t-s}$$

and

$$p(t,m) = \lambda_e \; U_{1t} + \delta \; U_{2t} + \text{ other variables as in (D),} \tag{E}$$

where $p(t,m)$ is $(log[P_{t+m}/P_t])/m - log(P_t/P_{t-1})$, m is the number of quarters in the forecast horizon, and the rest of the variables are as defined before. All regressions are estimated setting $k_1 = k_2 = k_3 = k_4 = 4$.

The full sample regression estimates, however, suggest strikingly different results. The coefficient that appears on the bond rate spread variable is now about one-third the size estimated in subsample regressions.[19] For forecast

[19] Mishkin (1990a) also finds that in full sample regressions the coefficients that appear on term spreads are generally smaller in size than those in pre-1979 regressions. Nonetheless,

Table 6 Long-Horizon Inflation Equations

Cointegrating Regressions: $BR_t = 4.2 + \dot{p}_t + U_{1t}$; $NFR_t = 2.5 + \dot{p}_t + U_{2t}$

Panel A: 1979Q4 to 1996Q4

Horizons in Quarters (m)	Equation (C) λ_c (t-value)	Equation (D[a]) λ_d (t-value)	Equation (E) λ_e (t-value)	δ (t-value)
4	0.21 (2.0)	0.18 (3.0)	0.29 (2.7)	−0.10 (1.1)[b]
8	0.31 (2.1)	0.28 (3.6)	0.58 (3.7)	−0.26 (2.2)[b]
12	0.35 (2.0)	0.31 (4.1)	0.61 (5.9)	−0.28 (2.9)[b]
16	0.42 (2.2)	0.37 (6.1)	0.73 (6.3)	−0.35 (3.5)[b]

Panel B: 1983Q1 to 1996Q4

Horizons in Quarters (m)	Equation (C) λ_c (t-value)	Equation (D[a]) λ_d (t-value)	Equation (E) λ_e (t-value)	δ (t-value)
4	0.08 (0.4)	0.14 (2.0)	0.08 (0.3)	0.07 (0.07)[b]
8	0.08 (0.9)	0.17 (2.7)	0.10 (4.4)	0.10 (1.1)[b]
12	0.11 (0.9)	0.21 (3.9)	0.30 (3.6)	−0.11 (1.7)[b]
16	0.14 (0.7)	0.28 (4.4)	0.42 (6.2)	−0.22 (2.9)[b]

[a] Additional variables included in equations (D) and (E) are always statistically significant as a group.

[b] The restriction $\lambda_e + \delta = 0$ is not consistent with data.

Notes: The cointegrating regressions are estimated over the period 1979Q4 to 1996Q4. The coefficients reported above are from regressions like those given in Table 5. See notes in Table 5.

horizons up to three years in the future, this coefficient is not statistically significant, and for somewhat longer horizons, it is marginally significant at the 10 percent level (see t-values on λ_c, λ_d, or λ_e in Table 5, panel A). Those estimates suggest there had been a significant deterioration in the predictive content of the bond rate spread for future inflation in the period since 1979. Furthermore, results are now sensitive to variables included in the conditioning set. If we ignore the current stance of Fed policy measured by the funds rate spread, then the bond rate spread has no predictive content for actual future inflation at any forecast horizon (see λ_d in Table 5, panel A). However, when the funds rate spread variable is included in the conditioning set, then in long-horizon inflation regressions, the bond rate spread variable appears with a positive coefficient. Yet in those same regressions the coefficient that appears on the funds rate spread is negative though barely statistically significant (see δ in Table 5, panel

his regressions pass the conventional test of parameter stability. The regressions estimated here, however, do not depict such parameter constancy.

A). This result is consistent with the presence of policy-induced movements in the real component of the funds rate and their subsequent negative effects on future inflation rates. In fact, the coefficients that appear on the bond rate and the funds rate spreads are equal in size but opposite in signs. Those estimates suggest that increases in the bond rate spread accompanied by equivalent increases in the funds rate spread have had no effect on actual future inflation rates.[20]

The first subperiod and full sample results discussed above suggest the presence of considerable subperiod instability. In order to gain more insight into subperiod differences, Table 6 presents estimates from long-horizon inflation equations for two post-1979 subperiods, 1979Q4 to 1996Q4 and 1983Q1 to 1996Q4. Those estimates permit the following inferences about the predictive content of the spread for future inflation in the post-1979 period. First, the predictive content of the spread for future inflation has deteriorated in the post-1979 period. The size of the coefficient that appears on the bond rate spread variable is not only small relative to its value found in the pre-1979 period, but its size declines further during the relative low inflation period of the late 1980s and the 1990s (compare values of λ_c in Tables 5 and 6). Second, the marginal predictive content of the spread for future inflation is now sensitive to variables included in regressions (compare values of λ_c and λ_e in panels A and B of Table 6). Third, the current stance of monetary policy measured by the funds rate spread correlates negatively with future inflation, indicating Fed policy was geared towards reducing inflation in the post-1979 period. In those regressions the bond rate spread variable remains significant, indicating the bond rate spread does contain information about future inflation. However, the results also indicate that actual future inflation may not accelerate following the rise in the bond rate spread if the Fed reacts aggressively by raising the funds rate (see estimates of λ_e and δ in Table 6).

The descriptive analysis of monetary policy in Goodfriend (1993) in fact indicates that since 1979 the Fed has had a disinflationary policy in force to reduce the trend rate of inflation and contain inflationary expectations. Accordingly, this Fed behavior may be at the source of deterioration in the predictive content of the bond rate for actual future inflation. To the extent that rising long-run inflationary expectations evidenced by the rise in the bond rate were triggered in part by news of strong actual or anticipated real growth, the Fed may have calmed those expectations by raising the funds rate. The induced tightening of monetary policy may have reduced inflationary expectations by reducing actual or anticipated real growth, thereby preventing any increase in actual inflation. Given such Fed behavior, observed increases in the bond rate

[20] This result is similar in spirit to the finding reported using cointegration and error-correction methodology.

do not necessarily indicate that actual inflation is going to accelerate in the near term.

3. CONCLUDING OBSERVATIONS

This article views the Fisher hypothesis as a long-run relationship with short-run variation in the real interest rate. The findings show that the bond rate is cointegrated with the inflation rate over the 1962Q2 to 1996Q4 period, which indicates that in the long run, permanent movements in actual inflation have been associated with permanent movements in the bond rate.

The short-run error-correction equations help identify ways in which the bond rate and inflation adjust in the short run. In the pre-1979 period, increases in the bond rate were followed by an acceleration in actual inflation, whereas that did not happen in the post-1979 period. In the latter period, short-run increases in the bond rate have usually been reversed, with no follow-up in actual inflation.

In the period since 1979, the Fed has made serious attempts to reduce the trend rate of inflation and contain inflationary expectations. Such Fed behavior may have prevented the short-run increases in inflationary expectations, as evidenced by increases in the bond rate, from finally producing higher actual inflation. These results imply that if the Fed retains its hard-won credibility for inflation stability, then the bond rate may reflect phenomena other than expected inflation, thereby undermining its usefulness as a precursor of actual future inflation.

REFERENCES

Campbell, John Y., and Pierre Perron. “Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots,” in Olivier J. Blanchard and Stanley Fischer, eds., *NBER Macroeconomics Annual 1991.* Cambridge, Mass.: MIT Press, 1991, pp. 141–201.

Engle, Robert F., and C. W. Granger. “Cointegration and Error-Correction: Representation, Estimation, and Testing,” *Econometrica,* vol. 55 (March 1987), pp. 251–76.

Engsted, Tom. “Does the Long-Term Interest Rate Predict Future Inflation? A Multicountry Analysis,” *The Review of Economics and Statistics,* vol. 77 (February 1995), pp. 42–54.

Goodfriend, Marvin. “Interest Rate Policy and the Inflation Scare Problem: 1979 to 1992,” Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 79 (Winter 1993), pp. 1–24.

Hall, Alastair. “Testing for a Unit Root in Time Series with Pretest Data-Based Model Selection,” *Journal of Business and Economic Statistics,* vol. 12 (October 1994), pp. 461–70.

Hodrick, Robert J., and Edward C. Prescott. “Postwar U.S. Business Cycles: An Empirical Investigation,” *Journal of Money, Credit, and Banking,* vol. 29 (February 1997), pp. 1–16.

Johansen, Soren, and Katarina Juselius. “Identification of the Long-Run and the Short-Run Structure: An Application to the ISLM Model,” *Journal of Econometrics,* vol. 63 (July 1994), pp. 7–36.

__________. “Maximum Likelihood Estimation and Inference on Cointegration—With Applications to the Demand for Money,” *Oxford Bulletin of Economics and Statistics,* vol. 52 (May 1990), pp. 169–210.

Jorion, Philippe, and Frederick Mishkin. “A Multicountry Comparison of Term Structure Forecasts at Long Horizons,” *Journal of Financial Economics,* vol. 29 (March 1991), pp. 59–80.

Kwiatkowski, Denis, Peter C. B. Phillips, Peter Schmidt, and Yongcheol Shin. “Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?” *Journal of Econometrics,* vol. 54 (October–December 1992), pp. 159–78.

McCallum, Bennett T. “Monetary Policy and the Term Structure of Interest Rates.” Manuscript. Carnegie-Mellon University, June 1994.

Mehra, Yash P. “An Error-Correction Model of the Long-Term Bond Rate,” Federal Reserve Bank of Richmond *Economic Review,* vol. 80 (Fall 1994), pp. 49–68.

Miller, Stephen M. “Monetary Dynamics: An Application of Cointegration and Error-Correction Modeling,” *Journal of Money, Credit, and Banking,* vol. 23 (May 1991), pp. 139–54.

Mishkin, Frederic S. “A Multicountry Study of the Information in the Shorter Maturity Term Structure about Future Inflation,” *Journal of International Money and Finance,* vol. 10 (March 1991), pp. 2–22.

__________. “The Information in the Longer Maturity Term Structure about Future Inflation,” *Quarterly Journal of Economics,* vol. 105 (August 1990a), pp. 815–28.

__________. “What Does the Term Structure Tell Us about Future Inflation?” *Journal of Monetary Economics,* vol. 25 (January 1990b), pp. 77–95.

Plosser, Charles I., and K. Geert Rouwenhorst. “International Term Structure and Real Economic Growth,” *Journal of Monetary Economics,* vol. 33 (February 1994), pp. 133–55.

Rudebusch, Glenn D. "Federal Reserve Interest Rate Targeting, Rational Expectations, and the Term Structure," *Journal of Monetary Economics,* vol. 35 (April 1995), pp. 245–74.

Sims, Christopher A. "Macroeconomics and Reality," *Econometrica,* vol. 48 (January 1980), pp. 1–49.

Stock, James H., and Mark W. Watson. "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica,* vol. 61 (July 1993), pp. 783–820.

Inventory Investment and the Business Cycle

Andreas Hornstein

When reporting on the current state of the economy, the business press gives considerable attention to changes in inventory investment. The reason for the media attention appears to be related to three issues. First, changes in inventory investment apparently account for a substantial fraction of changes in gross domestic product (GDP). Second, current changes in inventory investment are assumed to convey useful information about the near-term future of the economy. Third, there is a view that the inherent dynamics of inventory investment are destabilizing the economy. In this article I review some of the empirical regularities of inventory investment over the business cycle taking the first issue as a starting point.[1] The empirical regularities I choose to study are to some extent determined by particular theories of inventory investment, but any theory of inventory investment should be consistent with these regularities.

The argument that inventory investment is important for the business cycle is often based on the close relationship between changes in inventory investment and GDP during recessions. For example, Blinder (1981) and Blinder and Maccini (1991) argue that, in a typical U.S. recession, declining inventory investment accounts for most of the decline in GDP. In support of this claim, Table 1 documents the peak-to-trough decline of GDP and inventory investment during postwar U.S. recessions. This same peak-to-trough decline is apparent

I would like to thank Michelle Kezar and Andrew Olmem for research assistance and Mike Dotsey, Bob Hetzel, Tom Humphrey, and Alex Wolman for helpful comments. Any opinions expressed are mine and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] When appropriate, I will make some comments on the second issue, namely, whether inventory investment is useful for forecasting GDP. In the conclusion, I will remark briefly on the third issue, namely, whether inventory investment is destabilizing the economy.

Table 1 GDP and Inventory Investment in Postwar Recessions

GDP Peak to Trough	Change in GDP	Change in Inventory Investment
1948:4 to 1949:4	−24.4	−33.3
1953:2 to 1954:2	−48.8	−20.0
1957:3 to 1958:1	−81.4	−18.4
1960:3 to 1960:4	−40.7	−47.9
1969:3 to 1970:4	−20.3	−38.4
1973:4 to 1975:1	−146.2	−77.0
1980:1 to 1980:3	−116.7	−52.7
1981:3 to 1982:3	−140.9	−43.4
1990:2 to 1991:1	−124.1	−60.7

Notes: "Dates correspond to the largest peak-to-trough decline in GDP associated with each postwar recession. Each date is within one quarter of the quarter containing the peak or trough month as defined by the National Bureau of Economic Research."

Source: Fitzgerald 1997, Table 1, p. 12.

during the 1980–97 period, as shown in Figure 1.[2] Figure 1 also shows that inventory investment is a very noisy time series. During that period, inventory investment not only declines dramatically during recessions, it also declines substantially during expansions. For example, in the expansion years 1986 and 1988, inventory investment declined by almost as much as it did during the 1990 recession. This experience suggests that, while the observation that during recessions declining inventory investment accounts for much of declining GDP is interesting, it might not be very useful when we want to evaluate the role of inventory investment over the complete business cycle.[3]

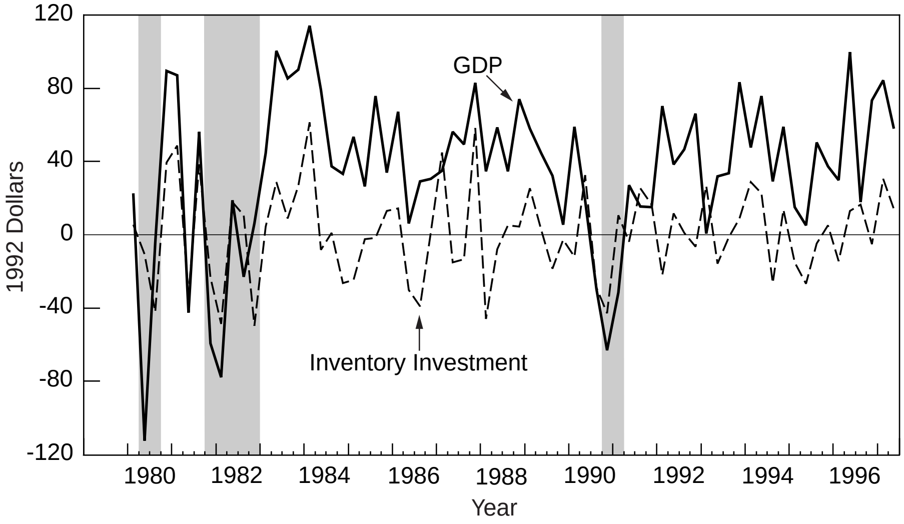
Rather than study the behavior of inventory investment for a particular phase of the business cycle, I choose to document the stylized facts of such investment over the entire business cycle using standard methods.[4] A stylized

[2] Shaded areas in Figure 1 represent National Bureau of Economic Research (NBER) recessions. We display only the last two NBER business cycles so that the graph is not overcrowded. The behavior of inventory investment and GDP from 1980 to the present is not qualitatively different from the earlier part of the postwar period.

[3] Because inventory investment is such a noisy time series, it is also unlikely that it contains useful information to forecast GDP growth. There are two additional pieces of evidence which suggest that inventory investment is not a particularly good predictor of future GDP growth. First, the Conference Board (1997) does not include inventory investment in its widely distributed list of leading economic indicators. Only the inventory/sales ratio is included and then as a lagging indicator. Second, if we forecast GDP growth using lagged GDP growth alone, we do better than if we include lagged GDP growth and changes in lagged inventory investment. That is, we do better in the sense that the first procedure has a lower mean squared forecast error.

[4] For the most part I review earlier work on inventory investment by Blinder (1981) and Blinder and Maccini (1991), using a different method to extract the business cycle components of time series.

Figure 1 Changes in GDP and Inventory Investment



fact is an observed empirical regularity between particular variables, which is of interest because economic theory predicts a certain pattern for it. One cannot look for stylized facts without the guidance of economic theory, but economic theory is also developed from the stylized facts uncovered. Since inventory investment ΔN is the difference between production Y and sales X, that is $\Delta N = Y - X$, the stylized facts discussed involve the behavior of these three interrelated variables.

1. MODELS OF INVENTORY INVESTMENT

The two leading economic theories of inventory investment are the production-smoothing model and the (S, s) inventory model. Both theories start with a single firm that solves a dynamic constrained-profit-maximization problem using inventory investment as one of the firm's decision variables.[5] The theories differ in how the implications for inventory investment, derived for an individual firm, are applied to the study of aggregate inventory investment.[6]

[5] These theories differ from the early behavioral models of inventory investment that are not explicitly based on fully specified dynamic optimization problems (Metzler 1941).

[6] For an extensive survey of theories of inventory investment, see Blinder and Maccini (1991).

A simple production-smoothing model starts with the assumption that a firm's production is subject to increasing marginal cost and that sales are exogenous. If the firm's sales are changing over time but its marginal cost schedule is constant, then the firm minimizes cost by smoothing production, and it reduces (increases) inventories whenever sales exceed (fall short of) production. Thus production is less volatile than sales, and inventory investment and sales tend to be negatively correlated. A firm with increasing marginal cost wants to use inventories to smooth production regardless of whether or not the changes in demand are foreseen. If demand changes randomly and the firm has to decide on current production before it knows what current demand is, the firm also uses inventories as a buffer stock and accordingly reduces (increases) inventory stocks whenever demand is unexpectedly high (low). This buffer-stock motive then reinforces the negative correlation between inventory investment and sales.

The previous argument assumes that the firm faces only demand variations. If, on the other hand, the firm predominantly faces supply shocks in the form of a changing marginal cost schedule, then the implications for inventory investment, production, and sales are very different. In order to minimize costs, the firm now increases (decreases) production and accumulates (reduces) inventories during times when marginal cost is low (high). Thus production is more volatile than sales, and inventory investment and production tend to be positively correlated.

So far the production-smoothing model described above applies to the behavior of an individual firm, rather than the behavior of aggregate variables. To understand the aggregate variables, one often uses the concept of a representative agent and interprets the behavior of aggregate variables in terms of the behavior of a large number of identical individual decision units. The simple production-smoothing model then predicts that production will be more (less) volatile than sales if supply shocks are more (less) important than demand shocks.[7]

A simple (S, s) inventory model assumes that the seller of a good does not himself produce the good. Instead, the seller orders the good from some producer and incurs a fixed cost when he places the order. Suppose that the marginal cost of ordering one more unit of the good is constant and that sales are exogenous. A seller who chooses the order size that minimizes total cost faces the following tradeoff. On the one hand, increasing the order size reduces the average or per-unit order cost because it spreads the fixed cost over more units of the good. On the other hand, an increased order size means that the seller forgoes additional interest income on the funds that have been used to finance the larger order. Given the optimal order size, the seller places an order

[7] Further work has studied the effects of serial correlation in demand shocks, stock-out avoidance, etc. Again, for a survey on this work, see Blinder and Maccini (1991) or Fitzgerald (1997).

whenever the inventory falls below a critical lower level s and the order brings inventories up to the higher level S. After that, sales reduce the inventory until the critical lower level s is reached again. If orders equal production, then production will be more volatile than sales. The relationship between sales and inventory investment is unclear.

Like the production-smoothing model, the (S, s) inventory model applies to an individual decision unit. Unlike the production-smoothing model, however, the notion of a representative agent cannot be used in order to understand the behavior of aggregate variables. The problem is that in the (S, s) inventory model, a firm's behavior is characterized by long periods of inactivity interrupted by short bursts of activity. While one may observe such discontinuous behavior for individual decision units, one does not observe it for aggregate variables. For this theory, aggregation has to be studied explicitly, and aggregate variables will not necessarily behave the same way as do the corresponding variables of individual decision units. In particular, in a model where individual firms follow (S, s) inventory policies, one cannot a priori say whether aggregate production or aggregate sales is more volatile or how aggregate inventory investment is correlated with aggregate sales. Fisher and Hornstein (1997) study the effects of technology and preference shocks on aggregate production, sales, and inventory investment in a general equilibrium model with a trade sector where individual firms use (S, s) inventory policies.[8] They find that for both types of shocks (1) production is more volatile than sales, and (2) inventory investment tends to be positively correlated with sales. They also find that preference and technology shocks differ in their effect on retail-price markups. In particular, retail-price markups are procyclical for technology shocks and countercyclical for preference shocks.

2. CYCLICAL COMPONENTS OF INVENTORY INVESTMENT

Up to this point I have used economic theory to identify potential stylized facts pertaining to inventory investment. To further evaluate the role of inventory investment over the business cycle, I will need an operational definition to identify business cycle movements in the data. Usually business cycles are identified with recurring expansions and contractions in economic activity that occur simultaneously over a wide range of sectors. Burns and Mitchell (1946, p. 3) state that

[8] In a general equilibrium model, shocks cannot be unambiguously classified as demand or supply disturbances. Usually shocks that affect the production technology are interpreted as supply disturbances, and shocks that affect the preferences of agents are interpreted as demand disturbances.

> . . . a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle; this sequence of changes is recurrent but not periodic; in duration business cycles vary from more than one year to ten or twelve years; they are not divisible into shorter cycles of similar character with amplitudes approximating their own.

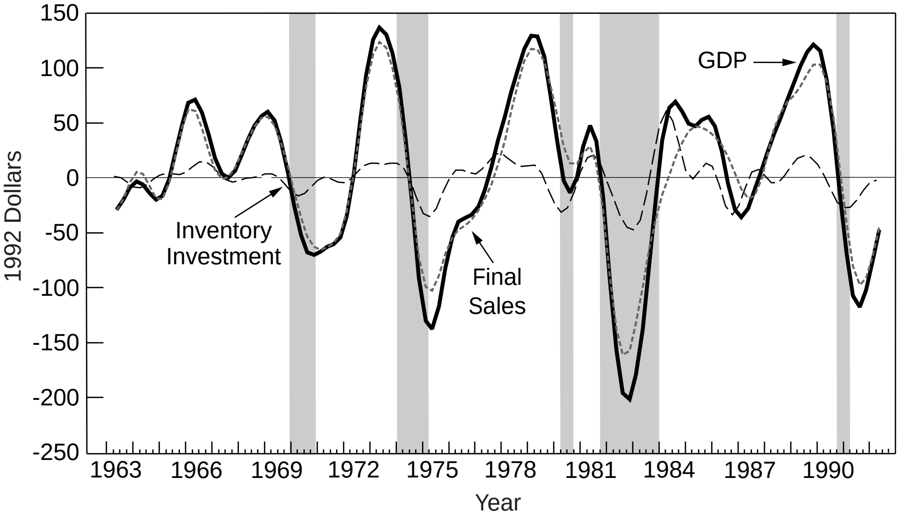
The business cycle is thus different from long-term trend and short-term irregular movements in the economy. Yet many of the economic variables are growing over time (GDP, sales) or are very erratic (inventory investment). Furthermore, since inventories can serve as a buffer stock to compensate for short-term movements in demand or supply, one may want to study the business cycle component and the short-term irregular component separately.

Bandpass filters, which essentially are moving averages, separate the time series of a variable into components with different periodicities (see Baxter and King [1995]).[9] Using this method, I construct the business cycle components of inventory investment, GDP, and final sales as displayed in Figure 2. First, one can see that, for business cycle movements, inventory investment contributes only a small part to GDP volatility. Second, GDP is more volatile than final sales, and final sales and inventory investment tend to increase and decrease together. Figure 3 plots the irregular components of inventory investment. Consistent with Figure 1's depiction of changes in GDP and inventory investment, Figure 3 shows that inventory investment accounts for a substantial fraction of the short-term volatility of GDP.

The two inventory models discussed above capture different features of the inventory holding problem. In any one sector of the economy, one of the features will play a bigger role. For example, when firms in the manufacturing sector choose the size of their finished goods inventories, the production-smoothing model seems to be more appropriate. But when firms in the trade sector make their order decisions, or firms in the manufacturing sector decide on the size of their material inventories, the (S, s) inventory model seems to be more appropriate. Our study of disaggregated data shows two things. On the one hand, it is difficult to attribute aggregate inventory investment volatility to particular sectors because inventory investment moves much the same in each sector. On the other hand, we find that although important features of the inventory holding problem differ systematically across sectors, the properties of inventory investment, production, and sales are remarkably similar across sectors; for business cycle movements, production is more volatile than sales, and inventory investment and sales are positively correlated. The only

[9] In Appendix A I describe the basic idea underlying the decomposition of a time series using bandpass filters.

Figure 2 Business Cycle Components of GDP, Final Sales, and Inventory Investment



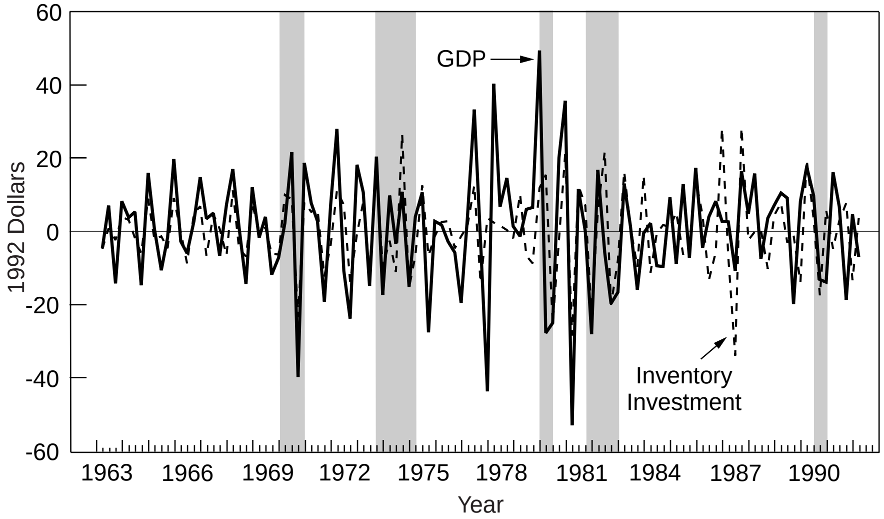
exception concerns the behavior of retail-price markups, which are not consistently procyclical or countercyclical across sectors.

3. STYLIZED FACTS OF INVENTORY INVESTMENT

The organization of the stylized facts is suggested by the predictions of the two basic models of inventory investment, the production-smoothing and the (S, s) inventory models. First, I document the behavior of aggregate variables, GDP, final sales, and inventory investment. Then I decompose aggregate inventory investment into sectors according to whether it is more likely that inventory decisions are influenced by the production-smoothing motive or the fixed order cost motive. Next, I examine the relative volatilities of production and sales and the correlation between inventory investment and sales. Finally, I study the behavior of the retail price index relative to the producer price index, that is, the retail-price markup.[10]

[10] Appendix B describes the time series used and how the business cycle and irregular component of each time series is constructed.

Figure 3 Irregular Components of GDP and Inventory Investment



Inventory Investment at the Aggregate Level

For quarterly changes in aggregate values of GDP, final sales, and inventory investment, Table 2 quantifies some of the observations made earlier in the introduction using Figure 1. The table shows that GDP is more volatile than final sales and that final sales and inventory investment are essentially uncorrelated (the correlation coefficient is 0.01). This observation is consistent with properties of the simple production-smoothing model and the (S, s) model, and for the former it implies that supply shocks must be relatively more important than demand shocks. Note that, consistent with conventional wisdom, changes in inventory investment account for a substantial part of the variance of changes in GDP, about 30 percent.[11]

When distinguishing between the business cycle and irregular components of a time series, one sees that GDP is more volatile than final sales for both components, whereas inventory investment is positively correlated with final sales for the business cycle component but negatively correlated for the irregular component. The correlation coefficients are, respectively, 0.54 and -0.2.

[11] Given that the variance of output is the sum of the variance of sales, the variance of inventory investment, and the covariance of sales and inventory investment, we can attribute output volatility to sales and inventory investment volatility because of the low sales/inventory investment correlation.

Table 2 GDP, Final Sales, and Inventory Investment

	First Difference of Levels		Business Cycle Component		Irregular Component	
NIA Component	**Variance**	**Percent**	**Variance**	**Percent**	**Variance**	**Percent**
GDP	1534.99		5479.39		255.41	
Final sales	1085.94	70.7	3961.99	72.3	190.19	74.5
Inventory investment	439.11	28.6	315.63	5.8	122.88	48.1
Covariance of final sales and inventory investment	9.94	0.6	1197.46	21.9	−57.92	−22.7

Again, the fact that production is more volatile than sales is consistent with the simple production-smoothing model when supply shocks dominate demand shocks, but now the model cannot easily account for the comovement of sales and inventory investment. In particular, the model does not predict the strong positive correlation between inventory investment and sales for the business cycle component. Moreover, the weak negative correlation for the irregular component seems to indicate that, over the short term, demand shocks dominate supply shocks, and inventories are used as a buffer stock. The (S, s) model is consistent with the properties of the business cycle component but does not predict the negative correlation between inventory investment and sales for the irregular component.

For the business cycle component, it is difficult to attribute GDP volatility to either sales or investment volatility since there is a strong positive correlation between these two components of GDP. Furthermore, relative to GDP, inventory investment is much less volatile for the business cycle component than it is for the irregular component. Inventory investment variance represents only 6 percent of GDP variance for the business cycle component but 50 percent of GDP variance for the irregular component. Thus it appears that inventory investment is less important for GDP volatility over the business cycle than it is for short-term fluctuations.[12]

[12] Since the calculation of changes in a variable, that is, its first differences, and the calculation of business cycle and irregular components of the same variable represent different data transformations, it is hardly surprising that they lead to different results. These observations are not inconsistent; they only reflect different properties of the data and the transformation used.

Appendix A shows how the business cycle and irregular components of a variable represent the frequency components of that same variable that fall within a particular frequency band and where each frequency receives the same weight. Calculating changes in a variable, that is, first differences, is another data transformation that includes all frequencies but gives more weight to higher frequencies (Baxter and King 1995). Thus first differences emphasize components with short periodicity relative to components with long periodicity, and therefore the properties of a first-differenced variable are more closely related to the properties of the irregular component than to the properties of the business cycle component of that variable.

Disaggregating Inventory Investment

Most of the results from the study of aggregate variables also apply when aggregate production, sales, and inventory investment are disaggregated into their sectoral components: manufacturing and trade. It is useful to study the sectoral components of inventory investment because the production-smoothing and the (S, s) models seem to be more or less appropriate for different types of inventories. For example, the production-smoothing model appears to be more appropriate for finished goods inventories in the manufacturing sector, whereas the fixed order cost model appears to be more appropriate for material inventories in the manufacturing sector and inventories in the wholesale and retail trade sector. This suggests that one should focus attention on the theory of inventory investment that is most appropriate for the sector that contributes the most to aggregate inventory investment volatility. Unfortunately, it turns out to be difficult to attribute aggregate inventory investment volatility to individual sectors.

Table 3 shows the variance of the components of total inventory investment: manufacturing and trade inventories. Over the business cycle, finished goods inventories in the manufacturing sector account for only about 10 percent of the total variance of inventory investment. On the other hand, inventories in the trade sectors and materials in the manufacturing sector account for about a quarter of total inventory volatility. Note that, although inventory investment in the trade sector accounts on average for more than half of total inventory investment, it accounts for less than 20 percent of the volatility of inventory investment over the business cycle. Any attempt to attribute the variance of total inventory investment to particular components, however, meets with limited success because more than half of total inventory volatility is due to the comovement of inventory investment components. In particular, within the manufacturing sector about 50 percent of total volatility is due to the comovement of finished goods, goods-in-process, and materials inventories. For total inventory investment volatility, about 40 percent of total volatility is due to the comovement of the individual components: manufacturing, retail, and wholesale trade. This observation is the main difference between our results and those of Blinder and Maccini (1991). They find that, of manufacturing inventory volatility, only 25 percent is due to the comovement of finished goods, materials, and goods-in-process inventories. And for total inventory investment, only 20 percent is due to covariance terms.

Blinder and Maccini (1991) define the business cycle component of a time series as fluctuations around a linear trend. In effect, their definition of the business cycle eliminates long-run growth components but not the irregular component, or short periodicity movements, from consideration. Table 3 reveals as much. It shows that these high-frequency movements are not highly correlated across sectors; that is, the results of Blinder and Maccini (1991) represent

Table 3 Variance Decomposition of Inventory Investment

Inventory Component	Percent of Total Investment	Business Cycle Component Variance	Business Cycle Component Percent	Irregular Component Variance	Irregular Component Percent
Manufacturing and trade		2.491		4.753	
Manufacturing	43.5	1.035	41.6	1.538	32.4
Finished goods	15.2	0.107	10.3	0.443	28.4
Goods-in-process	14.7	0.249	23.9	0.487	31.3
Materials and supplies	13.6	0.150	14.4	0.556	35.7
Covariance terms		0.535	51.4	0.072	4.6
Wholesale trade	26.5	0.172	6.9	0.924	19.4
Retail trade	30.0	0.312	12.5	1.990	41.9
Covariance terms		0.973	39.1	0.301	6.3

a mixture of the properties of business cycle and irregular components. Also, for the irregular component, inventory investment in the trade sector accounts for a much bigger share of overall inventory investment variance.

Production, Sales, and Inventory Investment at the Sectoral Level

The behavior of production, sales, and inventory investment is remarkably similar in the different sectors. In all sectors, production tends to be more volatile than sales, substantially so in the retail and wholesale trade sector. This is true for both the business cycle components and the irregular components (see Table 4), thus confirming Blinder and Maccini's (1991) results. Note also that over the business cycle, the durable goods sectors are much more volatile than the nondurable goods sectors. This is consistent with other work on sectorally disaggregated data (Hornstein and Praschnik 1997).

Table 5a documents the pattern of comovement between inventory investment and sales for the business cycle components. Over the business cycle, inventory investment and sales are positively correlated. What is of interest is that there are different patterns of lead-lag relationships between inventory investment and sales in the various sectors. For example, in the manufacturing sector, inventory investment in nondurable manufacturing is essentially uncorrelated with sales, but such investment in durable manufacturing bears a strong contemporaneous correlation with sales. In the wholesale trade sector, inventory investment in the durable sector is also contemporaneous with sales, but inventory investment in the nondurable sector leads sales by three months. Furthermore, in retail trade, inventory investment leads sales by three months,

Table 4 Relative Variance of Production *Y* and Sales *X*

	Business Cycle Component			Irregular Component		
Sector	Var (*Y*)	Var (*X*)	$\frac{Var(Y)}{Var(X)}$	Var (*Y*)	Var (*X*)	$\frac{Var(Y)}{Var(X)}$
Manufacturing	43.2	41.9	1.03	6.84	6.66	1.03
	47.9		1.14	7.36		1.11
Durables	20.0	19.2	1.04	3.59	3.37	1.07
	23.2		1.21	3.92		1.16
Nondurables	4.9	4.8	1.02	1.31	1.18	1.11
	4.9		1.01	1.38		1.17
Wholesale trade	11.49	10.3	1.12	1.48	1.74	1.11
Durables	6.0	5.0	1.19	2.14	0.50	1.18
Nondurables	1.6	1.6	1.03	1.45	0.90	1.01
Retail trade	9.9	8.1	1.23	1.64	1.78	1.21
Durables	3.9	3.1	1.28	1.43	1.08	1.26
Nondurables	1.5	1.2	1.18	2.35	0.27	1.17

Note: For the manufacturing sector, the first row refers to the narrow inventory definition (finished goods inventories only) and the second row refers to the broad inventory definition (finished goods and goods-in-process inventories).

Table 5a Comovement of Inventory Investment ΔN and Sales *X* for Business Cycle Components

	Correlation coefficient for X_t and ΔN_{t+s}, where $s =$								
	−4	−3	−2	−1	0	1	2	3	4
Manufacturing	0.09	0.15	0.20	0.25	0.30	0.33	0.36	0.38	0.38
	0.47	0.51	0.54	0.57	0.60	0.61	0.61	0.60	0.58
Durables	0.23	0.28	0.32	0.36	0.39	0.41	0.43	0.43	0.43
	0.59	0.62	0.65	0.66	0.67	0.67	0.66	0.64	0.61
Nondurables	−0.13	−0.08	−0.03	0.01	0.05	0.08	0.10	0.12	0.12
	−0.11	−0.08	−0.04	−0.01	0.01	0.03	0.04	0.04	0.03
Wholesale trade	0.43	0.44	0.44	0.43	0.39	0.34	0.29	0.23	0.16
Durables	0.49	0.51	0.53	0.53	0.52	0.50	0.47	0.43	0.37
Nondurables	0.19	0.17	0.14	0.09	0.03	−0.04	−0.12	−0.20	−0.27
Retail trade	0.55	0.54	0.53	0.50	0.47	0.43	0.37	0.31	0.24
Durables	0.54	0.54	0.53	0.51	0.48	0.44	0.38	0.32	0.25
Nondurables	0.35	0.34	0.32	0.30	0.28	0.25	0.22	0.18	0.14

Note: See Note to Table 4.

Table 5b Comovement of Sales X and Inventory Investment ΔN for Irregular Components

	Correlation coefficient for X_t and ΔN_{t+s}, where $s =$								
	−4	**−3**	**−2**	**−1**	**0**	**1**	**2**	**3**	**4**
Manufacturing	−0.14	−0.23	−0.15	−0.06	−0.08	0.18	0.13	0.15	0.20
	−0.11	−0.14	−0.06	0.04	−0.06	0.18	0.15	0.09	0.11
Durables	−0.11	−0.13	−0.07	0.02	0.04	0.18	0.09	0.03	−0.02
	−0.06	−0.07	−0.02	0.09	−0.02	0.17	0.14	0.02	0.01
Nondurables	−0.16	−0.17	−0.09	−0.05	−0.10	0.10	0.10	0.14	0.21
	−0.17	−0.13	−0.04	0.01	−0.08	0.10	0.11	0.10	0.16
Wholesale trade	−0.01	0.08	0.01	0.11	−0.04	−0.06	−0.06	−0.09	−0.00
Durables	−0.00	0.04	−0.13	0.04	−0.09	−0.03	0.08	0.08	0.01
Nondurables	−0.02	0.11	0.11	0.11	0.02	−0.18	−0.13	−0.05	−0.06
Retail trade	−0.01	−0.05	−0.03	−0.04	−0.23	0.05	0.18	0.07	0.06
Durables	−0.02	−0.03	0.06	−0.07	−0.34	0.01	0.15	0.05	0.07
Nondurables	−0.02	0.07	−0.15	0.02	−0.10	0.02	0.11	−0.05	0.04

Note: See Note to Table 4.

both for durable and nondurable goods.[13] This observation might be useful to differentiate between models of inventory investment across sectors.

For the irregular component, inventory investment and sales are essentially uncorrelated, with a tendency towards negative correlations (see Table 5b). In particular, for the retail trade sector, sales and inventory investment are somewhat negatively correlated. As for aggregate data, it appears as if inventory stocks are used to buffer unforeseen short-term fluctuations in sales.

The Cyclical Behavior of the Retail-Price Markup

One last variable, the retail-price markup, is of interest because extensions of simple (S, s) inventory models suggest that sellers have some control over the prices they set. Such control means that decisions on inventory investment, sales, and prices are interrelated. Fisher and Hornstein (1997) describe such an (S, s) inventory model for the retail sector. They argue that the cyclical behavior of the retail-price markup depends on whether supply or demand shocks are more important for a market. In particular, their model predicts that if productivity shocks to the suppliers of the retailers are predominant, then the retail-price markup should be positively correlated with sales. On the other

[13] The fact that inventory investment leads sales over the business cycle does not mean that inventory investment can be used to predict future sales. The reason is simply that the business cycle component of a variable represents a moving average of past and future values of the variable.

Table 6 Comovement of Sales X and Retail Markups M at Business Cycle Frequencies

	Correlation coefficient for X_t and M_{t+s}, where $s =$								
	−4	−3	−2	−1	0	1	2	3	4
All retail	0.61	0.60	0.58	0.56	0.54	0.51	0.48	0.45	0.41
Durable goods	0.28	0.27	0.25	0.23	0.21	0.18	0.15	0.11	0.07
Autos	0.03	0.01	−0.00	−0.02	−0.03	−0.05	−0.06	−0.06	−0.07
Furniture	0.13	0.09	0.04	−0.02	−0.07	−0.12	−0.18	−0.23	−0.27
Building mat.	−0.66	−0.70	−0.73	−0.75	−0.76	−0.76	−0.75	−0.74	−0.71
Nondurable	0.50	0.53	0.56	0.58	0.60	0.61	0.62	0.62	0.61
Food	0.37	0.34	0.30	0.26	0.21	0.16	0.11	0.07	0.04
Apparel	0.17	0.09	0.02	−0.06	−0.14	−0.21	−0.27	−0.33	−0.38
Others	0.48	0.42	0.35	0.28	0.21	0.15	0.09	0.03	−0.03

hand, if shocks to demand for the retailer's product are predominant, then the retail-price markup should be negatively correlated with sales.

The comovement over the business cycle between retail-price markups and sales for a selected number of products is documented in Table 6. Apparently there is no strong consistent pattern in the data. For nondurable goods, with the exception of apparels, the markup tends to be positively correlated with sales, and for durable goods the markup tends to be negatively correlated with sales. Of interest is the absence of any strong comovement for cars. One note of caution: the lack of correlation between markup and sales should not be taken as evidence for inflexible prices. The markup is defined as the ratio of retail prices to producer prices, both of which tend to be strongly correlated with sales over the business cycle. In particular, the retail price is negatively correlated with sales for all goods, and, with the exception of building materials, the producer price indexes are negatively correlated with sales.

Finally, we have not presented results for the irregular component because for these frequencies the markup is essentially uncorrelated with sales. In this case the markup is uncorrelated with sales, because both the retail price and the producer price index are uncorrelated with sales.

4. CONCLUSION

The description of the data above suggests that it is important to distinguish between the irregular and the business cycle components of inventory investment, production, and sales. Bearing this in mind, the findings can be summarized as follows. First, inventory investment fluctuations are not important for output fluctuations over the business cycle, but they are important for short-term output fluctuations. Second, over the business cycle, we cannot attribute total inventory investment volatility to its individual components because all components are highly correlated. Third, inventory investment is positively correlated with sales over the business cycle but tends to be uncorrelated or negatively correlated with sales for short-term fluctuations. Fourth, production tends to be more volatile than sales; this feature is common to all sectors, and it applies to business cycle and short-term fluctuations.

How well do the existing models of inventory investment match these stylized facts? The production-smoothing model is in principle consistent with the finding that production is more volatile than sales in the particular case where cost shocks are assumed to be more important than demand shocks. Essentially the production-smoothing model is used in order to say something about the relative importance of unobserved demand and supply shocks in the economy. Unfortunately, with direct observations on cost and demand shocks, the production-smoothing model often is no longer consistent with the stylized facts, given the observed relative volatility of shocks (for some recent work, see Durlauf and Maccini [1995]). Furthermore, the production-smoothing model has problems accounting for the comovement of sales and inventory investment, even if cost shocks are more volatile than demand shocks.

Less can be said about how well the (S, s) inventory framework conforms to the stylized facts because only recently has work begun that tries to incorporate this framework in quantitative general equilibrium models. For a simple general equilibrium model with (S, s) inventory policies, Fisher and Hornstein (1997) have shown that the model's quantitative implications are consistent with the stylized facts. But more work needs to be done.

Let me conclude with a remark on whether inventory investment can destabilize the economy.[14] Obviously the stylized facts reviewed in this article by themselves have nothing to say about this issue. Potential destabilization can only be addressed within some theory of inventory investment. For example, the fact that production appears to be more volatile than sales does not mean that, because of inventory investment, production is excessively volatile. If one believes that the production-smoothing model is a useful representation of the economy, then at least for a firm, this outcome is optimal if marginal cost varies over time.

Most inventory investment models, with few exceptions, are partial equilibrium in nature; that is, they describe the behavior of a firm/industry and take the behavior of the rest of the economy as given. A complete analysis of the role of inventory investment requires that the particular inventory investment model is embedded in a general equilibrium model in order to study how inventory investment affects the rest of the economy and vice versa. It is not clear that

[14] This is a well-known property of inventories in the traditional inventory-accelerator models (Metzler 1941).

inventory investment will be destabilizing in such a general equilibrium model or even what such destabilization means. One possible interpretation is that, with inventories, the equilibrium of an economy is no longer determinate. In this case, one could construct particular equilibria where output fluctuates even though the fundamentals of the economy do not change at all; however, such work remains to be done.[15]

APPENDIX A: THE CYCLICAL COMPONENTS OF A TIME SERIES

The decomposition of a time series into business cycle and irregular components using a bandpass filter is a statistical method based on the frequency domain analysis of time series.[16] Essentially, this method interprets a time series as the sum of a very large number of sine and cosine waves, and it isolates groups of waves within particular frequency bands. Rather than describing in detail this technique and the underlying statistical theory, I simply want to provide some insight on how it works. For an introduction to the analysis of time series in the frequency domain, see Harvey (1993) or Hamilton (1994). For a description of bandpass filters, see Baxter and King (1995).

Extracting Periodic Components from Deterministic Time Series . . .

In order to illustrate the problem, consider the following example. Define a variable Y_t as the sum of sine and cosine functions

$$Y_t = \sum_{i=1}^{3} [\alpha_i \cos(\omega_i t) + \beta_i \sin(\omega_i t)].$$

A sine (cosine) function has amplitude one and periodicity $T = 2\pi$. A function is periodic with period T if the function repeats itself every T periods.[17] For a periodic function, its frequency $1/T$ denotes how many cycles are completed within a unit of time. The transformation of the sine (cosine) function $\alpha \cos(\omega t)$ $[\beta \sin(\omega t)]$ has amplitude α (β), periodicity $T = 2\pi/\omega$, and frequency $\omega = 2\pi/T$.

[15] See Benhabib and Farmer (1997) for a survey on endogenous business cycle models.

[16] Other methods have been used to identify the business cycle components of time series, for example, stochastic trends or linear trends. One close relative of a bandpass filter for the business cycle is the Hodrick-Prescott filter, also described in Baxter and King (1995).

[17] The sine function satisfies $\sin(t + 2\pi j) = \sin(t)$ for all t and $j = \ldots, -1, 0, +1, \ldots$.

In the example, Y_t is the sum of three periodic functions and is itself periodic. Assume that the unit of time is a month and that the first component has a periodicity of 50 years ($\omega_1 = 2\pi/(12 \cdot 50)$); the second component has a periodicity of five years ($\omega_2 = 2\pi/(12 \cdot 5)$); and the third component has a periodicity of one year ($\omega_3 = 2\pi/(12 \cdot 1)$). For this example, the first component represents long-run trends (low-frequency movements); the second component represents a business cycle (medium-frequency movements); and the third component represents short-run fluctuations (high-frequency movements), like seasonal fluctuations. Suppose there is a finite number of observations on Y_t as shown in Figure A1: How can the three different components be extracted from Y_t?[18]

An ideal bandpass filter extracts the components of a time series whose frequencies are within a given frequency band (Baxter and King 1995). This filter assigns a weight of one to all frequencies that fall within the specified band and zero weight to all frequencies outside the specified band. The ideal bandpass filter is represented by a moving average with infinitely many leads and lags, $\hat{Y}_t = \sum_{s=-\infty,...,\infty} a_s Y_{t+s}$, and the filter is defined by its weights, $\{a_s\}_{s=-\infty,...,+\infty}$, which depend on the frequency band to be extracted from the time series. Since only a finite number of observations is available, the bandpass filter has to be approximated. It turns out that an approximate bandpass filter has the same moving average representation except that the weights are truncated, $\hat{Y}_t^S = \sum_{s=-S,...,S} a_s Y_{t+s}$, and the number of leads/lags S determines the approximation quality. Because the bandpass filter is approximate, it will pass some components with frequencies outside the specified frequency band, and it will not assign all frequencies within the specified frequency band the same weight. The approximation improves with the number of leads and lags included in the moving average term.

I follow Baxter and King (1995) and identify the business cycle with periodicities between one and one-half years and six years, and for monthly (quarterly) data I use an approximation involving 36 (12) leads and lags. The irregular component (trend component) is identified with the periodicities of less than one and one-half years (more than six years). When this procedure is applied to the time series of Y_t in Figure A1, the approximate bandpass filter extracts its low-, medium-, and high-frequency components quite well.[19]

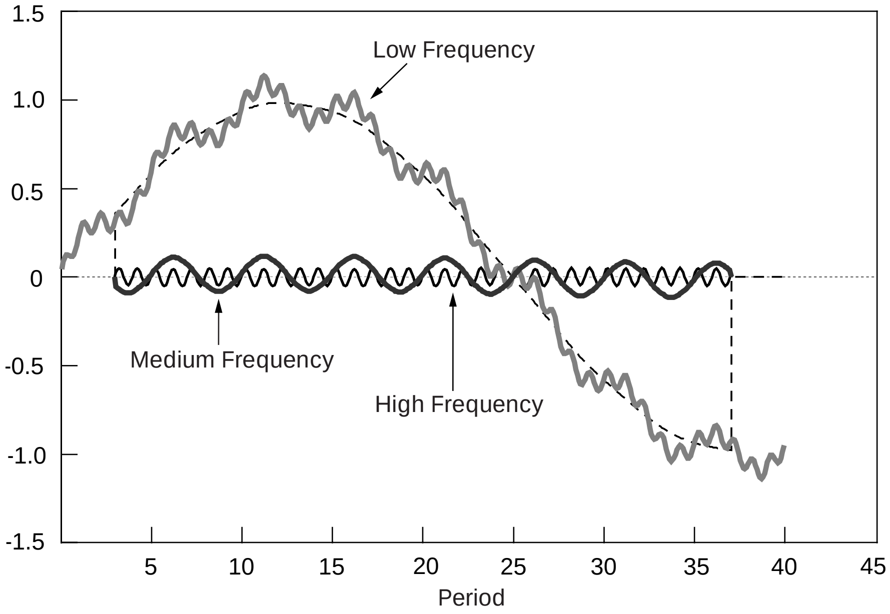
. . . and Stochastic Processes

The business cycle is not a deterministic process, which should be apparent from the graphs of GDP growth. Definitions of the business cycle, such as Burns and Mitchell's (1946, p. 3) above, recognize this fact and refer to ". . .

[18] I have set $\alpha_1 = 1$, $\alpha_2 = 0.1$, $\alpha_3 = 0.05$, and $\beta_i = 0$ for $i = 1, 2, 3$.

[19] The filtered series is not defined for the first and last S observations since the filter uses S leads and lags.

Figure A1 A Sine Function and its Cyclical Components



recurrent but not periodic . . .” movements. In particular, Lucas ([1977], 1989, p. 217) states that “. . . movements about trend . . . can be well described by a stochastically disturbed difference equation of very low order.” Yet, I have discussed the bandpass filter as a way to extract periodic components from a time series that is the sum of deterministic cycles.

This approach remains valid for the study of covariance stationary stochastic time series, because of the spectral representation theorem.[20] The theorem states that any covariance stationary time series can be written as the integral of randomly weighted sine and cosine functions

$$Y_t = \mu + \int_0^{\pi} [\alpha(\omega)\cos(\omega t + \beta(\omega)\sin(\omega t)]d\omega, \tag{1}$$

where the random variables $\alpha(\omega)$ and $\beta(\omega)$ are in a sense “mutually uncorrelated” with mean zero. The property that α and β are uncorrelated is useful because it allows us to attribute the variance of Y_t to its various components. Let

[20] A stochastic process Y_t is covariance stationary if the first and second moments of the process are time independent; that is, expected values are $E[Y_t] = \mu$ and $E[Y_t Y_{t-s}] = \rho_s$ for all t, s.

$0 \leq \omega^L < \omega^H \leq \pi$ and write the interval $[0, \pi]$ as the union of a low-frequency trend component $I^{TR} = [0, \omega^L]$, a medium-frequency business cycle component $I^{BC} = [\omega^L, \omega^H]$, and a high-frequency irregular component $I^{IR} = [\omega^H, \pi]$. The bandpass filter can be applied to the stochastic process Y_t and extract the components associated with each frequency band $I = [\omega_0, \omega_1]$. Because the sine and cosine functions at different frequencies are uncorrelated, the variance of Y_t is the sum of the variances of the nonoverlapping frequency components.

To get some idea of how a bandpass filter works for a stochastic process, consider the following example. Suppose the stochastic process Y_t is described by a low-order stochastic difference equation, in particular, Y_t is first-order autoregressive (AR(1)),

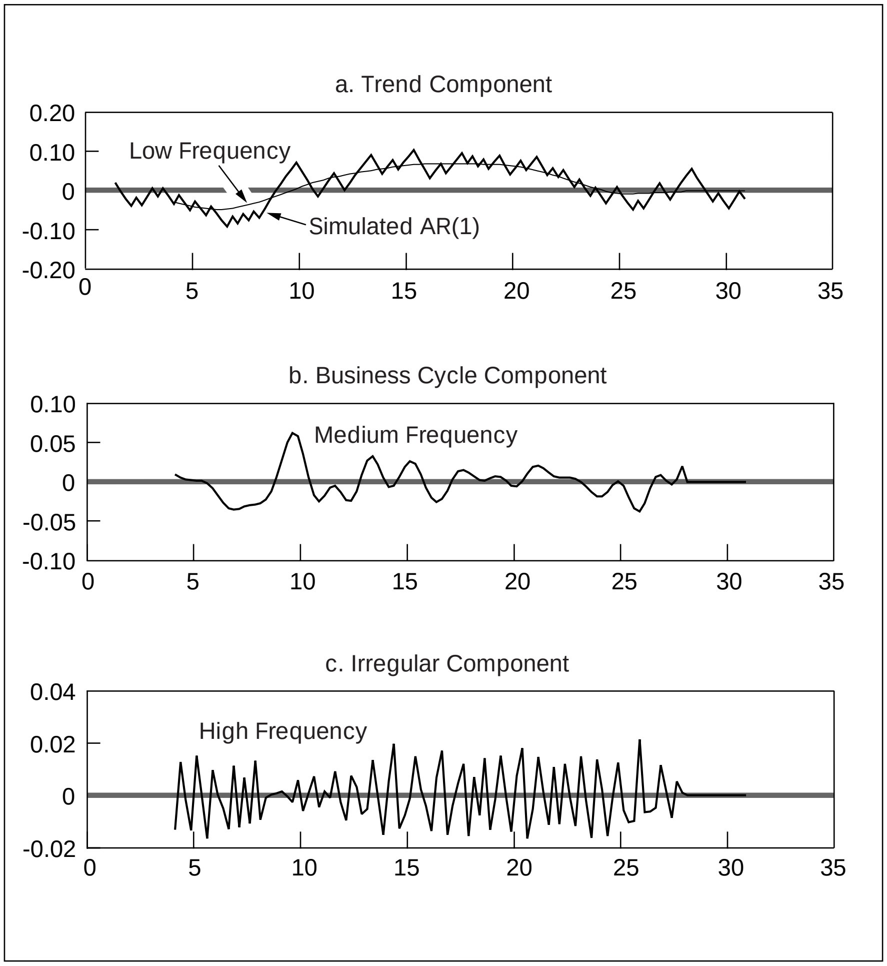
$$Y_t = 0.95 Y_{t-1} + \varepsilon_t \text{ for } t = 1, 2, \ \ldots\ ,$$

where Y_0 is given, and ε_t is an identically and independently distributed random variable that takes on values -0.1 or $+0.1$ with a probability of one-half each. Suppose that each time period represents a quarter. Make 120 independent draws of the random variable ε_t and construct a particular 30-year time path of this process $\{y_t\}_{t=0,...,30 \cdot 4}$. The result is something like Figure A2a. Recurrent but not periodic movements are clearly recognizable in the time path. As a next step apply a bandpass filter to this time path and extract the trend, business cycle, and irregular components of the time series, shown in Figure A2a-c. As is apparent from this figure, the business cycle component is quite smooth, with cycles between two and six years, whereas the irregular component has somewhat less amplitude and no particular cyclical pattern.

Finally, note that the spectral decomposition theorem applies to covariance stationary stochastic processes. In particular, this means that there should be no trend in the stochastic process; that is, the mean of the random variable Y_t should not change over time. As pointed out above, most of the economic time series do have a trend. In this context it is useful to know that a bandpass filter which excludes components with zero frequency, that is, infinite periodicity, also removes any linear and quadratic trend from a time series (Baxter and King 1995).[21] Thus a bandpass filter that isolates the business cycle and irregular components also eliminates linear and quadratic trends.

[21] It also removes any components that are integrated of order one or two.

Figure A2 AR(1) Stochastic Process and its Cyclical Components



APPENDIX B: THE DATA

The data used in this article are taken from DRI U.S. Central Database. For the study of aggregate inventory investment, I use quarterly data from 1960:1 to 1995:4 on GDP and the change in business inventories. Both series are in billions of chained (1992) dollars, seasonally adjusted. For the study of disaggregated inventory investment, I use monthly data from January 1960 to December 1996 for manufacturing and trade sales and inventories. All series are billions of chained (1992) dollars, seasonally adjusted. Inventories are end of period.

Production is defined as sales plus inventory investment. For the manufacturing sector I follow Blinder and Maccini (1991) and consider two definitions of output. For the narrow definition of output, I use only inventory investment in finished goods, and for the broad definition of output, I include inventory investment in goods-in-process as well.

The quantity index for a variable is usually obtained by deflating the nominal values with a price index. The quantity indexes for sales and inventories are not directly comparable because they are measured in different units. In particular, nominal sales are deflated with a "market" price index, while inventories are deflated with a "cost" price index. Since production in a sector is defined as the sum of sales and inventory investment, either inventories or sales have to be adjusted. For constant dollar quantity indexes, West (1983) suggests rescaling the inventory series using the base-period ratio of (business receipts)/(costs of goods sold) from corporate income tax returns. I follow West even though his procedure is not quite appropriate for my data set: I use a chain-type quantity index rather than the constant dollar quantity index that West uses. It does not appear as if my decision to follow West significantly affects the results. Since the scale factor is constant, the effects of a particular choice for the scale factor are limited to the properties of production relative to other variables. However, here I get similar results as Blinder and Maccini (1991).

For the study of the relative prices of retail goods to producer goods, I use monthly data from January 1967 to December 1996 for retail sales, implicit price deflators for retail sales, and the producer price index. All series are seasonally adjusted; retail sales are in billions of chained (1992) dollars. The commodity categories for retail data and producer price data are not the same, and I follow Blinder (1981) in the way the categories are linked:

Commodity	**Retail Sales/Prices**	**Producer Price Index**
Durable goods	Total durable goods	Durable goods
Cars	Automotive dealers	Passenger cars
Furniture	Furniture and audio video group	Furniture and household durables
Building materials	Building materials group	Lumber and wood products
Nondurable goods	Total nondurable goods	Consumer nondurables (less food)
Food	Food group	Processed foods and feeds
Apparel	Apparel group	Textile products and apparel
Other nondurable goods	Other nondurable goods	Consumer nondurables (less food)

The business cycle (irregular) component of a time series x_t is calculated as follows. Because production, sales, and inventory stocks are characterized by geometric growth, that is, a log-linear trend, I start out with the log transformation of the variables. First, we extract the business cycle (irregular) component $\ln \tilde{x}_t$ from the log of the time series, then we define the business cycle (irregular) component as $\hat{x}_t = x_t - \exp(\ln x_t - \ln \tilde{x}_t)$. For inventory investment the business cycle (irregular) component is defined as the first difference of the corresponding component of inventory stocks.

REFERENCES

Baxter, Marianne, and Robert G. King. “Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series,” Working Paper 5022. Cambridge, Mass.: National Bureau of Economic Research, February 1995.

Benhabib, Jess, and Roger E. A. Farmer. “Indeterminacy and Sunspots in Macroeconomics,” in *Handbook of Macroeconomics.* North Holland, forthcoming.

Blinder, Alan S. “Retail Inventory Behavior and Business Fluctuations,” *Brookings Papers on Economic Activity*, 2:1981, pp. 443–505.

__________, and Louis J. Maccini. “Taking Stock: A Critical Assessment of Recent Research on Inventories,” *Journal of Economic Perspectives*, vol. 5 (Winter 1991), pp. 73–96.

Burns, Arthur F., and Wesley C. Mitchell. *Measuring Business Cycles.* New York: National Bureau of Economic Research, 1946.

Conference Board. “Benchmark Revisions in the Composite Indexes,” *Business Cycle Indicators,* vol. 2 (December 1997), pp. 3–4.

Durlauf, Stephen N., and Louis J. Maccini. ”Measuring Noise in Inventory Models,” *Journal of Monetary Economics*, vol. 36 (January 1995), pp. 65–89.

Fisher, Jonas D. M., and Andreas Hornstein. “(S, s) Inventory Policies in General Equilibrium,” Federal Reserve Bank of Richmond Working Paper 97–7. May 1997.

Fitzgerald, Terry J. “Inventories and the Business Cycle: An Overview,” Federal Reserve Bank of Cleveland *Economic Review*, vol. 33 (Third Quarter 1997), pp. 11–22.

Hamilton, James D. *Time Series Analysis*. Princeton, N.J.: Princeton University Press, 1994.

Harvey, Andrew C. *Time Series Models*, 2d edition. Cambridge, Mass.: MIT Press, 1993.

Hornstein, Andreas, and Jack Praschnik. "Intermediate Inputs and Sectoral Comovement in the Business Cycle," *Journal of Monetary Economics*, vol. 40 (December 1997), pp. 573–95.

Lucas, Robert E. "Understanding Business Cycles," Carnegie-Rochester Conference Series on Public Policy, vol. 5 (1977), pp. 7–29, reprinted in Robert E. Lucas, *Studies in Business Cycle Theory*. Cambridge, Mass.: MIT Press, 1989, pp. 215–39.

Metzler, Lloyd A. "The Nature and Stability of Inventory Cycles," *Review of Economics and Statistics*, vol. 23 (Third Quarter 1941), pp. 113–29, reprinted in Robert A. Gordon and Lawrence R. Klein, eds., *Readings in Business Cycles*. Homewood, Ill.: Irwin, 1965.

West, Kenneth D. "A Note on the Econometric Use of Constant Dollar Inventory Series," *Economic Letters*, vol. 13 (Fourth Quarter 1983), pp. 337–41.

New Evidence Connecting Exchange Rates to Business Cycles

Alan C. Stockman

Virtually every theoretical model of exchange rates predicts that the real exchange rate between two countries (with floating nominal exchange rates) is correlated with the ratio of business-cycle conditions in the two countries. Yet almost no empirical evidence exists to support this prediction of the models. In fact, there is little empirical evidence that ties real exchange rates to *any* underlying economic conditions. Some well-known studies have concluded that exchange rates appear to have "a life of their own," perhaps moving with speculators' expectations far more than with changes in economic fundamentals. (See Flood and Rose [1995] for a prominent example.)

Contrary to this widely held contention, this article presents new evidence that exchange rates *are* connected with fundamentals, in particular with the relative gross domestic product (GDP) of each of the two countries involved, as predicted by nearly all exchange-rate theories. Moreover, they are related in the direction predicted by standard models: a country's currency tends to be depreciated in real terms when that country's real GDP is relatively high, and vice versa.

Why have previous studies not found this relationship between real exchange rates and ratios of real GDP? The answer is that the relationship is hard to detect with simple linear models, because it appears to be nonlinear and conditional on persistent movements (rather than purely transitory movements) in the data. Recent theoretical and empirical work has pointed to the potential

The author is the Marie C. and Joseph C. Wilson Professor of Economics, Chair of the Department of Economics at the University of Rochester, and a Research Associate with the National Bureau of Economic Research. The author thanks Mike Dotsey, Marvin Goodfriend, Tom Humphrey, and Wenli Li for comments that significantly improved the article. The views expressed are those of the author and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

importance of nonlinearities in exchange-rate data and has pointed the way to the evidence reported here.

1. WHAT THEORETICAL MODELS SAY

A common feature of nearly all models of exchange rates is the prediction that the real exchange rate between two countries is correlated with the real-GDP ratio between them.[1] Define the real exchange rate as the exchange-rate-adjusted ratio of price levels: if the nominal exchange rate e is the price of foreign money (in units of home money) and if P and P^* are the home and foreign price levels, then the real exchange rate is $q \equiv eP^*/P$. The real exchange rate measures the relative price of a basket of foreign goods in terms of a basket of home goods.[2] An increase in q means "real depreciation" of home currency (a decrease in the relative price of home goods); a decrease in q means "real appreciation" of home currency (an increase in the relative price of home goods). Define the output ratio y as home real GDP divided by foreign real GDP. Most theoretical models of exchange rates predict a positive relationship between y and q.

The economic reasoning behind this prediction is not difficult, though its details differ depending on the model. First, consider an equilibrium model of exchange rates as outlined in Stockman (1987). The real exchange rate in that model, as in other equilibrium models[3] and some sticky-price models,[4] equals the marginal rate of substitution (MRS) in consumption between home and foreign goods. Consider a model with two countries and two goods, one produced exclusively in each country. A representative consumer in the home country has the utility function $U(x, y)$, where x and y represent consumption of home and foreign goods by consumers in the home country. A representative consumer in the foreign country has the utility function $U^*(x^*, y^*)$, where x^* and y^* represent consumption of home and foreign goods by foreign consumers. Regardless of many other details of a model like this, an equilibrium will usually entail a condition that says

$$MRS(x, y) = MRS^*(x^*, y^*) = 1/q, \tag{1}$$

where

$$MRS(x, y) = U_X(x, y)/U_Y(x, y) \tag{2a}$$

[1] Some models, such as in Stockman and Tesar (1995), do not necessarily make this prediction because they postulate demand shocks with flexible prices.

[2] The baskets of goods are the baskets used to measure the price indexes P and P^*.

[3] See, for example, Stockman (1980, 1987), Lucas (1982), and Svensson (1985).

[4] See, for example, Obstfeld and Rogoff (1995) and Kollmann (1997). The equilibrium differs in other models, such as in Chari, Kehoe, and McGrattan (1998).

and

$$MRS^*(x^*, y^*) = U_X^*(x^*, y^*)/U_Y^*(x^*, y^*) \tag{2b}$$

are the equilibrium marginal rates of substitution between home and foreign goods, X and Y, for the home and foreign consumers. The difference between the operations of most models with flexible or sticky prices lies in the behavior of the consumptions, x and y, *inside* the MRS function. Short-run price stickiness affects consumption, so it affects the real exchange rate q.

First consider a model with flexible prices and suppose that production of the home good, X, rises exogenously, holding fixed the production of the foreign good, Y. This rise in the supply of the home good typically reduces its equilibrium relative price; i.e., it typically depreciates the home real exchange rate (it raises q). Therefore a rise in the ratio of home-country output to foreign-country output is associated with home-currency real depreciation.

A simple example occurs when the utility functions are the same in both countries and the elasticity of substitution between home and foreign goods is unity. Then a 10 percent increase in production of X reduces its relative price by 10 percent in equilibrium, and consumers in each country increase their consumption of X by 10 percent and leave their consumption of Y unchanged.[5] With any standard utility function, an increase in x, holding y fixed, reduces $MRS(x, y)$ and thereby causes home real depreciation (a rise in q).[6] This discussion has assumed that home output rises exogenously with no change in foreign output. However, the same reasoning and results apply when home and foreign

[5] This occurs regardless of the asset market structure of the model—the result holds with complete asset markets or no asset markets at all. See Stockman (1987) and Cole and Obstfeld (1991). More generally, whether the increase in supply of X raises or reduces home consumption of the foreign good Y depends on details of the model. For example, if the elasticity of substitution between the two goods is very low, then the demand for X may be very inelastic, and a 10 percent increase in the supply of X may reduce its price so much that home consumers reduce consumption of Y.

[6] The implication in this example that home output rises relative to foreign output may appear to depend on the use of different units of measurement for the two outputs—with home output measured in units of the home good X and foreign output measured in units of the foreign good Y. However, the model continues to predict a positive relationship between the output ratio and the real exchange rate if both outputs are expressed in common units, as long as the elasticity of substitution between the goods is less than unity (so that demands for the goods are elastic). For example, suppose both are measured in units of the home good X. The value of foreign output, measured in units of the home good X, is qy^s, where y^s denotes foreign output of good Y. Letting x^s denote home output of good X, the ratio of home-to-foreign output, expressed in common units, is x^s/qy^s. With unit-elastic demands, this ratio stays constant as x^s and q move together. When demands are less than unit-elastic, this ratio rises and falls together with x^s and q. So units of measurement of the output ratio become an important issue only if demands are elastic. This article uses national GDP data with each country's real GDP expressed in units of its own production bundle to calculate the output ratio y, so this measurement issue does not apply. However, the measurement issue would become important if a similar analysis calculated the output ratio with an exchange-rate-adjusted ratio of nominal GDP series.

output move together (as in the data) with output in one country exogenously rising more than output in the other country.[7]

Not surprisingly, the model's implication for comovements of the output ratio and the real exchange rate is *reversed* for exogenous changes in demand. After all, changes in supply generate negative comovements of the output of a good and its price; changes in demand generate positive comovements. For various reasons, most models of exchange rates of the type discussed above have relied on productivity shocks rather than demand shocks to drive the model.[8] This reliance results partly from the relative success of real business-cycle models driven by technology shocks and partly from the difficult task of identifying demand shocks in the data. (Fiscal policy changes appear to be much too small and infrequent in the data to explain either business cycles or exchange-rate changes; taste shocks are inherently unobservable, though potentially measurable through their effects on the economy.)

Despite the small role of demand shocks in flexible-price models, they have played the key role in another class of models—those with sluggish nominal price adjustment. These models predict, even with demand shocks, that output ratios and real exchange rates are positively correlated—a rise in home output relative to foreign output (a rise in y) occurs together with real depreciation of home currency (a rise in q). Again, the basic economic reasoning is straightforward and robust to many perturbations of details in the models.

When prices are sticky in the short run, a vast array of business-cycle models generates the prediction that an increase in aggregate demand raises real output in the short run. Corresponding models of open economies also predict short-run home-currency depreciation (the nominal exchange rate, e, rises), which translates into *real* depreciation (an increase in q) because of sticky prices. Together, these two results imply that aggregate demand shocks create a positive relationship between the output ratio y and real exchange rate q.

Despite differences in detail across sticky-price models, the reasons for their exchange-rate predictions share a common feature: in each model, the exchange rate is determined through an uncovered-interest-parity condition.

[7] This basic logic is not sufficient to generate an unambiguous prediction for the balance of trade or the current account of the balance of payments, as Stockman (1990) explains. To see why, imagine first that the increase in home production is temporary and that nothing that people do can make it permanent. In that case, people in the home country will typically want to save a large fraction of this temporary increase in income, and they will save by lending to (investing in) people in foreign countries. This international lending creates a surplus in the balance of trade and the current account. Now suppose instead that the increase in home production can be made permanent (or at least longer lasting) through investment to expand the economy's capital stock. In this case, investment demand rises and the equilibrium increase in investment (in the short run) may exceed the equilibrium increase in production; if so, the economy runs a deficit in its trade balance and current account.

[8] See Stockman and Tesar (1995) for a counterexample.

Uncovered interest parity says that the expected rate of change of the *nominal* exchange rate equals the difference in nominal interest rates across countries. This condition derives from two components: (1) the arbitrage condition of covered interest parity, which states that the forward exchange rate relative to the spot exchange rate equals the difference in interest rates (and which is well substantiated in the data), and (2) the hypothesis that the forward exchange rate equals the expected future spot exchange rate. The second component, unlike the first, appears to be falsified in the data for reasons that economists do not yet understand.[9] Its falsification does not necessarily invalidate theories that use it as a component (any more than quantum effects invalidate Newtonian physics) because the violations (which appear to be either time-varying risk premia or systematic forecast errors) could be small enough that they do not materially affect the theory.

The common reasoning in these models states that the increase in aggregate demand (perhaps resulting from a monetary shock) depreciates the expected *future* nominal exchange rate. Given the difference in nominal interest rates across countries, uncovered interest parity then requires a corresponding depreciation of the *current* nominal exchange rate. In some models, such as in Obstfeld and Rogoff (1995), the shock to aggregate demand does not affect the interest differential, so the current nominal exchange rate immediately depreciates to its expected long-run level.[10] In other models, such as in Dornbusch (1976) or Chari, Kehoe, and McGrattan (1998), the shock to aggregate demand reduces the home nominal interest rate relative to the foreign rate, requiring a short-run depreciation in excess of the long-run depreciation. In both cases, nominal depreciation implies real depreciation (a rise in q) in the short run because of sticky prices.[11] As a result, the models imply a positive correlation between the output ratio y and real exchange rate q.[12]

[9] With exchange rates expressed in logarithms, arbitrage would make this second component true in a world without uncertainty. Froot and Thaler (1990) discuss the puzzle of empirical falsification of this component.

[10] In the Obstfeld-Rogoff model, for example, the short-run response of the output ratio is proportional to the short-run response of the nominal and real exchange rates, with the factor of proportionality determined by the elasticity of substitution in consumption between various goods.

[11] Even a model in which the nominal interest rate rises in the short run (because the expected-inflation effect of a monetary shock exceeds the fall in the real interest rate) would predict currency depreciation if the increase in the home nominal interest rate were smaller than the percentage increase in the long-run exchange rate.

[12] For reasons similar to those explained in footnote 7, these models do not make unambiguous, robust predictions about the response of the trade balance or current account, or about comovements between these and other variables.

2. EMPIRICAL METHODS AND RESULTS

Data for the models consist of quarterly observations on exchange rates and real GDP for Australia, Canada, France, Italy, Japan, the Netherlands, Spain, Switzerland, the United Kingdom, and the United States, generally over the period 1974:1 through 1996:4. These countries appear in the sample as representative of developed economies with some heterogeneity in geography and circumstances, yet with consistent and reliable data. Germany is notable for its absence in the sample. Its exclusion is due to data issues associated with unification, which could play a particularly important role with our short sample of less than 25 years per country.

Define the nominal exchange rate e_{ab} as the domestic price in country a of the (foreign) currency of country b, and q_{ab} as the (natural) logarithm of the relative price of country b's goods in terms of country a's goods:

$$q_{ab} = \ln\left(\frac{e_{ab}p_b}{p_a}\right).$$

An increase in q_{ab} indicates "real depreciation" of currency a and "real appreciation" of currency b. Define y_{ab} as the (natural) logarithm of the ratio of output in country a relative to country b:

$$y_{ab} = \ln\left(\frac{GDP_a}{GDP_b}\right),$$

where GDP_a indicates real GDP in country a.

As noted above, standard models with sticky prices, such as in Dornbusch (1976) and its variants, including Obstfeld and Rogoff (1995) and Chari, Kehoe, and McGrattan (1998), imply a *positive* relationship between q_{ab} and y_{ab}. Specifically, a monetary expansion in country a raises y_{ab} (above its long-run equilibrium value) and raises q_{ab} (that is, it causes real depreciation of currency a).

Simple correlations are not consistent with this prediction. For example, simple correlations between q_{ab} and y_{ab} appear in Table 1, with the United States as the comparison country for each exchange rate. They are negative in seven out of eight cases. Only the correlation for Switzerland takes the predicted positive sign, and that correlation is smaller than the absolute values of four of the seven negative correlations.

While most empirical economic research involves linear methods, recent empirical and theoretical work motivates an alternative nonlinear approach to the data. Recent empirical papers present evidence of nonlinearities in univariate time-series models of exchange rates.[13] These papers examine nonlinear models of exchange rate adjustment toward purchasing-power-parity levels

[13] See O'Connell (1997); O'Connell and Wei (1997); Michael, Nobay, and Peel (1997); Obstfeld and Taylor (1997); and Taylor and Peel (1997).

Table 1 Simple Correlations

Italy	−0.48
Japan	−0.62
U.K.	−0.23
France	−0.49
Switzerland	0.38
Spain	−0.15
Netherlands	−0.80
Australia	−0.17

when adjustment takes place only or mainly outside a band around the long-run level of the exchange rate or when adjustment is simply more rapid because exchange rates begin farther from their long-run levels. Recent evidence strongly indicates nonlinearities in exchange-rate behavior, with faster mean reversion in exchange rates when they begin far from purchasing-power-parity levels. This result means that the half-life of real exchange rates (the time it takes them to return to their long-run levels following a one-time shock) is apparently lower than previous research had indicated. Whether the half-life is small enough for consistency with theoretical models remains unknown.[14]

Similarly, recent theoretical work suggests reasons for weaker adjustment of exchange rates toward purchasing power parity when the economy is close to its long-run equilibrium. Uppal and Sercu (1996) have explored a simple model of international arbitrage costs in product markets that implies little or no adjustment of real exchange rates toward long-run levels when the economy is close to long-run equilibrium. Proportional arbitrage costs create a *no-arbitrage band* around a long-run equilibrium, in which the real exchange rate can vary without tendency to return to a mean.[15] Outside that band,

[14] The theoretical models discussed above use the same shocks to generate business cycles and changes in real exchange rates, so they predict not only correlations between the two but similar half-lives of real exchange rates and output ratios. More research is required to test this prediction of the models.

[15] In a simple version of this model, countries are autarkic within the no-arbitrage band and the real exchange rate is indeterminate within the limits of that band. Consider a simple static model with a single homogeneous good randomly endowed and consumed in each of two countries. Imagine that the countries are identical ex ante in every respect. Suppose there are "iceberg" costs associated with shipping a good from one country to the other—when one unit of a good is exported by either country, only $k < 1$ goods arrive in the other country. (This situation resembles the commodity points discussed by Heckscher (1916) in analogy to the gold points of the classical gold standard, or zones of inaction in models of irreversible investment, or s-S inventory models.) Consider the equilibrium of such a model with complete markets in which consumers trade in asset markets prior to knowing the levels of random endowments. The equilibrium involves a no-arbitrage band—if endowments are sufficiently similar, then people consume their endowments, while if endowments are sufficiently dissimilar, then people with high endowments ship some of their goods to people with low endowments. Because of the "iceberg" costs of shipping, consumers choose not to trade on asset markets and not to equate consumption

arbitrage opportunities lead the real exchange rate back toward the band. This arbitrage-cost framework provides a loose interpretation of the recent empirical work mentioned above.[16] Ohanian and Stockman (1997) have extended the model developed by Uppal and his coauthors to a dynamic context.[17] While their model is directed at a different issue (explaining commonly suggested connections between exchange rates and international portfolio adjustments), it also implies that factors generating international portfolio adjustments inside the no-arbitrage band could mask connections between real exchange rates and other variables. As a result, connections between the real exchange rate and other variables, such as the output ratio, may be stronger outside that band. More generally, the signal-to-noise ratio may be larger when shocks are larger and drive either the output ratio or real exchange rate farther from its long-run equilibrium; we exploit this idea empirically below.

A key issue in analyzing exchange-rate data involves the choice of whether and how to filter those data. The correlations in Table 1 would be meaningless if either q_{ab} or y_{ab} were nonstationary. Unfortunately, standard statistical tests are unable to distinguish between the two main alternative hypotheses about these data: that they are trend-stationary or that they contain unit roots. This inability to distinguish the form of the trend in these data is not unique to this article—it plagues almost all analyses of real exchange-rate data and, indeed, macroeconomic analysis more generally. The hypothesis of trend-stationarity in q_{ab} means that the probability distribution of q_{ab} is stationary after a deterministic time trend has been removed from q_{ab}. The alternative hypothesis of a unit root in q_{ab} means that q_{ab} has a random trend, in the sense that the probability distribution of *changes* in q_{ab} is stationary. The same applies to trends in y_{ab}. One can investigate these trends by testing for the presence of a unit root in a standard unit-root test, such as the augmented Dickey-Fuller test. The results of unit-root tests for q_{ab} or y_{ab} are available from the author; however, they share a common characteristic with many unit-root tests in macroeconomic data in that they yield ambiguous results.

in the two countries. While a modified version of purchasing power parity holds outside the band, within the band, nominal and real exchange rates are indeterminate. With sticky prices or real costs of arbitrage outside the band, the real exchange rate can deviate from the band with only a slow tendency to return in the long run. It is this idea that forms the basis for the recent empirical studies of nonlinearities in exchange-rate behavior.

This simple model leaves many questions open. What are the arbitrage costs? How large are they? Evidence indicates that they are connected not only with distance but national borders—see Engel and Rogers (1996)—without providing reasons for this connection. These questions pose challenges for future research.

[16] The interpretation is loose for several reasons. One reason is that most recent empirical work uses linearly detrended (real or nominal) exchange-rate data. While changes in equilibrium relative prices may create such trends, the statistical models do not take into account the effects of this trend on the rate at which arbitrage would push the exchange rate back toward the (moving) band.

[17] See Uppal (1993); Uppal, Sercu, and Van Hulle (1995); and Uppal and Sercu (1996).

Unit-root tests have many well-known problems. One serious problem is that these tests have low power to reject unit roots (correctly, when the data in fact do *not* have unit roots) in short samples like those available for the current analysis, particularly when the true root is close to unity. For example, the power to reject a unit root is low in a short sample when a series follows a first-order autoregressive process, $y_t = \alpha y_{t-1} + u_t$, where $-1 < \alpha < 1$, but the root α is 0.95 or some other number close to 1. The power of these tests to reject unit roots correctly is also low when the data follow a more complicated (but stationary) time-series process, like the nonlinear processes studied in the papers mentioned in footnote 13. For example, Taylor and Peel (1997) show that nonlinear mean reversion can create a high probability of failing to reject unit roots when using standard methods, even when the data are actually stationary. For these reasons, this article follows most other research on real exchange rates by treating real exchange rates q as trend-stationary. For similar reasons, we also treat real-GDP ratios y as trend-stationary. This assumption makes results in this article more easily comparable with the bulk of other empirical work in the area.[18]

We begin with an examination of whether the detrended series, *DTq* and *DTy*, is more strongly related when one of the series is large (in absolute value) relative to its mean. Table 2 shows conditional correlations between detrended GDP ratios (*DTy*) and detrended real exchange rates (*DTq*) when the absolute value of the detrended real exchange rate is large. The first column of statistics shows unconditional correlations; the second column shows the correlations that are conditional on the absolute value of $DTq > 0.1$ (which means that the detrended real exchange rate is at least 10 percent above or below its mean). The interpretation of this table requires an economic model. One might think about an arbitrage model with a no-arbitrage band of unknown size and interpret the table as capturing situations in which the exchange rate is outside that arbitrage band. Alternatively, one might think that many factors other than business-cycle conditions affect the relationship between exchange rates and output ratios in normal times and that when the exchange rate is close to its mean, this “noise” makes it difficult to detect a consistent relationship. Under this interpretation, larger deviations of exchange rates from their means may indicate times when the “signal-to-noise” ratio is larger, allowing economists potentially to observe the relationships predicted by the theories outlined earlier.

[18] The usual assumption that the real exchange rate is trend-stationary, adopted here, implies a *failure* of long-run absolute purchasing power parity (though not relative purchasing power parity). The underlying assumption is that something—like differences in productivity growth causing differences in the trend relative price of nontraded goods, which are part of the bundles of goods included in the real exchange rate measure—cause the equilibrium real exchange rate to show a trend. The data clearly support the presence of *some* trend, whether stochastic or (as here) deterministic and linear.

Table 2 Correlations between Detrended Exchange Rates and GDP Ratios Conditional on Size of Exchange Rates

Country Pair	Unconditional	ABS(*DTq*) > 0.1	ABS(*DTq*) > 0.15	ABS(*DTq*) > 0.2
U.K., Japan	0.66	0.78	0.79	0.79
France, Japan	−0.21	−0.31	−0.01	−0.02
Italy, Japan	0.15	0.14	0.24	0.34
Switzerland, Japan	0.08	0.17	0.20	NA
Australia, Japan	0.18	0.29	NA	NA
Netherlands, Japan	−0.32	−0.45	NA	NA
Spain, Japan	−0.03	0.05	0.25	NA
Canada, Japan	0.41	0.50	0.46	NA
U.K., U.S.	−0.12	−0.27	−0.42	NA
France, U.S.	−0.21	−0.69	NA	NA
Italy, U.S.	0.01	0.01	−0.15	NA
Switzerland, U.S.	0.16	0.27	0.38	0.63
Australia, U.S.	−0.23	−0.59	0.21	0.26
Netherlands, U.S.	−0.47	−0.64	−0.63	NA
Spain, U.S.	−0.50	−0.57	NA	NA
Japan, U.S.	0.46	0.62	0.68	0.69

NA = Not available due to insufficient number of observations.

Table 2 examines the connection between output ratios and real exchange rates when the latter are far from their detrended means, measured as when the absolute value of the detrended log real exchange rate—ABS(*DTq*)—exceeds 0.1, 0.15, or 0.2. The results show, at best, a very weak connection between output ratios and exchange rates even when exchange rates are far from their means. Half (eight of 16) of the unconditional simple correlations are positive; only nine of 16 are positive when the absolute value of the detrended exchange rate exceeds its mean by 10 percent or more.

Table 3 is analogous to Table 2 in that it shows conditional correlations between detrended GDP ratios (*DTy*) and detrended real exchange rates (*DTq*). While Table 2 conditions on a large real exchange rate, Table 3 conditions on a large absolute value of the detrended output ratio. Again, one might use the arbitrage model to interpret this table as capturing situations in which the economy is outside the arbitrage band. Alternatively, one might think of large output ratios as indicating times when the signal-to-noise ratio is large enough that we could find evidence of the comovements predicted by the theories outlined above.

Like the results of Table 2, those in Table 3 fail to show any strong connection between output ratios and exchange rates even when output ratios are far from their means. The first column of statistics shows unconditional correlations between real exchange rates and output ratios; the second column shows the

Table 3 Correlations between Detrended Exchange Rates and GDP Ratios Conditional on Size of GDP Ratios

Country Pair	Unconditional	ABS(*DTy*) > 0.01	ABS(*DTy*) > 0.02	ABS(*DTy*) > 0.03
U.K., Japan	0.63	0.67	0.73	0.78
France, Japan	−0.19	−0.22	−0.32	−0.29
Italy, Japan	0.23	0.28	0.23	0.22
Switzerland, Japan	0.10	0.08	0.18	0.36
Australia, Japan	0.18	0.25	0.34	0.33
Netherlands, Japan	−0.32	−0.34	−0.32	−0.38
Spain, Japan	−0.03	−0.03	−0.06	−0.13
Canada, Japan	0.54	0.59	0.75	0.81
U.K., U.S.	−0.12	−0.14	−0.20	−0.10
France, U.S.	−0.21	−0.23	−0.22	−0.25
Italy, U.S.	0.03	0.02	0.10	0.25
Switzerland, U.S.	0.15	0.15	0.18	0.40
Australia, U.S.	−0.23	−0.28	−0.27	−0.11
Netherlands, U.S.	−0.47	−0.52	−0.53	−0.47
Spain, U.S.	−0.50	−0.55	−0.57	−0.64
Japan, U.S.	0.46	0.63	0.66	0.66

correlations conditional on the absolute value of $DTy > 0.01$ (which means that the detrended output ratio is at least 1 percent above or below its mean). Other columns show the results for larger values of output ratios. Evidently, the correlation is not much different when the real exchange rate is far from its mean (as in the last three columns of Table 2) or when the output ratio is far from its mean (as in the last three columns of Table 3) than when both are close to their means.

Tables 2 and 3 address the issue of large versus small values of detrended real exchange rates and output ratios but not issues of transitory versus more persistent changes in these variables. Business cycles refer to changes in real GDP that are sustained over at least several quarters. One might expect a stronger relationship between exchange rates and real GDP over business cycles and longer periods than over short periods, which may see many unrelated, transitory changes in output ratios and exchange rates.

We now turn to the connection between real exchange rates and output ratios when changes in either are sustained for several quarters. Table 4 shows that the results change substantially when we condition in a different way to capture business cycles. Although the table focuses on Japan as the base country, similar results appear when the United States is the base country. Column 1 shows the percentage of positive observations of a country's detrended real exchange rate over the entire sample. The numbers are all close to one-half. Column 2 shows the percentage of positive observations of a country's

Table 4 Percentage of *Positive* Detrended Real Exchange Rates When Detrended GDP Ratios are *Positive*

	1	2	3	4	5	6
		Percent of Positive Observations				**Marginal**
	in whole	**if**	**if**	**if**	**if**	**probability**
	sample	$DTy > 0$	$DTy > 0$	$DTy > 0$	$DTy > 0$	**level for**
			for two	**for four**	**for six**	**column 5**
Country Pair			**quarters**	**quarters**	**quarters**	
U.K., Japan	52	82	87	91	89	10^{-6}
Canada, Japan	46	66	73	74	75	0.002
Switzerland, Japan	53	68	75	73	88	0.002
U.S., Japan	54	70	77	81	72	0.02
France, Japan	54	51	54	56	64	0.18
Netherlands, Japan	52	52	50	50	65	0.26
Italy, Japan	52	56	58	55	58	0.54
Spain, Japan	52	49	50	51	45	0.73
Australia, Japan	45	47	48	38	52	1

detrended real exchange rate conditional upon its detrended GDP ratio being positive in the same quarter. In six of nine cases, the percentage of positive observations of the detrended exchange rate is larger than in the entire sample (column 1). Columns 3 through 5 show the same information conditional upon detrended GDP ratios being positive for the next two, four, or six consecutive quarters. The percentage of positive detrended exchange rates is generally larger when we condition on positive GDP ratios. This result occurs in eight out of nine cases in column 5, and in most cases the percentage is quite high. Column 6 shows the marginal probability level for rejecting the null hypothesis that the numbers in column 5 arise purely by chance (such as when one draws them randomly from a binomial distribution). This “sign test” (treating positive and negative observations as binomial draws) assumes that the underlying distribution is symmetric, which appears to be roughly true in these data. (The fact that the numbers in the first column are close to one-half is consistent with this assumption). In four of the nine cases (the United Kingdom, Canada, Switzerland, and the United States), we can strongly reject the null hypothesis that the connection between detrended real exchange rates and detrended GDP ratios arises purely by chance. Even in the five cases in which we cannot reject that null hypothesis, more than half of the observations are positive in four of the five cases (France, the Netherlands, Italy, and Australia). For example, column 5 shows that almost two-thirds of the observations are positive for both France and the Netherlands.

One can achieve greater statistical power by pooling the data; however, the probability distribution is exceedingly complicated when one allows for dependence across country pairs. Overall, nearly two-thirds of the observations

Table 5 Percentage of *Positive* Detrended GDP Ratios When Detrended Real Exchange Rates are *Positive*

	1	2	3	4	5	6
		Percent of	**Positive**	**Observations**		**Marginal**
	in whole	**if**	**if**	**if**	**if**	**probability**
	sample	$DTq > 0$	$DTq > 0$	$DTq > 0$	$DTq > 0$	**level for**
			for two	**for four**	**for six**	**column 5**
Country Pair			**quarters**	**quarters**	**quarters**	
U.K., Japan	42	70	72	77	77	0.01
Canada, Japan	38	60	63	71	75	0.01
U.S., Japan	45	63	72	78	72	0.04
Australia, Japan	53	58	57	59	65	0.11
Switzerland, Japan	59	65	67	63	68	0.14
France, Japan	52	48	51	53	62	0.38
Spain, Japan	48	41	45	48	54	0.84
Netherlands, Japan	52	43	42	43	42	0.65
Italy, Japan	41	50	53	52	47	1

on the detrended real exchange rate in column 5 of Table 4 are negative (with 264 observations in that column, the detrended exchange rate is negative in 183 instances and positive in 81 instances). The probability that this occurs purely by chance would be only three hundred-millionths of 1 percent if data on detrended GDP ratios and exchange rates in each row of the table were independent. Of course, the fact that Japan is involved in each comparison means that the data are probably dependent. However, even with dependence across observations, the results in this table show strong evidence that when a country's detrended GDP rises for six straight quarters relative to Japan, its currency tends to take a low value on foreign exchange markets (q is high).

Table 5 shows analogous calculations that are conditional upon sustained depreciation of real exchange rates. Column 1 of Table 5 shows the percentage of positive observations of a country's detrended output ratio over the entire sample. Column 2 shows the percentage of positive observations of a country's detrended output ratio conditional upon its detrended real exchange rate being positive in the same quarter. Except in the Netherlands (and Spain for durations shorter than six quarters), exchange rates show real depreciation significantly more often when the GDP ratio is high for several quarters (than in the overall sample). Unfortunately, the number of observations available for calculations in column 6 is sufficiently low that we can strongly reject the null hypothesis that this result arises purely by chance for only three of the nine cases in column 5—for the United Kingdom, Spain, and the United States.

Tables 4 and 5 examine groupings of observations in these data that are conditional upon whether one of the variables exceeds its mean for some duration. Table 6 extends this evidence by testing the null hypothesis that

Table 6 Test Results: Means of Detrended Exchange Rates Conditional on Signs of Detrended GDP Ratios

	1	2	3
Country Pair	**Mean detrended (log) exchange rate if detrended GDP ratio exceeds its median**	**Mean detrended (log) exchange rate if detrended GDP ratio is less than its median**	**t-statistic for null hypothesis that the means in columns 1 and 2 are equal**
U.K., Japan	0.08	−0.08	−11.0
Canada, Japan	0.07	−0.07	−9.0
U.S., Japan	0.05	−0.05	−6.4
Italy, Japan	0.01	−0.05	−4.3
Australia, U.S.	0.03	−0.02	−3.0
Switzerland, Japan	0.01	−0.00	−0.6
Spain, Japan	0.00	0.00	1.1
France, Japan	−0.01	0.01	2.3
Netherlands, Japan	−0.01	0.03	4.3

the mean detrended exchange rate does not depend on whether the detrended GDP ratio is above or below its median value. The large absolute values of the t-statistics in column 3 of Table 6 indicate rejection of that null hypothesis. In seven of nine cases, we can reject the hypothesis that the means of the real exchange rates do not depend on the GDP ratio. In five of those seven cases the direction of the difference in means is the direction predicted by standard models. The exceptions are France and the Netherlands. In most cases, the evidence in this table, as in Tables 4 and 5, clearly indicates that high GDP ratios and depreciated real exchange rates tend to occur together in the data.

Because the results of Tables 4 and 5 indicate that the connection between real exchange rates and GDP ratios is strongest when one of the series shows sustained movement away from its mean, Table 7 adds duration to the test shown in Table 6. Specifically, it repeats the test in Table 6 with the condition that the detrended GDP ratio is above or below its median value for six straight quarters. In every case we can reject the hypothesis that the means do not depend on the GDP ratio; in six of the nine cases, the connection is in the direction predicted by standard models. Spain now joins France and the Netherlands as exceptions. Table 7 provides even stronger evidence that high GDP ratios and depreciated real exchange rates tend to occur together.

Together, Tables 4 through 7 provide the strongest evidence currently available that changes in real exchange rates are connected to changes in real-GDP ratios over periods of several quarters, as predicted by nearly all existing models of exchange rates.

Table 7 Test Results: Means of Detrended Exchange Rates Conditional on Signs and Duration of Detrended GDP Ratios

	1	2	3
Country Pair	**Mean detrended (log) exchange rate if detrended GDP ratio exceeds its median for six straight quarters**	**Mean detrended (log) exchange rate if detrended GDP ratio is less than its median for six straight quarters**	**t-statistic for null hypothesis that the means in columns 1 and 2 are equal**
U.K., Japan	0.11	−0.13	−15.9
Canada, Japan	0.10	−0.10	−12.3
U.S., Japan	0.06	−0.05	−6.9
Italy, Japan	0.02	−0.11	−9.4
Australia, U.S.	0.04	−0.07	−6.8
Switzerland, Japan	0.06	−0.02	−8.3
Spain, Japan	−0.03	0.02	2.9
France, Japan	0.01	0.03	2.3
Netherlands, Japan	−0.01	0.04	5.1

3. CONCLUSIONS

The inability of economists to find strong statistical relationships between exchange rates and underlying economic conditions has been a huge puzzle in international economics. In particular, standard models of both the sticky-price and flexible-price varieties predict that real depreciations of a country's currency tend to occur along with increases in its output relative to foreign output. The findings reported here are probably the strongest evidence yet that this relationship appears in the data. The same findings show that the relationship is nonlinear and conditional. These empirical results raise a set of new questions for future research, particularly regarding related nonlinear and conditional connections between exchange rates and other variables. It appears that exchange rates, after all, do not have lives of their own.

REFERENCES

Chari, V. V., Patrick J. Kehoe, and Ellen R. McGrattan. "Monetary Shocks and Real Exchange Rates in Sticky Price Models of International Business Cycles." Federal Reserve Bank of Minnesota Staff Report 223. Revised January 1998.

Cole, Harold L., and Maurice Obstfeld. "Commodity Trade and International Risk Sharing: How Much do Financial Markets Matter?" *Journal of Monetary Economics,* vol. 28 (August 1991), pp. 3–24.

Dornbusch, Rudiger. "Expectations and Exchange Rate Dynamics," *Journal of Political Economy,* vol. 84 (December 1976), pp. 1161–76.

Engel, Charles, and John Rogers. "How Wide is the Border?" *American Economic Review,* vol. 86 (December 1996), pp. 1112–25.

Flood, Robert, and Andrew Rose. "Fixing Exchange Rates: A Virtual Quest for Fundamentals," *Journal of Monetary Economics,* vol. 36 (August 1995), pp. 3–37.

Froot, K., and R. H. Thaler. "Anomalies: Foreign Exchange," *Journal of Economic Perspectives,* vol. 4 (Summer 1990), pp. 179–92.

Hecksher, Eli. "Växelkursens grundval vid pappersmynfot," *Ekonomisk Tidskrift,* vol. 18, 1916.

Kollmann, Robert. "The Exchange Rate in a Dynamic Optimizing Current Account Model with Nominal Rigidities: A Quantitative Investigation," IMF Working Paper 97/7.

Lucas, Robert. "Interest Rates and Currency Prices in a Two-Country World," *Journal of Monetary Economics,* vol. 10 (November 1982), pp. 335–59.

Obstfeld, Maurice, and Kenneth Rogoff. "Exchange Rate Dynamics Redux," *Journal of Political Economy,* vol. 103 (June 1995), pp. 624–60.

Obstfeld, Maurice, and Alan Taylor. "Nonlinear Aspects of Goods-Market Arbitrage and Adjustment: Heckscher's Commodity Points Revisited," *Journal of the Japanese and International Economics,* vol. 11 (1997), pp. 441–79.

O'Connell, Paul. "Market Frictions and Relative Traded Goods Prices," unpublished, Harvard University, 1997.

__________, and S. J. Wei. "The Bigger They Are, the Harder They Fall: How Price Differences between U.S. Cities are Arbitraged," unpublished, Harvard University, April 1997.

Ohanian, Lee E., and Alan C. Stockman. "Arbitrage Costs and Exchange Rates," unpublished, University of Rochester, April 1997.

Panos, Michael, A. Robert Nobay, and David A. Peel. "Transactions Costs and Nonlinear Adjustment in Real Exchange Rates: An Empirical Investigation," *Journal of Political Economy,* vol. 105 (August 1997), pp. 862–79.

Stockman, Alan C. "International Transmission and Real Business Cycle Models," *American Economic Review,* vol. 80 (May 1990), pp. 134–38.

__________. “The Equilibrium Approach to Exchange Rates,” Federal Reserve Bank of Richmond *Economic Review,* vol. 73 (March–April 1987), pp. 12–31.

__________. “A Theory of Exchange Rate Determination,” *Journal of Political Economy,* vol. 88 (August 1980), pp. 673–98.

__________, and Linda Tesar. “Tastes and Technology in a Two-Country Model of the Business Cycle: Explaining International Comovements,” *American Economic Review,* vol. 85 (March 1995), pp. 168–85.

Svensson, Lars E. O. “Currency Prices, Terms of Trade, and Interest Rates: A General Equilibrium Asset-Pricing, Cash-in-Advance Approach,” *Journal of International Economics,* vol. 18 (February 1985), pp. 17–41.

Taylor, Mark P., and David A. Peel. “Nonlinearities in Real Exchange Rate Adjustment during the Recent Float: Empirical Evidence and Monte Carlo Analysis,” Working Paper. Oxford: University College, 1997.

Uppal, Raman. “A General Equilibrium Model of International Portfolio Choice,” *Journal of Finance,* vol. 48 (June 1993), pp. 529–53.

__________, and P. Sercu. “Exchange Rate Volatility and Trade: A General Equilibrium Analysis,” Working Paper. University of British Columbia, 1996.

__________, __________, and C. Van Hulle. “The Exchange Rate in the Presence of Transaction Costs: Implications for Tests of Purchasing Power Parity,” *Journal of Finance,* vol. 50 (September 1995), pp. 1309–19.

Zimmermann, C. “International Business Cycles and Exchange Rates,” Working Paper 33. Centre de Recherche sur l’Emploi et les Fluctuations Economiques, Université du Québec à Montréal, December 1994.