Selling Federal Reserve Payment Services: One Price Fits All?

John A. Weinberg

In a large modern economy, there is a vast and constant movement of funds in the conduct of commerce and finance. The channels through which these funds move constitute the payment system, which, ultimately, forms a network connecting all participants in the economy. In dollar value, the bulk of this movement is not in cash but in the form of instructions for the crediting and debiting of accounts held with public or private financial institutions.[1] As a network for sending and receiving instructions, the payment system bears a resemblance to transportation and, especially, communication systems. Accordingly, many of the issues and questions that arise in discussions of markets for payment services have parallels in discussions of these other markets.

Markets that are characterized as networks are often thought to be driven by the existence of economies of scale. In the presence of scale economies, the average cost of providing services declines with the size of the network and the volume of traffic it carries. The belief in such economies has motivated a long history of direct government involvement and intervention in network markets, from the operation of the postal service to the regulation of telecommunications and transportation networks.

Much of the evolution of the structure of markets for payment services has been driven by the desire of participants to take advantage of the economies of network expansion. The most fundamental example is the replacement of a system in which payments are made in currency directly between individuals to one in which payments are made through accounts with financial intermediaries. Specifically, a check-based payment system opened the door to

This paper has benefited from the helpful comments of Bill Cullison, Jeffrey Lacker, Bruce Summers, John Walter, and Tom Humphrey. The views expressed herein are the author's and do not represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] For a detailed description of the payment system, see Blommestein and Summers (1994).

network efficiencies to be gained through the centralized exchange of checks among banks in clearinghouses.[2] More recently, some payments have moved from checks into electronic forms of transmission. For instance, the use of Automated Clearinghouse (ACH) payments, for such purposes as payroll direct deposit, tripled in the number of transactions processed annually (from around 800 million to around 2.4 billion) from 1986 to 1992.[3]

In addition to technological factors, the evolving market structure in the payment system has been greatly influenced by the policy of the Federal Reserve System. Prior to 1980, many payment services were provided free of charge by the Federal Reserve to its member banks. As a result, a majority of payments cleared through the Federal Reserve, either directly or through correspondent banks. The Monetary Control Act of 1980, among its provisions, required Federal Reserve services to be made available, at a price, to all institutions. The Reserve Banks were instructed to set prices to cover all direct and indirect costs incurred in the provision of services. Since the institution of pricing, the Reserve Banks have experienced losses in market share to private providers. In check processing, for instance, renewed growth has occurred in the activities of clearinghouses on local, regional and, most recently, national levels. The resulting loss of market share by the Reserve Banks has been most significant among larger institutions.[4]

In the provision of ACH services, the Fed's position is somewhat more dominant than in check services. The Federal Reserve processed about 94 percent of all transactions in 1992 (McAndrews 1994). Private alternatives continue to develop, however. As in the case of check processing, new competition and the potential for institutions to engage in direct (nonintermediated) exchanges are focused on large-volume ACH users.

In the changing payment services environment, there have been a number of proposals for the restructuring of Fed pricing. Proposals for *market-sensitive* pricing tend to suggest advantageous pricing terms to large-volume users of services. Any such scheme amounts to some form of *price discrimination*. This term is purely descriptive: it applies to any pricing other than the setting of a single price per unit sold that is available to all buyers. The simplest example, referred to as *two-part pricing*, involves charging all buyers the same combination of a fixed fee and a per-unit price. When two-part pricing does not "discriminate enough," more complex schemes can be used. Examples include a per-unit price that varies with the quantity purchased and a schedule of combinations of fixed and per-unit charges among which buyers can choose.

[2] Goodfriend (1990) discusses this change and how banks developed institutions for enhancing payments efficiency and dealing with the resulting credit risk.

[3] McAndrews (1994) describes the growth in ACH payments.

[4] The General Accounting Office (1989) found that between 1983 and 1987 the Federal Reserve lost market share only among banks with over $750 million in assets.

Price discrimination in response to market competition raises some important questions about Federal Reserve pricing policy. For instance, are the Reserve Banks' "business interests" in conflict with their public policy responsibilities? Additional questions arise from the fact that price discrimination has the tendency to favor some institutions, particularly larger institutions, over others. Should the equal treatment of all banks be part of Fed pricing policy? Of course, at the most basic level is the question of whether the Fed should participate at all as a competitor with private providers of payment services.

This article argues that the public interest may be best served by a Federal Reserve pricing policy that is responsive to competition, within certain limits. This argument is based on the presumption that an important goal for Federal Reserve policy is the resource efficiency of the payment system. An efficiency perspective dictates that a loss of market share by the Federal Reserve is neither good nor bad per se. What matters is the overall cost efficiency of the market. If the Federal Reserve is replaced by providers with lower costs, then such a change should be accommodated. The goal of pricing policy, however, should be that only efficiency-enhancing losses are experienced.[5]

The central concept employed in this article is that of *sustainable prices*.[6] Sustainable prices are prices designed to sustain an efficient allocation of production by giving no buyer an incentive to seek to obtain the product from an alternative source. The following section briefly describes the organization and pricing in the markets for check clearing and ACH services. These two markets can be broadly characterized by a high volume of low-value transactions. As such, they make relatively intensive use of resources in transmitting payment instructions and constitute large markets for transmission services. The subsequent sections develop the notion of sustainable prices and use it to draw conclusions about Fed pricing policy. In particular, sustainable pricing can provide a guide for determining when market-sensitive pricing by the Fed is and is not in the public interest.

It is important to note that resource efficiency is not the Federal Reserve's only public policy interest in the payment services market. Indeed, the Fed's primary concern is with the overall safety and reliability of the system. This concern is expressed in the Fed's regulatory oversight of arrangements used for payment settlement. It is along the dimension of efficiency, however, that the Fed's role as a provider of many payment services should be evaluated. The Fed's participation should be determined by its ability to provide services in a cost-effective manner.

---

[5] While this article focuses on pricing, the terms of competition among alternative providers are affected by a variety of other factors. For instance, in 1994 the Board of Governors adopted a requirement of same-day settlement of checks presented by private collecting banks that put private-sector processing on a more equal footing with Fed processing.

[6] See Spulber (1989).

1. TWO PAYMENT SERVICES MARKETS IN BRIEF: CHECKS AND ACH

The concept of sustainable prices, as developed below, applies to concentrated markets.[7] Hence, it is useful to establish at the outset that markets for payment services tend to be fairly concentrated. In most of these markets, the Federal Reserve has a significant market share, while in some markets, the Fed's share is dominant. A brief description of the structure of two markets follows.

In 1992, over 72 billion noncash payments were made in the United States. Of these, 80 percent were made by check.[8] Checks are written on more than 15,000 banks and other depository institutions. In about 30 percent of all transactions made with checks, the recipient deposits the payment in an account in the bank on which the check is written. The clearing of these "on-us" items is a simple matter; the bank merely debits the account of the payor and credits the account of the payee (subject, of course, to the payor's account having sufficient funds). The remaining 70 percent of check payments must clear between banks. This clearing can proceed directly: a payee bank can send the check to a payor bank in exchange for funds. Alternatively, check clearing can make use of one or more of a number of intermediary services.[9] One such service is that provided by a clearinghouse. In a clearinghouse arrangement, a number of institutions agree to exchange checks drawn on each other at a specified place and time. Hence, a clearinghouse resembles multilateral direct exchange, except in the way that payments are cleared. With each exchange of checks, a clearinghouse member pays its net debit position or receives its net credit.

If a bank participates in a clearinghouse of any size or if it engages in direct exchange with a large number of banks, it must have the capacity to sort the checks it receives by payor bank. This task is performed by specialized equipment, reader-sorter machines. If a bank chooses not to invest in sorting capacity, it can, instead, send unsorted or incompletely sorted checks to an intermediary institution that completes the collection process. Both the Federal Reserve Banks and private collecting banks play this role. The collecting bank, private or Fed, may sort and send checks to payor banks or to subsequent collecting banks. For instance, a Federal Reserve office sends within-district checks to payor banks and out-of-district items to their respective Fed offices. In 1992, the Fed handled over 19 billion checks, about half of all checks requiring interbank clearing.

The resource costs in the check-collection process are dominated by two cost categories: the sorting and transportation of checks. Direct, bilateral

[7] For a treatment of the wide variety of theories of behavior in concentrated markets, see Tirole (1989).

[8] The data cited in this section are from the Bank for International Settlements (1993).

[9] The various paths for check clearing are reversed when a payor bank sends a "return item" (a check returned because of insufficient funds).

exchange of checks is the most costly means of clearing since it requires the payee bank to sort and ship to a large number of endpoints. Concentration of both activities can lead to cost savings. A group of banks that regularly receive checks drawn on each other can economize through a clearinghouse arrangement. Hence, the typical clearinghouse is composed of relatively large institutions within a metropolitan area. When an institution does not internalize the economies of concentration, it can instead purchase sorting and transportation services from entities that can take advantage of the cost efficiencies available.

The use of Automated Clearinghouse transactions is relatively new. In an ACH payment, the payor (or the payee with preauthorization by the payor) gives direct instructions to the payor's bank for the transfer of funds. Modern electronic information technology has made this means of transfer particularly cost-effective for recurring payments of set value. Accordingly, a growing fraction of the work force has wage and salary payments directly deposited into bank accounts by ACH. Other payments that might be made by ACH include mortgage payments and insurance premiums.

As with checks, ACH payments must clear between banks when the payor and the payee do not have accounts with the same institution. Clearing is facilitated if the payor and payee bank share an electronic connection over which instructions can be sent. Transactions can be made by direct bilateral exchange, through a private clearinghouse, or through the Fed. The first two options are likely to be used primarily by pairs or groups of banks that share a large number of payments. That is, private ACH transactions have been carried out primarily within geographic regions, while for interregional payments, the Fed has been the dominant provider. This market structure may be subject to change, however, as a private, national ACH initiative has recently begun competing with the Fed. In 1992, 94 percent of approximately 1.8 billion ACH transactions were made through the Fed.

Current pricing of Federal Reserve check and ACH services is a form of two-part pricing, a combination of a fixed fee and a per-unit price. In check services, the fixed charge is the *cash letter charge*. A cash letter is a collection of checks deposited with the Fed. The cash letter charge and the per-item fee vary with the amount of sorting that has already been done by the depositing bank and with the locations of the banks on which the deposited checks are drawn. The Federal Reserve Bank of Richmond's price structure for 1994 includes cash letter charges between $2 and $3 for most checks, while per-item fees range from less than 1¢ to 6¢.[10] These different fee combinations apply to varying amounts of sorting that may be necessary.

[10] Larger cash letter and per-item charges are assessed for some special categories of checks.

The prices for ACH services also vary with the particular services provided. The basic fee structure in the Richmond Fed's 1994 price list includes a participation fee of $20 per account per month and transaction (per-item) fees of 1¢ per intradistrict item and 1.4¢ per interdistrict item. In addition, a bank must have electronic access to the system. Access is priced with a monthly fee that ranges from $30 to $1000, depending on the type of connection maintained. While electronic access allows institutions to receive other services as well, at least part of the access fee can be considered the fixed cost of engaging in ACH transactions.

2. NATURAL MONOPOLY[11]

The main concepts to be employed can be demonstrated with a simple example of a single service that can be provided by one or more sellers. Let q_i be the quantity provided to the ith out of N buyers. Denote by q the array of quantities provided to all the buyers, $q = (q_1, q_2, \ldots q_N)$, and let Q be the sum of the q_i. The total cost incurred by a single seller in providing the service is given by

$$C(Q) = F + \sum_{1}^{N} f_i + v(Q). \tag{1}$$

The fixed cost has two components. A general cost of F, the *common* fixed cost, is incurred by any seller providing any quantity of the service (e.g., the cost of maintaining an accounting and communication system for ACH transfers). In addition, there may be a cost of f_i specific to the relationship with buyer i (the cost of an individual bank's electronic connection to the system). The variable-cost function, $v(Q)$, is increasing and convex.[12]

The basic ideas can be presented for the simple case in which only the common fixed cost, F, is present in equation (1). In this case, the relationship between total cost and output might be represented as in Figure 1. The corresponding relationship between average cost and output is shown in Figure 2. This U-shaped average-cost curve exhibits economies of scale as long as

[11] The case of natural monopoly is developed for expository purposes. The concept of sustainable pricing can be extended to any market structure. The application to concentrated, nonmonopoly markets closely parallels the case of natural monopoly. For instance, if all sellers can operate at minimum average cost, then that minimum cost is the sustainable price.

[12] It is worth pointing out that the N quantities (q_i) specified above could just as easily be interpreted as quantities of N different products. In that case, the variable-cost function $v(Q)$ might be replaced by a sum of separate cost functions, $v(q_i)$, for each of the individual products. The concepts developed here to analyze pricing of a single product in the presence of economies of scale are directly applicable to the pricing of a set of products in the presence of *economies of scope*. Economies of scope are said to exist when the costs of joint production of a set of products is less than the sum of the costs of separate production.

Figure 1 Total Cost Curve

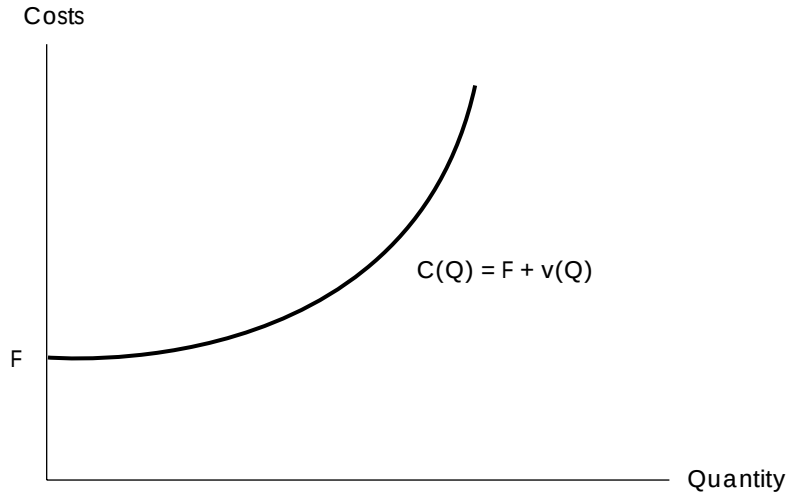
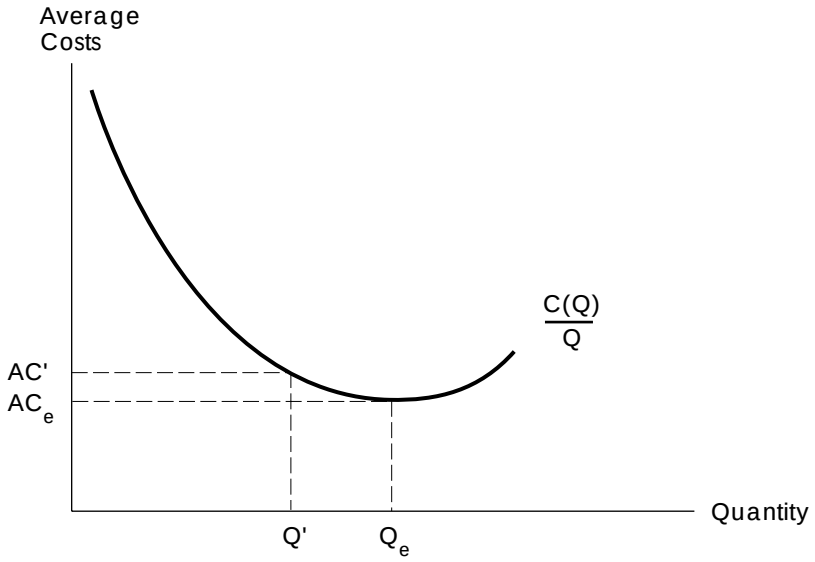


Figure 2 Average Cost Curve



Note: Figure 1 displays a total cost curve with fixed cost F and convex variable-cost function $v(Q)$. Figure 2 shows the corresponding average-cost function. The quantity Q_e is the "efficient scale" at which minimum average cost AC_e is achieved.

total output is less than the level labeled Q_e.[13] This level of output, at which average cost is minimized, is referred to as the *efficient scale* for the production of the service. Above the efficient scale, as average cost rises, there are diseconomies of scale. For now, it is assumed that all sellers and potential sellers have identical cost structures.

As in any market, pricing is affected by the structure of the market—e.g., the number and relative sizes of sellers. Market structure is, in turn, affected by the nature of the cost function for producing the service. If the total quantity demanded in this market was very large relative to Q_e, then competitive pricing and free entry among providers of the service would tend to result in a market composed of a large number of providers, each producing about Q_e. The price in this competitive market would tend toward AC_e in Figure 2. That is, when efficient scale is small relative to the size of the market, the invisible hand of competition works well; production costs are minimized and price just covers costs.

At the opposite extreme is the case in which a single seller's efficient scale (Q_e) is at least as large as the total quantity of service demanded by the market. In this case, competition among active providers cannot enhance the efficiency of production. Any division of output among sellers will only serve to raise the overall economic costs of providing the service, by duplicating the fixed costs. This is a case of *natural monopoly.* Under the belief that competition is infeasible, price regulation is often imposed on industries which are thought to operate under the conditions of natural monopoly.

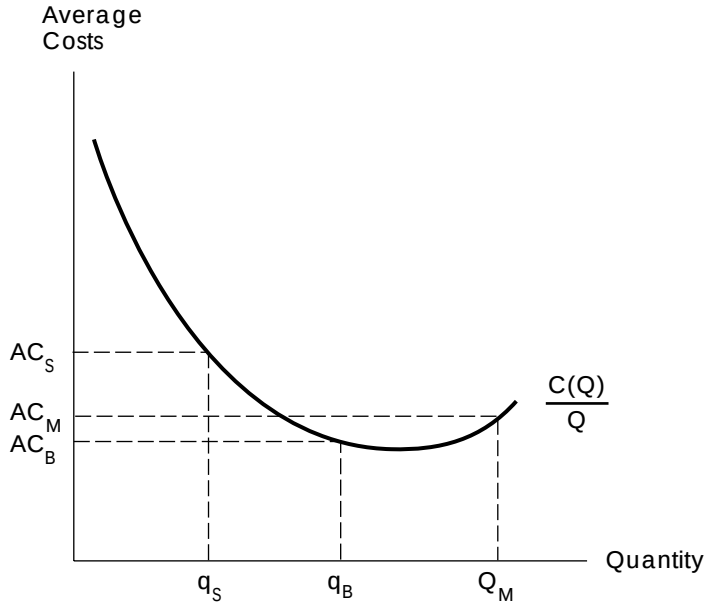
A natural focus for the pricing of the product sold by a natural monopoly, subject to the requirement that revenues just match costs, is to set a per-unit price equal to the average cost of producing the total industry output. Suppose, in Figure 2, that this output level is Q'. Suppose further that the quantity demanded is independent of price. This assumption is not essential but allows us to focus on the issue of whether or how the market quantity is divided among sellers. The price that just covers costs is AC'. Note that this price is greater than the marginal cost of production, since average cost is declining at Q'; when average cost is declining, marginal cost is less than average cost. Since price deviates from marginal cost, average-cost pricing in such cases is sometimes referred to as *second-best* pricing; "first-best" pricing would equate price to marginal cost, but would result in revenues less than costs. Second-best pricing maximizes net social benefits subject to the constraint that total revenues from the sale of the product just equal total costs.

[13] The average-cost curves should be understood as long-run average-cost curves. Although all factors of production are variable in the long-run, fixed costs are possible in the long-run if a minimum (positive) level of some input is necessary for production of any positive amount of output. For instance, to send telephone messages between two points, one must have, at least, one telephone line connecting those points. This represents a fixed cost, even in the long run.

Clearly, average-cost pricing in Figure 2 leaves no opportunity for a competitor to attract some piece of the market and at least cover its costs. Any piece of the market will involve average production costs greater than AC'. In order to win customers, however, a competitor would have to offer a price below AC'. In this case, uniform average-cost pricing (a per-unit price equal to AC' available to all buyers) is not vulnerable to the entry of competitors.

There is another case that falls into the category of natural monopoly for which pricing is more problematic. This case can be illustrated by a simple example in which there are two buyers, Big (B) and Small (S). Buyer B uses q_B units of the service, while S uses q_S. The total market quantity, then, is $q_B+q_S = Q_M$. Again, in the present example, quantity demanded is independent of price, except that each buyer seeks the lowest-cost supplier. The situation is depicted in Figure 3. Market quantity lies in the range of diseconomies of scale, and the average cost of serving the whole market is greater than the cost of serving just buyer B ($AC_M > AC_B$). Although market quantity exceeds efficient scale, the market is still a natural monopoly; any division of the market would result in higher total production costs. The average cost of serving only buyer S is greater than the average cost of serving the whole market ($AC_S > AC_M$).

Figure 3 Natural Monopoly with Quantity Greater than Efficient Scale



Note: Although market quantity, $Q_M = q_S + q_B$, is greater than efficient scale, the market is still a natural monopoly; the cost of serving the entire market is less than the combined cost of serving the market in any set of separate "pieces."

In this example, a simple price structure would set a uniform price equal to AC_M, the average cost of serving the entire market. If there are no legal barriers to entry, however, this price will induce a competitor to seek to gain a portion of the market. Specifically, a competitor can target buyer B, offering a price between AC_M and AC_B, the average cost of serving just buyer B. This strategy allows the competitor to take advantage of the economies of scale available in serving the large-volume user. Indeed, if no competitor were forthcoming and if buyer B had access to the necessary technology, then the buyer would be prompted to provide the service in-house.

A couple of comments on the competitor's pricing strategy are useful to bear in mind. First, the competitor must have reason to believe that the incumbent monopolist cannot or will not rapidly adjust prices in response to the competitor's move. Such a belief might be justified if the incumbent's pricing is subject to a cumbersome administrative procedure. Second, the competitor must be able to offer the lower price to a restricted set of buyers. If targeting a segment of the market requires making private deals with individual buyers, the competitor's task will be simpler if it is possible to identify a relatively small number of buyers with large enough volume to take substantial advantage of available economies of scale.

If the large-volume user defects to a competing source for the service, what becomes of the small-volume user? If the incumbent continues to offer the service at the price AC_M, then buyer S is just as well off as before. This price, however, no longer covers costs, which are now AC_S. Assuming the incumbent must cover costs, its price must rise. If it is resigned to serving only the remaining customer, S, then the incumbent must set its price at AC_S. Note that the end result may be an inefficient market structure. If there are two sellers operating, one serving buyer B and the other serving buyer S, then the duplication of fixed costs in serving the market constitutes social waste. The story may not end here. The incumbent may seek to win back some or all of the market share lost. This counterattack may ultimately succeed, but even temporary production by more than the efficient number of sellers is socially inefficient.

3. SUSTAINABLE PRICES

Are there pricing strategies for the incumbent that leave no room for encroachment by competitors? In the above example, the incumbent was vulnerable, because one buyer was charged a price that was greater than the cost of serving that buyer alone. The cost of serving only some subset of the buyers in a market is referred to as the *stand-alone* cost for those buyers. Accordingly, a set of buyers will be receptive to alternative sources of a service unless they face a price that is no greater than their stand-alone cost. A pricing scheme that meets this requirement for all sets of buyers is called a *sustainable* pricing

scheme. Sustainable prices leave no opportunity for a competitor with identical costs to capture any segment of the market.

How should one set prices that just cover costs and result in efficient production? In the case of natural monopoly, efficient production requires a single producer. In this case, the task is to find prices that recover costs and are sustainable. When market quantity is smaller than efficient scale, as in Figure 2, a uniform (per-unit) price equal to the average cost of producing the market quantity does the job. When market quantity is greater than efficient scale, as in Figure 3, there is no uniform (nondiscriminating) price that can satisfy both sustainability and cost recovery. On the other hand, a variety of nonuniform price structures can achieve the desired goals. One simple form for such pricing would be to give each buyer (or class of buyers) a distinct price. While this approach may not be practical in all circumstances, it is used here for illustrative purposes.

As defined above, sustainable prices generate total revenue that is necessarily no greater than total cost, since total cost is the stand-alone cost for the whole market. Sustainability, however, does not rule out prices that yield total revenue less than total cost. The Federal Reserve Banks operate under the requirement, from the Pricing Principles developed by the Board of Governors pursuant to the Monetary Control Act, that revenues be sufficient to at least cover all costs (the cost-matching requirement). Adding this condition to sustainability necessarily results in revenues that exactly match total costs.

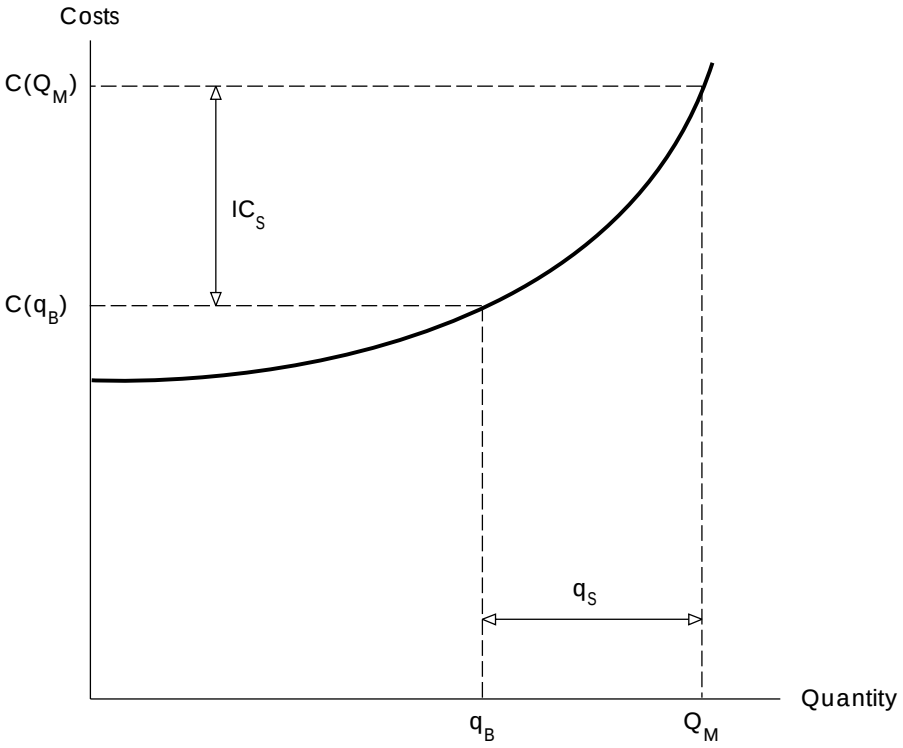
One implication of cost-matching, sustainable pricing is that at least one buyer must be given a price lower than stand-alone cost. Suppose, in the example of Figure 3, that buyer B is charged its stand-alone cost, in the form of a per-unit price of $AC_B = [F + v(q_B)]/q_B$. If both buyers are to be served, the revenue that needs to be collected from buyer S in order to just recover total costs is

$$[F + v(q_B + q_S)] - [F + v(q_B)]. \tag{2}$$

Here, the first term is the total cost of serving both customers, while the second term is the stand-alone cost of serving the large-volume customer. The difference between these two terms is referred to as the *incremental* cost of serving customer S. This cost is denoted by IC_S in Figure 4. Hence, the revenue needed from buyer S can be collected with a per-unit price equal to IC_S/q_S. If buyer S is charged anything less than this price, then in order to recover costs, the seller must charge more than AC_B. If B is charged more than AC_B, a competitor will take B's business.

It is important to note that incremental cost, as the term is used here, is not the same as marginal cost. The former, as indicated by equation (2), is the cost of providing a particular quantity to a particular buyer, given the quantity being provided to other buyers. The latter is simply the cost of providing an additional unit of the product, without regard to the identity of the recipient. It

Figure 4 Incremental Cost

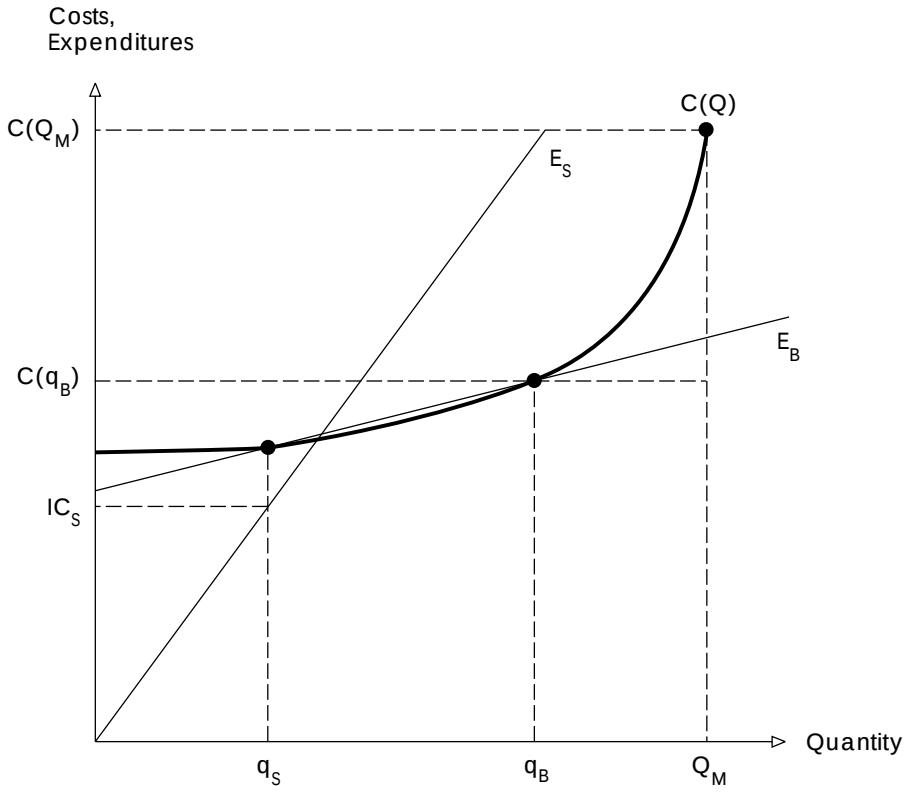


Note: $C(q_B)$ is the stand-alone cost of serving buyer B, while $C(Q_M)$ is the total cost of serving the entire market. The difference between the two, denoted IC_S, is the incremental cost of serving buyer S.

is possible to have a pricing structure in which the (marginal) price charged to some buyers is less than marginal cost while no buyer's average price is less than its average incremental cost.

The price discrimination just described requires that buyers be segmented into groups according to some observable characteristic. This task may not always be straightforward. For instance, the quantity of services used by an institution may be subject to significant change over time. In that case, setting a price to a buyer based on the buyer's previous behavior may not yield the desired results of tailoring prices to current demand conditions. Fortunately, the desired segmentation can typically be achieved by pricing schedules that allow buyers to self-select into groups. One example is "option pricing," in which buyers are given a choice between a schedule with a high fixed charge and low fee per unit and a schedule with a low fixed charge and high fee per unit. For the two-buyer example, Figure 5 illustrates the total expenditure

Figure 5 Option Pricing



Note: The lines labeled E_S and E_B give the total expenditures (as a function of quantity purchased) resulting from buying services under the two alternative options. Under one option, given by E_S, the buyer pays no fixed fee and pays a per-unit price of $P_S = IC_S/q_S$, which is the slope of the line E_S. This price generates expenditures by buyer S equal to incremental cost. The other option, given by E_B, includes a positive fixed fee and a lower per-unit price (slope). The key features of the schedule E_B are that it meets the total cost curve $C(Q)$ at the quantity q_B and that it lies below E_S at q_B. Hence, buyer B prefers the schedule E_B and has a total expenditure equal to stand-alone cost, $C(q_B)$. The fixed fee in the schedule E_B must be (and is, as drawn) high enough so that buyer S prefers the schedule E_S.

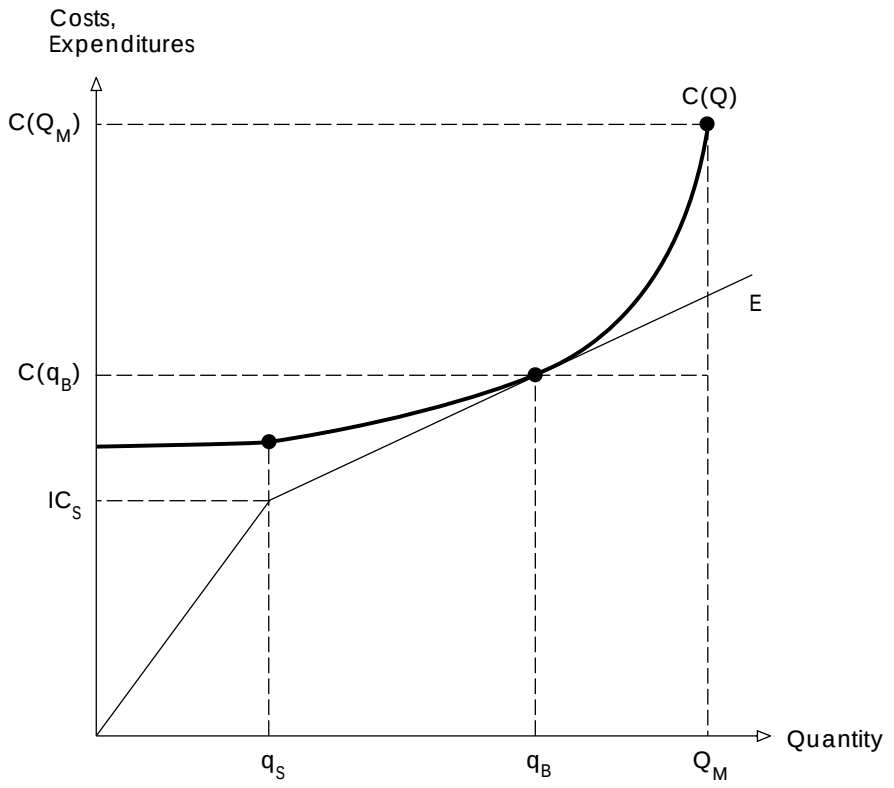
schedules generated by such pricing options. In this example, the low fixed fee is actually set at zero and combined with a per-unit price of $P_S = IC_S/q_S$. An individual buyer will choose the option for which its own total expenditures are the smallest. Hence, the low-volume buyer selects the low fixed charge and high per-unit fee, which is constructed so that total expenditures by buyer S just cover incremental costs. Buyer B selects the other schedule, resulting in expenditures equal to stand-alone cost.

The desired results also could be achieved by a pricing schedule with volume discounts, in which the per-unit fee varies with the quantity purchased.

The total expenditure schedule from one such price structure is given in Figure 6. This price schedule sets a price of IC_S/q_S for the first q_S units purchased. For each additional unit purchased beyond that threshold, the buyer pays the lower price of $[C(q_B) - IC_S]/(q_B - q_S)$. As Figure 6 indicates, this volume-discounting scheme is designed to collect exactly stand-alone cost from buyer B and incremental cost from buyer S.

An option pricing scheme, like that presented in Figure 5, has recently been adopted on a trial basis by some Federal Reserve Banks for some check-processing services. Most prices for Reserve Bank services, as discussed in Section 1 above, are a simpler form of two-part pricing, including a single fixed fee and a single per-item charge. This form of pricing is not as flexible

Figure 6 Volume Discount



Note: Like Figure 5, Figure 6 presents a pricing arrangement that results in expenditures equal to stand-alone cost for buyer B and incremental cost for buyer S. This arrangement involves no fixed fee. The per-unit price is IC_S/q_S on the first q_S units purchased. For each subsequent unit purchased, a lower price is charged. This lower price is equal to $[C(q_B) - IC_S]/(q_B - q_S)$. The resulting total expenditure, as a function of quantity purchased, is given by the kinked line, *E*.

as an option pricing schedule. Accordingly, sustainability may or may not be achievable with simple two-part pricing.

In summary, sustainability requires that no buyer (or set of buyers) faces prices greater than stand-alone costs. Adding the cost-matching condition requires that no buyer (or set of buyers) faces prices less than incremental costs. That is, each buyer must face a price between stand-alone and incremental cost. Within these restrictions, the relative pricing to different buyers can be treated in a variety of ways. In the examples above, buyer B is charged stand-alone cost and buyer S is charged incremental cost. Consequently, all of the fixed costs of production are allocated to buyer B. This allocation could be reversed, charging stand-alone cost to buyer S and incremental cost to buyer B. Price schedules that allocate some fixed costs to each buyer face both buyers with prices in between stand-alone and incremental costs. The concept of sustainability, by itself, gives no guide to the choice among these alternatives. The next section suggests that the possibility of technological differences among alternative providers of a service can help to sharpen the choice.

4. BYPASS AND TARGETED COMPETITION

The forgoing development of sustainable prices assumes that the same technology is available both to the incumbent firm (the Fed) and to any potential or actual competitors. Hence the relevant cost standard for deterring entry is the stand-alone cost of a market segment. There may be instances in which some segment of the market can be served with a technology different from that used by the incumbent seller. In such cases, the term "stand-alone cost," as defined above, is somewhat of a misnomer. This cost is the cost to a buyer, or group of buyers, of obtaining services from a source with a cost structure identical to the incumbent's. When the alternative to the incumbent's network involves a substantially different technology, the buyer's option is not so much to "stand alone" as to "bypass" the network.

An example, as described by Einhorn (1987), is in the provision of long-distance telephone services. Most long-distance calls are routed through the local telephone company, for which a charge is assessed. Large-volume callers, however, may exercise the option to bypass the local company and connect directly with their long-distance provider. The technology for bypassing the local network and, therefore, the costs associated with doing so are different from those associated with connecting through the network.

One can also think of obtaining payment services from alternative sources as bypassing the Federal Reserve network. For instance, a local check clearinghouse utilizing centralized exchange of items involves a different pattern of sorting and transportation expenditures from that arising in the use of the Fed's check-clearing services. Of course, in this regard, the most stark example of bypassing the Fed's network is direct bilateral exchange of payments.

The presence of a bypass alternative places additional constraints on the pricing choices facing the incumbent seller. Continuing with the example of Section 2, suppose that there are two buyers with quantities demanded of q_S and q_B and the sum of these quantities is denoted Q_M. Buyers B and S can bypass the incumbent and receive the service at a cost of C_B^* and C_S^*, respectively. The incumbent seller's total cost of serving the entire market is $C_M = F + v(Q_M)$, while buyer B (S) alone can be served by the incumbent for the stand-alone cost of $C_B = F + v(q_B)$ ($C_S = F + v[q_S]$). Suppose that the bypass technology is potentially attractive only to a large-volume buyer, so that $C_B^* < C_B$ but $C_S^* > C_S$. On the other hand, suppose that the market is still a natural monopoly. This is so if when buyer B bypasses the system, total costs rise, or $C_M < C_S + C_B^*$. This last statement is equivalent to saying that the incumbent's incremental cost of serving buyer B is less than the bypass cost, since $IC_B = C_M - C_S < C_B^*$.

Under the conditions just described, the incumbent is limited in how much of the common fixed costs can be allocated to buyers with a viable bypass option. Recall that allocating all of the (common) fixed costs to a single buyer amounts to charging that buyer its stand-alone cost. Here, such a price to buyer B would induce B to bypass the incumbent, to the detriment of market efficiency.

The incumbent's pricing problem is further complicated if there is some uncertainty about the viability of the bypass technology. Suppose, for instance, that the incumbent is reasonably sure that $C_S^* > C_S$ but is uncertain as to the value of C_B^*. If buyer B's bypass cost is so low that it is less than the incremental cost of serving B, then it is efficient to let buyer B bypass. Indeed, in this case, there are no prices that the incumbent can set to cover costs and guarantee against the loss of buyer B. On the other hand, it is still possible to price in such a way that buyer B will be lost only if bypass is efficient. Specifically, pricing to buyer B at incremental cost and buyer S at stand-alone cost will succeed in covering costs regardless of whether B is retained. Further, by comparing its bypass and incremental costs, B makes its choice in a way that minimizes the total (social) costs of serving the market.

To summarize, the presence of bypass options that are potentially attractive to some individual buyers or groups of buyers limits the ways in which fixed costs can be recovered from buyers. Especially when the value of the bypass option is not fully known by the incumbent, prices to segments of the market that are likely to have the most attractive bypass options should be pushed down to the incremental cost of serving those segments. Such an allocation of fixed costs is likely not to coincide with the allocations implied by standard accounting practices, and it may strike some (especially other buyers) as inequitable. It is important, however, to consider the alternative. If the incumbent seeks to recover some of the fixed cost from likely candidates for bypass, those buyers may turn to alternative sources even when it is socially inefficient to do so. If their business is lost, the incumbent will still have to recover fixed

costs from the remaining buyers. This result would be less cost-efficient and no more equitable than the result of sustainable pricing that recovers all fixed costs from those buyers with the least attractive alternatives.

5. ADDITIONAL COMMENTS ON SUSTAINABLE PRICES

While this article has presented sustainable pricing as a tool for evaluating pricing from a public policy point of view, the concept originated as a predictive notion in the theory of "contestable markets."[14] Contestable markets theory holds that in the presence of potential competition, incumbent firms will not be able to charge anything other than sustainable prices; any attempt to charge unsustainable prices would quickly prompt entry of and loss of market share to a competing seller. In other words, even in a natural monopoly, there are no economic rents earned by an incumbent monopolist. This is a strong conclusion that has not been broadly accepted without qualification. Most importantly, one cannot discuss the effects of potential entry without considering the likely response to entry by the incumbent seller. In its purest form, contestable markets theory assumes that the incumbent can alter its price in response to entry only with some lag. The incumbent's inability to respond quickly leaves an opportunity for an entrant to capture, at least temporarily, some part of the market should the incumbent's prices be unsustainable. Ultimately, the incumbent may regain the market, but the absence of sunk costs implies that even temporary profit opportunities will be exploited by entrants.

In an unregulated market populated only by private firms, there is little reason to suppose that firms do not have a great deal of flexibility in adjusting their prices to competitive conditions. The situation of a Reserve Bank, however, may come closer to that imagined by the contestable markets theory. Clearly, the process necessary to adjust pricing policy is time-consuming. Reserve Banks must set and publish prices once each year. Further, the Board of Governors' Pricing Principles, adopted pursuant to the Monetary Control Act, state that substantive changes in the structure of prices or services offered shall be made subject to public comment. Volume-based pricing for check services (on a limited basis) was approved by the Board in November 1993 and became effective in January 1994, "subject to additional staff analysis and public comment" (Board of Governors of the Federal Reserve 1994a).

When price adjustment is subject to lags, then prices that are not sustainable can attract entry, even if the market cannot efficiently support the additional seller(s) in the long run. Hence, the use of unsustainable prices can attract excessive entry when entrants can take advantage of an incumbent's administrative delays in responding to competition.

[14] See Baumol, Panzar, and Willig (1982).

It is also useful to compare sustainable prices to a pricing concept often used in discussions of regulatory price setting. In such discussions, one approach is to seek prices that maximize social welfare, subject to a zero-profit constraint for the seller. The "social welfare" to be maximized is a measure of the benefits (e.g., utility or profits) received by buyers. The resulting prices are referred to as *Ramsey* prices, because their derivation follows Ramsey's (1927) formulation of optimal taxation. Sustainability is a stronger constraint than zero profits. Hence, Ramsey prices will not, in general, coincide with sustainable prices. Accordingly, the former might be more applicable to the problem of setting prices in the public interest when an incumbent seller is protected from competition by legal barriers to entry.

Unlike Ramsey prices, the notion of sustainability used here is entirely cost-based; it does not take into account a measure of the benefits generated by the provision of payment services. A cost-based specification of sustainability is exact when demands for services are assumed to be perfectly inelastic. The more general specification would require that the net value provided to any group of buyers (benefits to buyers less payments to seller) be no less than the greatest net value those buyers could obtain from an alternate source. While this generalization is a direct extension of the basic idea, measures of benefits on the demand side of a market may be difficult to obtain. Hence, the cost-based notion of sustainability may remain useful as a practical approximation to the more general concept.

6. ARE PAYMENT SERVICES MARKETS NATURAL MONOPOLIES?

Sustainable pricing is presented above in the context of a market that is a natural monopoly. Neither of the markets discussed above, check and ACH services, is a monopoly, although the ACH market comes close. Even the market for check services, however, is fairly concentrated; in any given geographic region, the Federal Reserve serves a significant share of the market for intraregional processing. Market structure is determined in part by the degree of scale economies relative to the size of the market. Hence, a concentrated market is likely to be one in which demand and technology conditions are such that only a small number of sellers is viable. In such a market, the analysis of sustainable pricing closely parallels that of natural monopoly.

The analysis offered in this article does assume that a seller's efficient scale is at least a sizeable fraction of the size of the market. Hence the applicability of the pricing principles proposed above is partly an empirical matter. Specifically, what evidence exists on the significance of scale economies? There have been a number of studies of the Federal Reserve's check-collection services,

aimed at addressing this question.[15] These studies tend to find fairly weak scale economies in the observed range of production levels.[16] Such findings might seem at odds with the narrative description of the experience in check processing (and in ACH services), which seems to parallel the analysis of Section 2; average-cost pricing to the market as a whole led to the defection of high-volume users of the services. One possible conclusion is that the alternative means used by defecting customers do, in fact, deliver the services with lower real resource costs. That is, these users may have access to a superior bypass technology. In that case, the Fed's loss of market share would be efficiency-enhancing. On the other hand, the analysis of Figure 3 refers to a case in which the incumbent operates above efficient scale. If this case were an accurate description of the Fed priced-services environment, then one would not expect to find empirical evidence of widespread, unexploited economies of scale.

Aside from economies associated with check processing, there may be scale efficiencies in the distribution and transportation of processed checks. Fixed costs that are specific to each endpoint served may result in markets where efficient scale is a sizeable fraction of the relevant market.

There is also the possibility that economies exist less in increasing the scale of production of any given service than in the joint provision of multiple services. This is the most common use of the term "economies of scope." For instance, a single electronic connection to a Reserve Bank can allow a customer to use ACH services and other electronic services, including new electronic check-collection options. To the extent that scope economies exist, it may not make sense to talk about market structure, pricing, and cost recovery on a product-by-product basis. The concept of sustainable prices, however, can be directly extended to an environment with economies of scope. Consider the pricing of an array of services. For such pricing to be both sustainable and cost-matching, no service to any buyer or group of buyers can be priced above stand-alone cost or below incremental cost. Here, stand-alone cost is the cost of providing only a specific subset of the services to a specific subset of the buyers. Similarly, incremental cost refers to the added cost of a specific subset of services to a subset of buyers, given the services already being provided to other buyers. As before, choices among sustainable price configurations amount to choices among possible allocations of common fixed costs across buyers and services. If a particular service is targeted for competition (for instance, because of the availability of a bypass technology specific to that service), then that service's price should be set at incremental cost.

Even if the structure of cost and demand is such that these markets are not natural monopolies, the concept of sustainable prices can still provide a useful

[15] A recent example is Bauer and Hancock (1992).

[16] If, as these studies suggest, the average-cost curve is relatively flat at its minimum, then average-cost pricing should be close to sustainability.

benchmark for Reserve Bank pricing policy. The cost structure in a market might be such that the efficient number of sellers is greater than one but still small. For instance, if the market quantity sold tends to be about three times the efficient scale of production, then the efficient number of sellers is three. In such a "natural oligopoly," pricing behavior tends to be the result of a complicated dynamic game. Here, the administrative structure that governs Reserve Bank pricing can be advantageous in that it may give the Federal Reserve the ability to precommit to a pricing strategy over a long horizon.[17] When the Fed is one of several competitors, it can contribute to the efficiency of the market by adopting a clear pricing policy to which other sellers can react. Specifically, the Fed could make it known that it stands ready to sell to any market segment at no greater than stand-alone cost and no less than incremental cost. Within these bounds, it will adjust pricing to respond to competition, moving prices in more competitive segments toward incremental cost. Such a strategy makes it clear that market gains by competitors that reduce overall social costs will not be contested, while those that raise costs will not be accommodated. Under sustainable pricing, a seller cannot preserve market share that is not justified by its technological capabilities.

7. SUSTAINABLE PRICING AND THE MONETARY CONTROL ACT

The move toward market-sensitive pricing that responds to competitive conditions might raise questions about the role of the Federal Reserve in the provision of payment services. To what extent should a Reserve Bank behave like a private business? Does an attempt by a Reserve Bank to maintain its share of the market interfere with its public policy objectives with regard to the payment system? To the latter question, the discussion in this article suggests the answer, "Not necessarily." By letting its prices be guided by the notion of sustainability, the Federal Reserve establishes a benchmark for the market place. If competition targets a particular segment of the market, that segment should be served by the Federal Reserve at incremental cost. Then, any gains in market share by competitors will also be in the public interest. It is also worth noting that no market segment is being subsidized by another as long as no price is less than incremental cost.

It is important to note that the pricing behavior suggested herein is, in many cases, not the behavior one would expect from a private business. That is, the resulting pricing is not the pricing that would prevail in the market if the Fed played no operational role. Private businesses are motivated by long-run profit

[17] A treatment of the benefit of precommitment in oligopoly pricing games can be found in Tirole (1989).

maximization. This may lead to deviations from sustainable prices in a number of ways. First, an incumbent firm facing potential entry can set prices to any market segment above stand-alone cost, as long as the incumbent has adequate flexibility to adjust its prices in response to entry. In other words, revenues can more than cover costs. Second, there may be situations in which a private firm will be willing and able to set prices that fail to recover all costs in the short run. Suppose, for example, that two firms find themselves in competition in a market that has the cost and demand structure of a natural monopoly.[18] In the long run, only one of the firms can remain in the market. To determine which firm will survive, the two might engage in a "war of attrition" in which prices are below costs and losses are incurred until one firm chooses to exit. The short-run pricing would necessarily involve some prices to some market segments below incremental cost. While the Monetary Control Act does allow the Fed to have revenues that fall below costs in the short run, the Board of Governors has adopted the policy of setting prices each year with the aim of recovering all anticipated costs for that year (Board of Governors of the Federal Reserve 1994b).

Unlike the pricing behavior of a private business, market-sensitive, sustainable pricing is motivated not by profit maximization, but by an interest in the overall efficiency of the market for payment services. This motivation drives pricing as close as possible to the "first-best" result of marginal-cost pricing of all products to all buyers. The constraints that keep pricing away from that goal are the need to cover costs and the need to ensure that market share is lost only when the loss results in lowering the resource costs of serving the entire market.

Does a pricing policy that results in disparate treatment of banks conflict with the goals of Congress in writing the pricing requirement into the Monetary Control Act? The language of the Act instructs the Federal Reserve to "give due regard to . . . the adequate level of [services] nationwide." Since the sustainable pricing schemes outlined above tend to involve average and marginal prices that decline with the volume of services purchased, it appears that such pricing will favor large institutions, because small banks would pay a higher average price. Hence, disparate treatment, in the form of higher average prices, might be thought of as impeding smaller institutions' access to services. The language in the Monetary Control Act could conceivably be interpreted as prohibiting pricing that faces some institutions with a greater cost of access to services. In the presence of economies of scale or scope, pricing that achieves equal treatment of all buyers and just recovers costs is typically not sustainable. Hence, if the Monetary Control Act is interpreted strictly as mandating

[18] For instance, a market that could previously support two firms might experience a permanent decline in demand.

equal treatment, the Federal Reserve could find itself in an intractable bind; if uniform, unsustainable prices result in significant loss of business, the Reserve Banks could have difficulty covering costs without raising prices to remaining buyers. The result would be equal treatment by the Fed but disparate treatment by the market as a whole.

While the language of the Monetary Control Act may or may not be read as providing a mandate for equal treatment, it does seem to dictate a continued role for the Fed in the provision of payment services. Without such a legislated dictum, one might legitimately wonder whether there is a necessary role for the Fed in these markets. Indeed, the central result in the theory of contestable markets, as noted above, is that the force of potential competition among private businesses is sufficient to yield sustainable prices. On this point, experience in deregulated transportation and telecommunication markets has been inconclusive. These markets tend to be highly concentrated, and strategic interaction may tend to result in fluctuation between collusive and aggressively competitive behavior. In such an environment, it is conceivable that a single large provider committed to a sustainable pricing policy could provide a stabilizing influence on the market while promoting an efficient market structure.

8. CONCLUSION

This article proposes a general principle for evaluating Reserve Bank pricing strategies. The concept of sustainable pricing under conditions of scale and scope economies appears to be a useful tool. Sustainable prices that just cover total costs price all services to all customers in between their stand-alone and incremental costs. When competition from private-market providers is focused on a subset of services and customers, sustainability retains enough flexibility to respond to competitive pressures by pushing some prices down to incremental cost. This response is particularly appropriate in conditions of uncertainty about competitors' costs.

A strategy of market-sensitive sustainable pricing would result in loss of business to competitors only when such loss is efficient. Hence, this strategy provides a guideline for responding to competition in a way that respects the requirements of the Monetary Control Act while promoting efficiency in the delivery of payment services. If the Federal Reserve is going to be in the payment services business, it should use its position as a provider motivated by the public interest to guide the market in the direction of efficiency. Sustainable prices provide market participants with a benchmark for assessing the cost effectiveness of alternative modes of service delivery. Following this benchmark may or may not stem the Fed's loss of market share, but maintaining market share should not be a goal of Fed policy. The Fed's market share should be whatever is consistent with the efficient operation of the payment system.

REFERENCES

Bank for International Settlements. *Payment Systems in Eleven Developed Countries.* Bank Administration Institute, 1993.

Bauer, Paul, and Diana Hancock. "The Efficiency of the Federal Reserve Payments System," Working Paper. Washington: Board of Governors of the Federal Reserve, 1992.

Baumol, William J., John C. Panzar, and Robert D. Willig. *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovic, 1982.

Blommestein, Hans J., and Bruce J. Summers. "Banking and the Payments System," in Bruce J. Summers, ed., *The Payment System: Design, Management and Supervision*. Washington: International Monetary Fund, 1994.

Board of Governors of the Federal Reserve System. *80th Annual Report,* 1993. Washington: Board of Governors, 1994a.

_____. *Federal Reserve Regulatory Service,* Vol. III. Washington: Board of Governors, 1994b, locator number 7-134.

Einhorn, Michael A. "Optimality and Sustainability: Regulation and Intermodal Competition in Telecommunications," *Rand Journal of Economics,* vol. 18 (Winter 1987), pp. 550–63.

Goodfriend, Marvin S. "Money, Credit, Banking and Payment System Policy," in David B. Humphrey, ed., *The U.S. Payment System: Efficiency, Risk and the Role of the Federal Reserve*. Boston: Kluwer Academic Publishers, 1990, pp. 247–277.

McAndrews, James. "The Automated Clearinghouse System: Moving Toward Electronic Payment," Federal Reserve Bank of Philadelphia *Business Review,* July/August 1994, pp. 15–23.

Ramsey, Frank P. "A Contribution to the Theory of Taxation," *Economic Journal,* vol. 37 (March 1927), pp. 47–61.

Spulber, Daniel F. *Regulation and Markets*. Cambridge, Mass.: MIT Press, 1989.

Tirole, Jean. *The Theory of Industrial Organization.* Cambridge, Mass.: MIT Press, 1989.

United States General Accounting Office. "Check Collection: Competitive Fairness Is an Elusive Goal," Report to Congressional Committees, May 1989.

Were Bank Examiners Too Strict with New England and California Banks?

Robert M. Darin and John R. Walter

> Massachusetts Gov. Michael Dukakis accused the Comptroller of the Currency of "enforcing stricter standards in New England than in the rest of the country." . . . New England's elected officials are . . . concern[ed] that regulators are pushing their once vibrant region into a recession by forcing banks to *increase loan reserves* [emphasis added], which, in turn, is causing them to tighten credit standards.
>
> [T]here is widespread concern that the medicine might be worse than the disease. Bankers fear that regulators who were heavily criticized for not acting quickly when Texas banks were collapsing are now overreacting in New England.
>
> In a reprise of the kind of regulatory crackdown already experienced in the East, California bankers report that federal agencies . . . have been harsh this year.
>
> *American Banker*

During the early 1990s bank examiners were frequently accused of being too strict with banks in New England, thereby contributing to a credit crunch in the region.[1] If supervisors of New England banks were being unusually strict, they may have been reacting to public complaints of lax supervision of the savings and loan industry in the 1980s. Such complaints were rife as New England banks' loan problems were surfacing. As California's economy began a slowdown and banks there began to experience significant loan losses, examiners of California banks also were accused of being too strict. Unusually strict examination practices could have contributed

The views expressed are those of the authors and do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] For reports of examiner strictness, see *American Banker,* April 20, 1990, p. 1; April 25, 1990, p. 1; and August 16, 1991, p. 1; *The Economist,* April 7, 1990, p. 94; or *The Wall Street Journal,* April 12, 1990, p. A16.

to the large declines in bank loans and the severity of the economic downturns in New England and California.[2] Several studies have found evidence that the large declines in bank lending in New England were, in part, the result of constraints on bank lending imposed by regulatory capital standards (Peek and Rosengren 1992, 1993; Bernanke and Lown 1991).[3] These studies focus on whether capital constraints faced by New England banks during that region's economic troubles produced declines in bank lending. The capital constraints in many cases resulted from large additions to reserves for loan losses. The studies make no attempt to determine if bank examiners were inappropriately strict in the amount of additions to reserves for loan losses they required of banks, though Bernanke and Lown (1991) do briefly examine supervisory strictness and conclude that New England banks were not subject to overzealous supervision.

This article looks for evidence of excessive examiner strictness as manifested in the amount of reserves for loan losses New England and California banks were required to maintain. Here, strictness refers to the required level of reserves for loan losses relative to expected loan losses. While requiring banks to maintain a certain level of reserves for loan losses is only one of several ways examiner strictness can manifest itself, it is one of the most important. Allowing banks to hold reserves for loan losses that are too small relative to expected future losses, or, equivalently, allowing them to overvalue their loan portfolios, may increase bank failure costs borne by the deposit insurance fund. On the other hand, excessive strictness may lead to unnecessary cutbacks in bank lending. Such indeed is the contention of those criticizing examiners of New England and California banks. To test for loan loss reserve account strictness, we compare the ratio of reserves for loan losses to nonperforming loans for New England and California banks to the average ratio for all U.S. banks. If examiners were being unusually strict in the amount of loan loss reserves they required of New England and California banks, the ratio for these banks should have exceeded that of the average U.S. bank at the time of the hypothesized strictness. We also examine how the average ratio for New England and California banks changed in the periods before and during the hypothesized strictness. If examiners were unusually strict, the ratios for these banks should have increased to unusually high levels compared with past years. Last, we compare the ratio for banks in New England and California to the ratio for banks affected by oil-industry problems of the mid-1980s. If

[2] According to a 1991 survey, 56 percent of surveyed small banks in northern and central California indicated that they had denied loans during the year because of the strict regulatory environment. The Western Independent Bankers Association and the Secura Group conducted the survey, and the *American Banker* reported the results in its November 20, 1991, issue.

[3] Peek and Rosengren (1993) go so far as to conclude that their evidence suggests that "New England did suffer from a regulatory-induced credit crunch" (p. 28).

examiners were unusually strict with New England and California banks during the 1990s, the ratios for New England and California banks should exceed those of banks in the oil-industry-dependent region during economic difficulties. We also broaden our measure of supervisory strictness beyond the simple reserves-to-nonperforming ratio and test again for signs of examiner strictness.

Section 1 deals with bank problem-loan accounting and notes how examiner strictness may influence reported results. In Section 2 we describe our measures of examiner strictness. In Section 3 we report the results of our analysis using the measures mentioned above. According to such measures, we find little evidence that supervisors were too strict with banks in New England and California. To the contrary, we find that banks in New England and California seem to have received relatively lenient treatment. Finally, in Section 4 we examine some possible reservations to our analysis.

1. EXAMINER STRICTNESS AND BANK ACCOUNTING FOR PROBLEM LOANS

One category of problem-loan data is *reserves for loan losses.* Reserves for loan losses are reported by all banks to federal regulators in quarterly financial statements known as "call reports," more technically named "Consolidated Reports of Condition and Income," which consist of a balance sheet, an income statement, and other financial information. The primary function of the reserve for loan losses account is to adjust the reported value of the loan portfolio for expected future credit losses. A bank's reserve for loan losses should equal its, or its examiner's, best estimate of the dollar value of expected losses of principal on its portfolio of loans. If a bank maintains its reserve account at a level equal to this estimate, then total loans less reserves, or net loans, is the best estimate of the collectible value of the loan portfolio. On bank financial statements, net loans are added to other assets to arrive at total assets. The reserve account is established and maintained by periodic charges to an expense account denoted "provision for loan losses."[4]

Additions to the loan loss reserve account, like other expenses, reduce net income. Under normal circumstances, a bank's operating income is sufficient to cover additions to loan loss reserves and other expenses. Sometimes, as occurred at many New England banks during the late 1980s, additions to loan loss reserves exceed income. In such cases, adding to loan loss reserves reduces capital.

It seems likely that pressures on examiners to be strict or lenient will manifest themselves in reserves for loan losses required of banks. During bank examinations, examiners verify the adequacy of loan loss reserves and often require banks to increase the size of the account. Examiners exercise a good deal

[4] See Walter (1991) for further discussion of loan loss reserves.

of judgment and discretion when determining what constitutes an adequate level of loan loss reserves. Such judgment is necessary because many bank loans are heterogeneous and the signals of impending loan losses vary from loan to loan. But it leaves scope for examiner decisions to be influenced by pressures to be strict or lenient. New England bank examiners were criticized for excessive strictness in the early 1990s. Such strictness was attributed to fears of repeating past mistakes. On the other hand, throughout much of the 1980s, many observers expressed concerns that examiners were being too lenient with banks that held nonperforming less-developed-country (LDC) loans. Such banks were seen as holding loan loss reserves that were low relative to expected losses on the loans. Bank supervisors may have thought that by giving them additional time to collect nonperforming loans or to supplement reserves for loan losses, the banks would be able to avoid shrinking their loan portfolios or even failing. Supervisors also may have been under some political pressure to "go easy" on LDC-exposed banks. Had examiners forced the LDC-exposed banks to quickly add reserves to cover expected loan losses, the necessary additions to reserves could have virtually eliminated the equity of some of these banks (Mengle and Walter 1991). Ultimately, the exposed banks made large additions to reserves for LDC loans beginning in 1987.

Another category of banks' problem-loan data is *nonperforming loans*. According to federal bank regulatory definitions, nonperforming loans ("past-due and nonaccrual loans" on bank call reports) are those for which the borrower is 30 days or more late on contracted interest or principal payments and those on nonaccrual status. Loans 30 days or more late are further classified as 30 to 90 days past due and 90 days or more past due. Regulators require banks to stop accruing interest on loans, or place them on nonaccrual status, if the borrower's financial condition has deteriorated, if payment in full is not expected, or if the loan has been in default 90 days or more.[5] Few loans are placed on nonaccrual status unless they are past due, since the first sign that the financial condition of the borrower has deteriorated or that payment in full is not expected generally is the failure to make timely interest or principal payments.

Verifying the appropriateness of the loan loss reserve account during examinations typically involves a significant amount of examiner judgment and discretion. Little discretion, however, is involved in determining whether or not a loan should be reported as nonperforming. For most loans, if the borrower is current on interest and principal payments, the loan is not reported as nonperforming. If, on the other hand, the borrower is more than 30 days past

[5] A loan 90 days or more late generally must be placed on nonaccrual status unless (1) it is a consumer installment loan, (2) it is secured by a mortgage on a one- to four-family property, or (3) it is well secured and in the process of being collected. Loans that are 90 days or more late that fall under one of the excluded categories are reported as "loans past due 90 days or more."

due, the loan will be reported as nonperforming. Occasionally loans may be placed on nonaccrual status even though they are not past due. The examiner or bank may believe that even though the borrower is current on payments, the borrower may be unable ultimately to repay the entire loan. In such cases, nonperforming loans will be enlarged based on examiner or bank discretion.

The final category of problem-loan data discussed is loan *charge-offs*. When it is apparent that all or a portion of a loan will be uncollectible, the loan is charged off. The amount of the charge-off will equal the book value of the loan when the bank or its examiner believes the loan is likely to be a total loss. The charge-off will be less than book value when the bank or its examiner believes that some of the loan's principal value will be recovered, say, from foreclosure on collateral. When a charge-off is taken, some or all of the book value is removed from the bank's books and the same amount is deducted from the reserve for loan losses account. In most cases, loans more than 180 days past due are charged off. On the other hand, there is a good deal of bank or examiner judgment involved in charging off a loan that is less than 180 days past due. Any recovery of an amount previously charged off is added to the reserve balance upon its collection.

2. MEASURES OF EXAMINER STRICTNESS

As discussed earlier, examiners have considerable latitude to determine the appropriate level of loan loss reserves, so that pressures to be more or less strict may influence the amount of reserves held. Ideally, a test for excessive examiner strictness would compare the bank's loan loss reserve to a knowledgeable but impartial party's estimate of future loan losses. Using this test, the examiner's strictness would be measured by the ratio of the bank's reserves to the impartial party's loss estimate. If the ratio is significantly less than one, the bank has underreported reserves and its examiner may have been too lenient. If the ratio is approximately one, then the bank has properly reported reserves and its examiner has been fair. If the ratio is significantly greater than one, then the bank has overreported reserves, possibly because the bank's examiner has been too strict. While bank financial statements report loan loss reserve figures, they do not report impartial loan loss estimates. In our analysis, we use banks' reported nonperforming loans as a proxy for the impartial party's estimate of future loan losses.[6]

[6] While many researchers count as nonperforming only those loans past due 90 days or more and those in nonaccrual status, our measure of nonperforming loans also includes loans past due 30 to 90 days. We have chosen to be more inclusive because we believe that the component consisting of loans 30 to 90 days past due provides additional information about future loan losses. Our empirical results are not dependent on including this component.

We choose the nonperforming loans figure as a proxy because it is unlikely to be influenced by examiner strictness yet is likely to be highly correlated with an impartial party's estimate of future loan losses. As discussed earlier, the amount of reported nonperforming loans is subject to little examiner judgment. Thus, like the impartial party's loan loss estimate, it is unlikely to be influenced by pressures on examiners to be lenient or strict. Since nonperforming loans are known to be troubled, when the amount of such loans held by the bank increases, an impartial party would increase his estimate of eventual loan losses for most banks. For all U.S. banks, from 1983 to 1993, nonperforming loans and net charge-offs (charge-offs less recoveries on previously charged-off loans) during the following four quarters were highly correlated, with a correlation coefficient of 0.87. Other research supports the hypothesis that nonperforming loans have power in predicting future losses. Berger, King, and O'Brien (1991) regress charge-offs on loan loss reserves and nonperforming loans, using data for all U.S. banks from 1982 through 1989. They conclude that "the nonperformance measures [nonperforming loans] add significantly to the information about future bank performance beyond loan loss reserves" (p. 769).[7] In related work, Avery, Hanweck, and Kwast (1985), Hirschhorn (1986), and Cole and Gunther (1993) find that nonperforming loans help predict bank failures.

Unfortunately, we cannot simply examine the average ratio of reserves to nonperforming loans for a region and conclude that if the ratio is greater than one, the region's examiners were unusually strict, and if the ratio is less than one, they were unusually lenient. Typically the ratio is significantly lower than one because a portion of nonperforming loans is likely to be completely or partially repaid and only the remainder will result in a loss. A priori, we do not know what levels of the reserves-nonperforming loans ratio indicate that examiners have been "lenient," "fair," or "strict" for a given bank or group of banks. Instead, to draw conclusions regarding examiner strictness we analyze the reserves-nonperforming loans ratio for New England and California banks relative to the ratio for three control groups. First, we compare the ratio for New England and California banks to the ratio for all U.S. banks in the same time period. We assume that examiners were fair for the average of all U.S. banks. Second, we compare the reserves-nonperforming loans ratio for New England and California banks to past years' average levels of the ratio. In doing so, we

[7] The Berger, King, and O'Brien measure of nonperforming loans differs slightly from our measure of nonperforming loans. Berger, King, and O'Brien include as nonperforming loans those past due 90 days or more, those on nonaccrual status, and renegotiated loans. Renegotiated loans are those loans for which the bank has reduced interest or principal payments because of the deterioration of the financial position of the borrower. We exclude from our analysis renegotiated loans but include loans past due 30 to 90 days. Any differences in results should be minor because Berger, King, and O'Brien estimate that renegotiated loans have a relatively small, and in some of their regressions insignificant, influence on later loan charge-offs, and because, as noted earlier, our results are largely unchanged by the inclusion of loans past due 30 to 90 days.

assume that before troubled times, examiners were fair with New England and California banks. Finally, we compare the ratio for New England and California banks to the ratio for banks in the "oil region" during the period of that region's economic difficulties. We compute these ratios relative to the U.S. average. As for oil-region banks, we assume their examiners were fair, or at least not strict, since no complaints of such strictness were heard at the time of distress. Now any one of these assumptions alone may be subject to question. But if all three comparisons point to the same conclusion about examiner strictness, then we can be fairly confident of our conclusions. We would conclude that there is evidence that examiners in New England or California were unusually strict if the loan loss ratio for banks in these regions was significantly above (1) the ratio for all U.S. banks, (2) past levels of the ratio for New England and California, and (3) the ratio for the oil region.

We also test for examiner strictness using a second measure, the ratio of loan loss reserves to loan charge-offs occurring later (RES_t/CO_{t+i}). The ratio allows us to avoid a potential bias caused by a change in the definition of nonperforming loans. Approximately when loan problems of New England banks reached their peak, examiners began to require more frequently that banks place loans current on principal and interest payments on nonaccrual status. Before then, examiners only infrequently required banks to report any current loans on nonaccrual status. Current loans placed on nonaccrual status were commonly referred to as "performing nonperforming loans." Typically, such loans were suspect because collateral values had fallen significantly or because there was some indication that the borrower would be unable to make continued payments. Examiners generally required extra reserve backing for these loans. Because a performing borrower is more likely than a nonperforming borrower to repay a loan, the amount of loan loss reserves for performing nonperforming loans should be somewhat lower than for other nonperforming loans. Thus, the reserves-nonperforming loans ratio may have a downward bias beginning when examiners increased the frequency with which they declared performing loans nonperforming. Unfortunately, bank financial statements do not segregate performing nonperforming loans, so we cannot adjust for the bias. The RES_t/CO_{t+i} measure avoids this bias since it does not employ nonperforming loans at all.

The RES_t/CO_{t+i} ratio also allows us to test the robustness of our conclusions regarding nonperforming loans. That is, it provides a measure that requires no proxy of an impartial party's estimate of loan losses. This could be important because when using nonperforming loans as a proxy for an impartial party's estimate of loan losses, we in effect assume that a dollar of nonperforming loans always leads an impartial examiner to require each and every bank to hold approximately the same level of reserves. If this is not true—in other words, if a dollar in nonperforming loans leads an impartial examiner to require fewer reserves in one region than in others—then our reserves-nonperforming loans ratio may give us biased results. One can imagine, for example, that in a

region that has experienced a perceived temporary economic shock, impartial examiners might require lower loan loss reserves per dollar of nonperforming loans than in other regions. Since we cannot directly test the accuracy of nonperforming loans as a proxy, testing for examiner strictness with a measure that is not dependent on this proxy provides the best opportunity to test the robustness of our strictness conclusions. If RES_t/CO_{t+i}-based strictness conclusions confirm those from the reserves-nonperforming loans measure, then we can be more certain of the robustness of our conclusions.

3. ANALYSIS USING MEASURES OF EXAMINER STRICTNESS

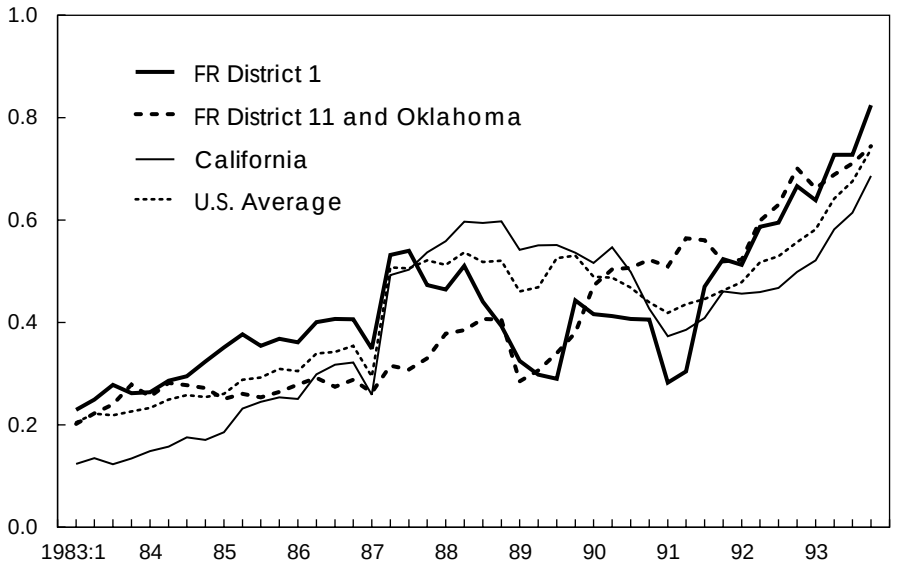
Figure 1 displays weighted-average reserves-nonperforming loans ratios for banks in New England (**Federal Reserve District 1**), in the "oil region" (**FR District 11 and Oklahoma**), in California, and throughout the United States (observations are quarterly).[8] The weighted average is the sum of loan loss reserves for all banks in a region divided by the sum of all nonperforming loans for all such banks.[9] The figure shows that during the periods when examiners were supposedly too strict, New England and California banks' average reserves-nonperforming loans ratios were not unusually high relative to (1) the U.S. average, (2) past levels achieved in the two regions, or (3) the experience of the oil region. On the contrary, New England and California reserves-nonperforming loans ratios were somewhat low.

Figure 1 shows that after remaining above the U.S. average ratio from 1983 until 1987, New England's reserves-nonperforming loans ratio fell below the U.S. average ratio and remained well below it until 1991, when it rose slightly above it. The period when the ratio was low relative to the U.S. average corresponds with New England's economic troubles, which were worst between 1987 and 1992. Therefore, New England banks' reserves-nonperforming loans ratio was low for some years before examiners were criticized for unusual strictness (mostly in 1990). Even after the New England ratio rose above the U.S. average ratio, it tracked that average fairly closely. Figure 1 also shows that as economic difficulties were hitting California in 1990, the California

[8] Our "oil region" (**FR District 11 and Oklahoma** in Figure 1) includes banks in Oklahoma, from Federal Reserve District 10, and banks in Federal Reserve District 11. This combination means that banks in states most affected by petroleum-industry problems are grouped together.

[9] We display weighted results in these graphs because displaying the average of individual banks' ratios produces results that might be distorted by a small number of banks with very few nonperforming loans. The reserves-nonperforming loans ratios of these banks are extremely high because they had almost no nonperforming loans but over time maintained a significant amount of loan loss reserves, producing ratios as high as 1700.

Figure 1 Ratio of Reserves to Nonperforming Loans



reserves-nonperforming loans ratio fell below the U.S. average ratio. It remained slightly below the U.S. average ratio through 1993.[10]

In 1990, at the time of the hypothesized strictness, New England banks' reserves-nonperforming loans ratio was below its 1987 level and only about equivalent to the level reached in 1985 and 1986. The ratio did begin to increase rapidly in 1991, but it did not regain its 1987 level until late 1991. Likewise, California banks' average reserves-nonperforming loans ratio did not return to

[10] The dips in the **FR District 11 and Oklahoma** line in Figure 1 beginning in the first quarter of 1989 and in the **FR District 1** line beginning in the first quarter of 1991 are the result of the fairly unique way the FDIC handled some large bank failures in the late 1980s and early 1990s. Failures of large banking companies in Texas and New England (First Republic Bank Corp. of Texas, MCORP of Texas, Texas American, First American Bank and Trust of Texas, and Bank of New England) were handled for an interim period, usually less than a year, by placing the assets and deposits of failed banks in "bridge banks" set up and owned by the FDIC, until a buyer could be found. These bridge banks included on their books most of the failing banks' nonperforming loans but minimum loan loss reserves, so that their presence in our Figure 1 data set causes the reserves-nonperforming loans ratio to be low for the periods of their existence. Removing bridge banks from the data used to construct Figure 1 eliminates most of the 1989 dip in the **FR District 11 and Oklahoma** line and all of the 1991 dip in the **FR District 1** line and causes the **FR District 1** line to rise above the **U.S. Average** line one quarter earlier (second quarter rather than third quarter 1991) than shown in Figure 1. Other than these changes, eliminating the "bridge bank effect" leaves Figure 1 essentially unchanged.

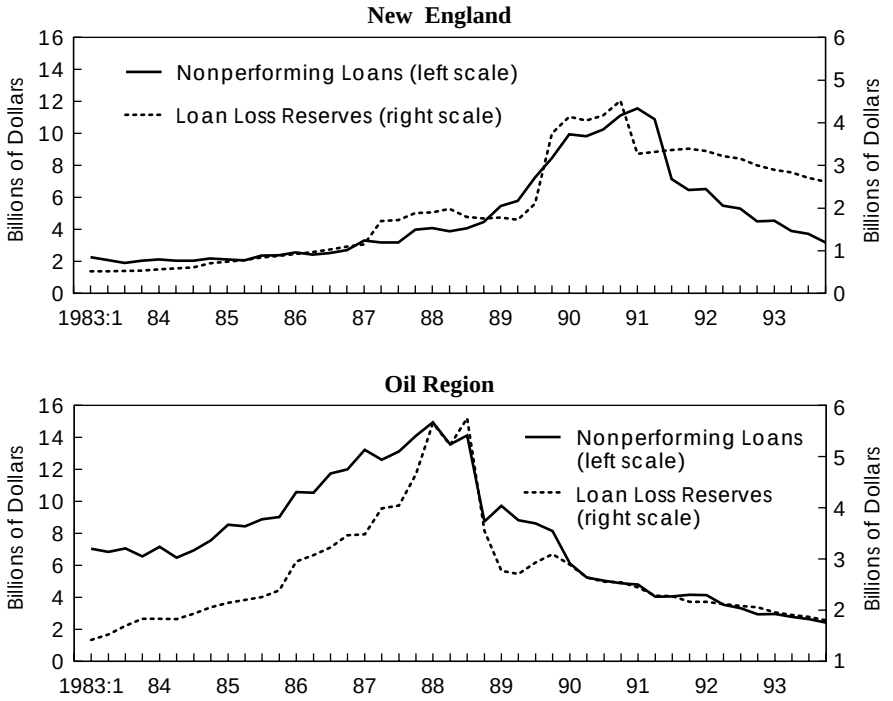
levels attained in 1990 until 1993, well after examiner strictness was supposed to have occurred. Therefore, at the time of the hypothesized strictness, banks in New England and California had fairly low reserves-nonperforming loans ratios relative to pre-recession levels.

The oil region's economic problems were worst in the 1984 through 1989 period. As shown in Figure 1, the reserves-nonperforming loans ratio of the region's banks did not fall until 1989, one year after nonperforming loans reached their peak. In contrast to the experience of banks in New England and California, oil-region banks showed little decline in the ratio of reserves to nonperforming loans during troubles in the region. Though their reserves-nonperforming loans ratio was below the U.S. average ratio throughout much of the latter half of the 1980s, they maintained a fairly consistent increase throughout the 1980s and early 1990s.

One of the most dramatic features of Figure 1 is the increase in the reserves-nonperforming loans ratio in the second quarter of 1987. Virtually all of the increase is accounted for by $18 billion of additions to loan loss reserves made by large banks to provide for anticipated losses on LDC loans (McLaughlin and Wolfson 1988). These additions began on May 20, 1987, when Citicorp added $3 billion to loan loss reserves to cover expected LDC loan losses. Following the Citicorp addition, other large banks throughout the country made sizable additions to loan loss reserves to cover expected future losses on LDC loans. As shown in Figure 1, the ratios for New England banks and California banks also jumped in the second quarter of 1987, as banks in these areas also made large additions for LDC loans. The ratio did not increase significantly in the oil region because exposure to LDC debt was minimal in that region.

Even though New England and oil-region banks were in similar shape in terms of the percentage of loans that were troubled, banks in these two regions displayed very different behavior as nonperforming loans first began to increase. Figure 2 displays the dollar amount of (1) nonperforming loans and (2) reserves for loan losses for New England and the oil region. On average, oil-region banks added substantially to reserves as soon as nonperforming loans began to rise in 1984. In New England, by contrast, loan loss reserves remained essentially unchanged as nonperforming loans almost doubled between mid-1987 and late 1989. In the first seven quarters of that period, these loans increased 82 percent in New England while reserves rose only 2 percent. Conversely, comparable seven-quarter figures for the oil region show that nonperforming loans relative to assets rose 63 percent while reserves rose 62 percent. Banks in New England did not add significantly to reserves for almost two years after the onset of rising problem loans. As a fraction of nonperforming loans, New England banks' large quarterly additions to reserves in late 1989 and 1990 were significantly greater than any quarterly additions made by oil-region banks. Additions made by banks in New England were viewed as evidence that examiners were being too stringent. But they only brought the reserves to nonperforming loans ratio

Figure 2 Loan Loss Reserves and Nonperforming Loans



at New England banks up to the level of oil-region banks at a comparable point in that region's fortunes.

Differing reactions in New England and the oil region may have resulted from different signals of future losses available to the two regions. The collapse of OPEC and oil prices in the early 1980s may have given oil-region banks and their examiners early and clear warning of long-lasting loan problems in that region, leading them to make early additions to loan loss reserves. In New England, on the other hand, signs of persistent loan problems may have become clear only as more and more loans became nonperforming.

When compared to the average of all U.S. banks, weighted-average reserves-nonperforming loans ratios for New England and California banks provide no evidence of unusual examiner strictness. Indeed, they give some indication of examiner lenience, especially in the period before the hypothesized strictness. Likewise, there is no evidence of unusual examiner strictness when comparing reserves-nonperforming loans ratios for New England and California during the periods of hypothesized examiner strictness to the average

ratios generated by banks in these regions before the onset of their economic troubles. Finally, in comparison to results produced by banks in the oil region, New England and California banks do not appear to have been treated strictly.

Figure 1 seems to point fairly consistently to the conclusion that New England and California banks were not forced to add excessively to reserves and might have even been treated leniently by examiners. Seeking additional confirmation, we employed regression analysis to determine whether regions were statistically significantly different from the average for all U.S. banks. We ran regressions using as dependent variables the log of the quotient of two ratios, namely, reserves to nonperforming loans for individual banks and for the average of all U.S. banks. We employed as independent variables dummies for banks' regions. Expressed this way, our regression counts each bank's individual reserves-nonperforming loans ratio equally, regardless of the size of the bank. The regression equation appears as follows:

$$\log(\text{RATIO}_{it}/\text{RATIO}_{\text{U.S.}t}) = B1 * RG_1 + B2 * RG_2 + \cdots + B13 * RG_{13} + e_{it}$$

$$\text{RATIO}_{it} = \frac{(RES_{\text{Q1}} + RES_{\text{Q2}} + RES_{\text{Q3}} + RES_{\text{Q4}})}{(NPL_{\text{Q1}} + NPL_{\text{Q2}} + NPL_{\text{Q3}} + NPL_{\text{Q4}})}$$

$=$ Bank i's average reserves-nonperforming loans ratio in year t.[11]

$\text{RATIO}_{\text{U.S.}t} =$ Arithmetic average of all U.S. banks' RATIO_i in year t.

The independent variables are all dummy variables: RG_d is a dummy variable equal to one if a bank is in Federal Reserve District number d and zero otherwise. The state of California is entered as regional dummy 13 (and is excluded from Federal Reserve District 12) because California was especially plagued by the recession of the early 1990s, while other Twelfth District states were relatively better off. $d = 1, 2, \ldots 13$.

The regression equation was run once for each year 1983 through 1993. Because banks, or their supervisors, may take several quarters to adjust the reserve account in response to a change in nonperforming loans, reserves-nonperforming loans ratios were calculated using annual averages. Since every region was represented by a dummy, constants were omitted from the regression. Table 1 displays the results of these regressions. The coefficient on each region's dummy is a measure of how location influences the deviation of a bank's reserves-nonperforming loans ratio from the U.S. average ratio. The t-statistics are test statistics for the hypothesis that the region dummy coefficients equal zero. In other words, they test the hypothesis that there is

[11] Banks that do not produce call reports for all four quarters in a year, either because of failure, merger, or de novo entry, are removed from the regression calculation for the year. The "bridge bank effect" (discussed in footnote 10) therefore does not influence our regression results.

no relationship between location in the region and the deviation of a bank's reserves-nonperforming loans ratio from the comparable U.S. average ratio.

The regression results corroborate the trends apparent in Figure 1. Banks in New England and California had low reserves-nonperforming loans ratios compared to banks nationwide around the time of their economic troubles and during and after the periods examiners were criticized for being excessively strict. Specifically, the regressions show very large (in absolute value) and statistically significant negative coefficients for New England and California during these periods. Banks in Federal Reserve District 1 (New England) had a coefficient of $-.68$ in 1990, near the trough of the New England recession. This was by far the largest absolute coefficient of any region in any year. The reserves-nonperforming loans ratio for banks in New England fell significantly below the average ratio for all U.S. banks in 1988, soon after loan troubles began to surface in New England. The ratio then declined further through 1990. It began recovering in 1991 but remained significantly below the U.S. average through 1993. The coefficients for California also became very highly negative in the early 1990s. In contrast, the lowest coefficient registered in the oil region (Federal Reserve District 11 and Oklahoma) was $-.23$. Until 1987 the reserves-nonperforming loans ratio for banks in the oil region was significantly above or only slightly below the U.S average. From 1987 through 1990, the reserves-nonperforming loans ratio for oil-region banks was statistically significantly lower than the average ratio for all U.S. banks, although the absolute value of the coefficient for the oil region was much smaller than that of coefficients for New England and California.[12]

The regression analysis provides evidence that New England and California banks were not forced by excessively strict examiners to overreserve. It shows that New England and California banks had much lower reserves relative to nonperforming loans than average for all U.S. banks before, during, and after the time examiners were being criticized for excessive strictness. The analysis also shows that relative to the U.S. average, underreserving was much greater in New England and California than it had been earlier in the oil region. The regressions do indicate that some underreserving during economic troubles may be normal, since it seems to have occurred in New England, California, and in the oil region. Such could be the case because it may take some time for banks to recognize and set aside income for problem loans, or for examiners to examine banks and force them to increase reserves for loan losses.

When we further investigate examiner strictness using our second strictness measure, the ratio of loan loss reserves today to loan charge-offs tomorrow

[12] One might conjecture that different size banks may reserve for loan losses in different ways, on average. If this is the case, then our results may have been influenced by differences in the size distribution of banks across the regions. To test for this, we regressed our ratio on size dummies and found no consistent relationships.

Table 1

Regression equation: $\log(\text{RATIO}_{it}/\text{RATIO}_{\text{U.S.}t}) = B1 * RG_1 + B2 * RG_2 + \cdots + B13 * RG_{13} + e_{it}$

Regional Dummies	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**1992**	**1993**
Fed Reserve District 1	0.12	0.19	0.21	0.23	0.19	−0.23	−0.62	−0.68	−0.56	−0.41	−0.35
	(2.19)[b]	(3.62)[c]	(4.08)[c]	(4.54)[c]	(3.64)[c]	(−4.10)[c]	(−10.96)[c]	(−11.71)[c]	(−9.09)[c]	(−6.10)[c]	(−4.99)[c]
Fed Reserve District 2	0.15	0.17	0.28	0.27	0.15	−0.10	−0.29	−0.43	−0.35	−0.38	−0.41
	(2.68)[c]	(3.09)[c]	(5.44)[c]	(5.57)[c]	(3.08)[c]	(−1.91)[a]	(−5.59)[c]	(−8.35)[c]	(−6.83)[c]	(−7.05)[c]	(−7.35)[c]
Fed Reserve District 3	−0.18	−0.08	0.01	0.04	−0.04	−0.16	−0.23	−0.35	−0.38	−0.29	−0.34
	(−3.40)[c]	(−1.69)[a]	(0.20)	(0.81)	(−0.94)	(−3.15)[c]	(−4.66)[c]	(−7.21)[c]	(−7.87)[c]	(−5.97)[c]	(−6.51)[c]
Fed Reserve District 4	−0.15	−0.08	−0.04	−0.05	−0.10	−0.18	−0.23	−0.23	−0.23	−0.18	−0.13
	(−4.00)[c]	(−2.04)[b]	(−1.08)	(−1.51)	(−2.97)[c]	(−4.88)[c]	(−6.15)[c]	(−6.08)[c]	(−6.03)[c]	(−4.67)[c]	(−3.23)[c]
Fed Reserve District 5	0.06	0.10	0.10	0.11	0.01	−0.05	−0.14	−0.20	−0.26	−0.19	−0.18
	(1.47)	(2.86)[c]	(2.83)[c]	(3.40)[c]	(0.36)	(−1.33)	(−3.93)[c]	(−5.87)[c]	(−7.60)[c]	(−5.55)[c]	(−4.88)[c]
Fed Reserve District 6	0.07	0.07	0.03	0.00	−0.08	−0.11	−0.17	−0.23	−0.24	−0.16	−0.11
	(3.11)[c]	(3.26)[c]	(1.22)	(0.02)	(−3.59)[c]	(−4.86)[c]	(−7.16)[c]	(−9.89)[c]	(−10.68)[c]	(−7.07)[c]	(−4.56)[c]
Fed Reserve District 7	−0.03	−0.04	0.00	0.11	0.20	0.22	0.19	0.18	0.12	0.06	0.03
	(−1.42)	(−2.38)[b]	(0.12)	(6.61)[c]	(11.98)[c]	(12.51)[c]	(10.61)[c]	(9.91)[c]	(6.39)[c]	(2.96)[c]	(1.45)
Fed Reserve District 8	−0.04	−0.04	−0.02	0.03	0.00	−0.03	−0.02	−0.10	−0.11	−0.06	0.01
	(−1.46)	(−1.74)[a]	(−0.94)	(1.30)	(0.14)	(−1.38)	(−0.92)	(−4.18)[c]	(−4.68)[c]	(−2.49)[b]	(0.30)
Fed Reserve District 9	−0.26	−0.32	−0.34	−0.31	−0.16	−0.01	0.04	0.08	0.09	0.10	0.07
	(−10.26)[c]	(−13.70)[c]	(−15.30)[c]	(−14.48)[c]	(−6.93)[c]	(−0.32)	(1.56)	(3.11)[c]	(3.66)[c]	(3.78)[c]	(2.60)[c]

Fed Reserve District 10 (excluding Oklahoma)	0.16 (7.45)[c]	0.07 (3.26)[b]	0.10 (5.13)[c]	0.11 (6.20)[c]	0.18 (9.72)[c]	0.27 (13.22)[c]	0.35 (16.79)[c]	0.38 (18.23)[c]	0.45 (21.32)[c]	0.40 (18.59)[c]	0.38 (16.80)[c]
Fed Reserve District 11 (including Oklahoma)	0.15 (7.71)[c]	0.19 (10.65)[c]	0.08 (4.99)[c]	−0.07 (−4.31)[c]	−0.20 (−12.12)[c]	−0.23 (−12.27)[c]	−0.18 (−8.68)[c]	−0.05 (−2.36)[b]	0.04 (2.07)[b]	0.03 (1.22)	−0.03 (−1.36)
Fed Reserve District 12 (excluding California)	−0.48 (−9.40)[c]	−0.40 (−8.30)[c]	−0.38 (−8.17)[c]	−0.31 (−6.88)[c]	−0.29 (−6.02)[c]	−0.29 (−5.54)[c]	−0.17 (−3.20)[c]	−0.03 (−0.53)	0.08 (1.61)	0.14 (2.61)[c]	0.29 (5.36)[c]
California	−0.24 (−4.81)[c]	−0.08 (−1.93)[a]	0.04 (1.05)	0.01 (0.24)	0.06 (1.50)	0.04 (1.01)	0.10 (2.36)[b]	0.05 (1.16)	−0.26 (−6.42)[c]	−0.43 (−10.15)[c]	−0.45 (−10.16)[c]
F-Statistic	29.76	33.96	31.60	32.74	39.76	45.79	58.94	68.67	72.26	54.70	46.24
F-Significance	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Number of Observations	13,889	13,841	13,870	13,715	13,221	12,643	12,269	11,925	11,556	11,140	10,674

[a] Significant at the 10 percent level.

[b] Significant at the 5 percent level.

[c] Significant at the 1 percent level.

Notes: Banks that did not produce call reports for all four quarters in a year were removed from the regression calculation for the year. Otherwise, regressions include all U.S. banks. t-statistics are in parentheses.

(RES_t/CO_{t+i}), our nonperforming loans-based results are confirmed. Analysis using RES_t/CO_{t+i} indicates that New England and California banks might have received lenient treatment at the hands of their examiners, or at least did not receive overly strict treatment.

Figure 3 displays the weighted-average RES_t/CO_{t+i} ratio graphed by region. The figure shows that New England banks' ratio fell below the U.S. average ratio beginning in 1988 and remained below it until early 1991, when New England's ratio moved slightly above the U.S. average ratio.[13,14] The figure also shows that the average ratio for California banks remained significantly above the U.S. average ratio until mid-1990, but then fell below. The average ratio for New England banks began falling in 1986, remained until 1992 well below levels maintained between 1983 and 1986, and never rose above the early 1986 level. California banks' average RES_t/CO_{t+i} ratio shows a similar pattern across time. In the oil-industry region, the ratio remained well below the U.S. average ratio until 1988. Then it either conformed to the average or rose above it throughout the remainder of the region's difficulties. These graphs indicate that compared to their U.S. and oil-region counterparts, as well as themselves in better times, New England and California banks may have been underreserved during much of each region's slowdown.

Table 2 displays results from regressions using RES_t/CO_{t+i} instead of *RES*/*NPL* but otherwise equivalent to the *RES*/*NPL* regression presented earlier. As was the case with the earlier regressions, the largest negative coefficients are associated with the economic difficulties experienced in New England and California in the late 1980s and early 1990s. Significantly negative coefficients continued even as complaints were surfacing of excessive examiner strictness in these regions. Banks in the oil region had RES_t/CO_{t+i} ratios well below the U.S. average ratio throughout the study period, though negative coefficients for the oil region were never as large as they were for New England and California at their worst. These results corroborate results from the nonperforming loans regressions and graphs.

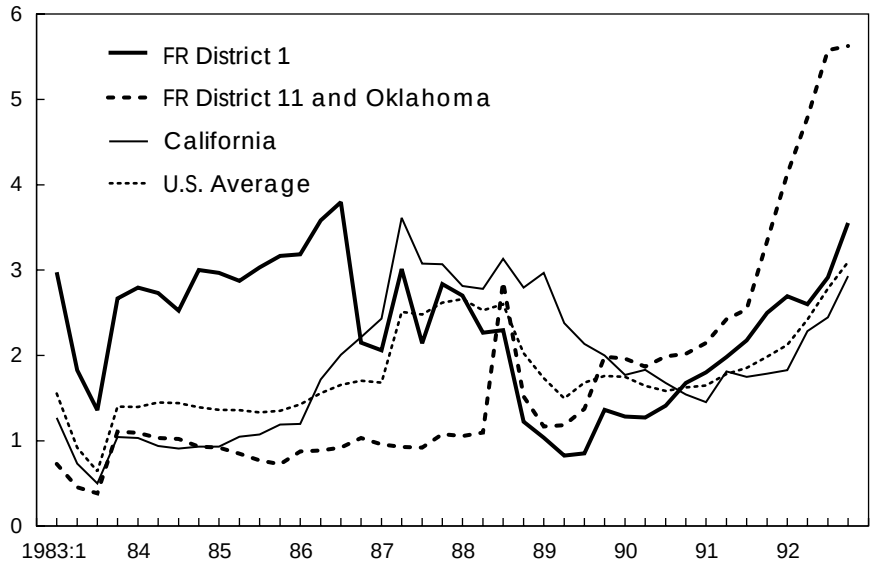
4. CAVEATS

Our regression analyses presented in Tables 1 and 2, and our comparisons of ratios for New England and California banks to the average ratio for all U.S. banks, could incorrectly classify regional examiners as more lenient than they were. If examiners were inappropriately strict in their standards for determining

[13] In our calculations, we excluded any bank that was not operating in all of the five quarters (one quarter when reserves are observed plus the following four quarters when charge-offs for the bank are observed) used to calculate a ratio.

[14] We also compared reserves to charge-offs over the next eight quarters and found similar results.

Figure 3 Ratio of Reserves to Eventual Net Charge-Offs



reserves with *all* U.S. banks, New England and California banks might seem to have received lenient treatment by comparison even though they too were subject to inappropriately strict treatment. Bizer (1993) finds evidence that U.S. bank supervisors, on average, became stricter in their confidential bank ratings after 1989 as compared with before 1989. No studies exist of examiner strictness in standards for determining loan loss reserves. However, stories appearing around 1990 in the banking press reported a perception among bankers and borrowers that these standards were made stricter for banks throughout the country, not just in New England and California. We believe it unlikely that our measures incorrectly indicate lenient treatment of New England and California banks.

For one thing, our data do not support the conclusion that examiners throughout the United States became stricter in loan loss reserves standards in 1989 or 1990. To be sure, the U.S. average reserves-nonperforming loans ratio line shown in Figure 1 reaches a local minimum in the first quarter of 1991 and rises consistently afterward. But it does not rise to unusually high levels until 1993. Similarly, in Figure 3 the U.S. average RES_t/CO_{t+i} line reaches a local minimum in 1989, remains relatively flat until 1991, and then begins rising. True, as of the end of 1992, it does rise above the previous highs achieved in 1987 and 1988, but not far above. It therefore seems unlikely that unusual examiner strictness occurred for the average U.S. bank before late 1992.

Table 2
Regression equation: $\log(\text{RATIO}_{it}/\text{RATIO}_{\text{U.S.}t})$, where RATIO = $(\text{loan loss reserves})_t/(\text{net charge-offs})_{t+i}$

Regional Dummies	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**1992**
Fed Reserve District 1	1.11	1.35	1.31	0.99	0.61	−0.42	−1.19	−1.14	−0.71	−0.55
	(13.23)[c]	(16.78)[c]	(16.48)[c]	(12.10)[c]	(7.36)[c]	(−5.07)[c]	(−14.10)[c]	(−12.35)[c]	(−7.34)[c]	(−5.23)[c]
Fed Reserve District 2	0.75	1.12	1.10	0.76	0.38	0.04	−0.64	−0.64	−0.47	−0.56
	(9.09)[c]	(14.28)[c]	(14.42)[c]	(9.56)[c]	(4.87)[c]	(0.53)	(−8.30)[c]	(−8.34)[c]	(−5.95)[c]	(−6.66)[c]
Fed Reserve District 3	1.05	1.22	1.31	0.99	0.82	0.52	0.10	−0.13	−0.16	−0.05
	(13.10)[c]	(16.09)[c]	(17.87)[c]	(13.06)[c]	(10.80)[c]	(7.21)[c]	(1.42)	(−1.82)[a]	(−2.22)[b]	(−0.58)
Fed Reserve District 4	0.40	0.55	0.50	0.36	0.15	−0.05	−0.09	−0.11	−0.16	−0.13
	(6.98)[c]	(10.26)[c]	(9.50)[c]	(6.72)[c]	(2.77)[c]	(−0.90)	(−1.55)	(−1.99)[b]	(−2.82)[c]	(−2.04)[b]
Fed Reserve District 5	0.67	0.74	0.74	0.46	0.34	0.12	−0.15	−0.25	−0.23	−0.02
	(11.79)[c]	(14.08)[c]	(14.61)[c]	(9.01)[c]	(6.56)[c]	(2.27)[b]	(−2.91)[c]	(−4.94)[c]	(−4.47)[c]	(−0.31)
Fed Reserve District 6	0.14	0.15	0.10	−0.08	−0.21	−0.29	−0.37	−0.43	−0.28	−0.22
	(3.95)[c]	(4.52)[c]	(3.06)[c]	(−2.27)[b]	(−6.01)[c]	(−8.41)[c]	(−10.87)[c]	(−12.46)[c]	(−8.01)[c]	(−5.87)[c]
Fed Reserve District 7	0.16	0.05	0.19	0.46	0.53	0.50	0.46	0.32	0.27	0.18
	(5.91)[c]	(1.84)[a]	(8.03)[c]	(17.93)[c]	(19.26)[c]	(17.56)[c]	(15.69)[c]	(10.88)[c]	(9.08)[c]	(5.31)[c]
Fed Reserve District 8	0.05	0.04	0.11	0.22	0.22	0.18	0.13	0.09	0.08	0.13
	(1.34)	(1.12)	(3.38)[c]	(6.41)[c]	(6.04)[c]	(4.88)[c]	(3.36)[c]	(2.33)[b]	(2.13)[b]	(2.92)[c]
Fed Reserve District 9	−0.08	−0.22	−0.28	−0.12	0.07	0.17	0.27	0.42	0.37	0.35
	(−2.25)[b]	(−6.62)[c]	(−8.61)[c]	(−3.58)[c]	(1.90)[a]	(4.54)[c]	(6.89)[c]	(10.72)[c]	(8.89)[c]	(7.65)[c]

Fed Reserve District 10 (excluding Oklahoma)	−0.50 (−15.98)[c]	−0.53 (−18.32)[c]	−0.35 (−12.59)[c]	−0.25 (−8.71)[c]	−0.10 (−3.08)[c]	0.04 (−1.14)	0.16 (4.68)[c]	0.26 (7.46)[c]	0.31 (8.19)[c]	0.16 (3.84)[c]
Fed Reserve District 11 (including Oklahoma)	−0.39 (−13.91)[c]	−0.32 (−12.66)[c]	−0.55 (−23.00)[c]	−0.71 (−28.93)[c]	−0.79 (−29.55)[c]	−0.64 (−22.33)[c]	−0.31 (−10.07)[c]	−0.09 (−2.84)[c]	−0.07 (−2.00)[b]	−0.08 (−2.17)[b]
Fed Reserve District 12 (excluding California)	−0.06 (−0.84)	0.05 (0.78)	0.03 (0.45)	−0.18 (−2.47)[b]	−0.25 (−3.36)[c]	−0.17 (−2.19)[b]	−0.14 (−1.63)	−0.04 (−0.44)	−0.05 (−0.63)	−0.01 (−0.12)
California	−0.28 (−3.87)[c]	0.03 (0.45)	0.16 (2.66)[c]	0.18 (2.97)[c]	0.13 (2.01)[b]	0.32 (4.97)[c]	0.29 (4.43)[c]	−0.25 (−3.89)[c]	−0.68 (−10.97)[c]	−1.03 (−15.22)[c]
F−Statistic	87.61	124.13	150.51	141.73	122.36	79.76	65.31	55.90	41.83	35.16
F−Significance	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Number of Observations	12,363	12,744	12,726	11,718	10,750	10,400	10,106	9,860	9,172	8,109

[a] Significant at the 10 percent level.

[b] Significant at the 5 percent level.

[c] Significant at the 1 percent level.

Notes: Banks that did not produce call reports for all five quarters (one quarter when reserves are observed plus the following four quarters when charge-offs are observed) used to calculate a year's ratio and banks that had no net charge-offs were removed from the regression calculation for the year. Otherwise, regressions include all U.S. banks. Year indicates time period when reserves were held. t-statistics are in parentheses.

Second, in comparing ratios for New England and California banks to ratios for control groups other than the U.S., we find that our measures also indicate no unusual strictness, and possibly lenience, for New England and California banks. As noted earlier, the reserves-nonperforming loans and RES_t/CO_{t+i} ratios achieved by New England banks in the late 1980s and early 1990s remained at or below the levels produced by New England banks before the 1987 decline. This indicates that unless examiners were inappropriately strict with New England banks in 1986 and 1987, they apparently were not in 1990 and 1991. A similar argument may be made for California banks. The reserves-nonperforming loans ratios for the oil region and New England were approximately equivalent at similar times during their difficulties. (Nonperforming loans peaked in the third quarter of 1988 for the oil region and the first quarter of 1991 for New England.) So, unless examiners were inappropriately strict in reserve standards for oil-region banks, they apparently were not with New England banks.

While it seems unlikely that comparisons with U.S. bank averages bias our analysis, our reserves-nonperforming loans ratio may understate the degree of examiner strictness in another way. A bank with an unusually large ratio of loan charge-offs to nonperforming loans could have a low ratio of reserves to nonperforming loans even though it is not underreserved and has not undergone unusually lenient examination. The high charge-off bank is likely to have a relatively low reserves-nonperforming loans ratio for two reasons. First, if the nonperforming loans charged off tend to be those with the greatest expected losses and therefore those with the greatest proportion of reserves, it is likely that charge-offs will lower the proportion of reserves to nonperforming loans. Second, when a portion of a nonperforming loan is charged off, the remainder of the nonperforming loan may have a lower-than-normal expected loss and require few loan loss reserves.

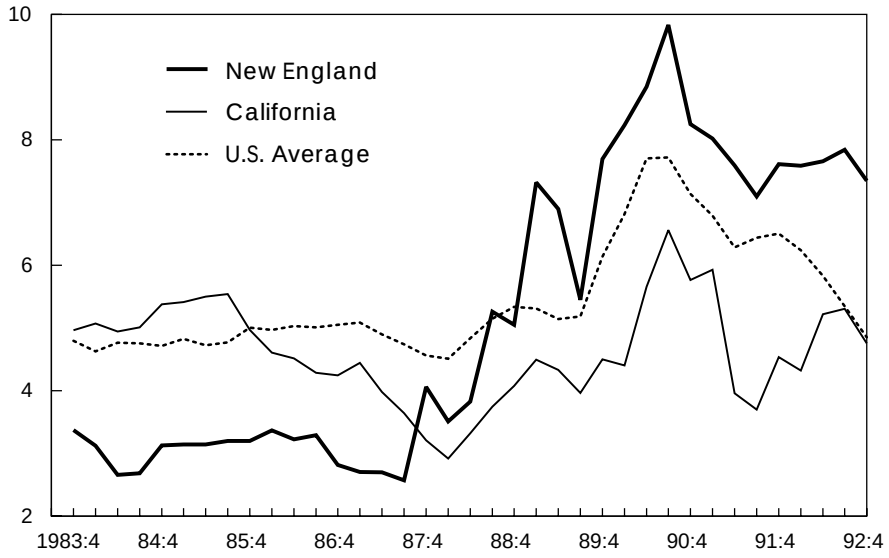
The bias in the reserves-nonperforming loans measure that can occur when loan charge-offs are unusually large can be minimized by modifying the reserves-nonperforming loans ratio. Loan charge-offs are added back to reserves for loan losses and to nonperforming loans so that the measure of examiner strictness becomes $(RES+CO)/(NPL+CO)$. This modification reverses the bias introduced by loan charge-offs on reserves and nonperforming loans. When the regressions presented in Table 1 were rerun with the $(RES + CO)/(NPL + CO)$ ratio substituted for the reserves-nonperforming loans ratio as the dependent variable, coefficients and their significance levels were virtually identical to those generated with the simpler reserves-nonperforming loans ratio. This result indicates that our original reserves-nonperforming loans ratio suffered from little if any bias from unusually large charge-offs. It follows that the reserves-nonperforming loans ratio probably underestimates examiner strictness little.

The RES_t/CO_{t+i} ratio also may understate examiner strictness because of examiner charge-off procedures. Since examiners have some discretion in

determining the required amount of loan charge-offs, it is possible that a tendency to be excessively strict might show up in the amount both of loan loss reserves and banks' charge-offs. If so, then CO_{t+i} in the RES_t/CO_{t+i} ratio would increase, causing that ratio to indicate either a decline or no change in examiner strictness when in fact examiners increased strictness.

Charged-off loan *recovery* data for New England and California banks, however, provides no evidence that examiners were excessively strict in the amount of charge-offs they required. As noted earlier, funds collected on loans previously charged off (for example from the sale of foreclosed properties or from repayments made by delinquent borrowers) are called *recoveries.* Their dollar amounts are reported in quarterly call reports. Excessive charge-off strictness means that examiners are forcing banks to charge off loans that ultimately will be repaid, or to charge off greater proportions of loans than ultimately will be lost on these loans. Therefore, an inappropriate increase in charge-off strictness should lead to an increase in later recoveries. Figure 4 graphs one year's average charge-offs divided by the following year's average recoveries (CO_t/REC_{t+i}) for New England, California, and the United States. Suppose the decline in the New England RES_t/CO_{t+i} line in Figure 3, or the rise in the line

Figure 4 Ratio of Charge-Offs to Eventual Recoveries



Notes: Each charge-off figure is the four-quarter sum of charge-offs in the current quarter and three previous quarters. Each recovery figure is the four-quarter sum of recoveries beginning in the following quarter.

to levels only slightly higher than the U.S. average after 1990, was the result of examiners being unusually strict in charge-off procedures. If so, then one would expect the CO_t/REC_{t+i} line in Figure 4 also to *fall* after 1987 and be unusually *low* relative to the U.S. average. Figure 4, however, shows that New England banks' CO_t/REC_{t+i} line *increased* from 1987 through 1990. Then, consistent with the U.S. average line, the New England line peaked in mid-1990 and declined for several quarters before leveling off well *above* the U.S. average line. California banks follow the same pattern as New England banks and U.S. banks, though from a lower level. The consistently lower-than-U.S.-average CO_t/REC_{t+i} ratio exhibited by California banks could indicate that California banks' examiners consistently applied stricter charge-off requirements than the average for all U.S. banks. It is unlikely then that New England and California banks' RES_t/CO_{t+i} ratio was artificially depressed during economic difficulties in those regions, since examiners apparently were not unusually strict in the charge-offs they required during those periods.

5. CONCLUSIONS

We have developed and examined several measures of supervisory strictness. We find little evidence that bank supervisors were too strict with New England and California banks. To the contrary, by our measures, examiners treated New England and California banks less strictly in times of trouble than the average U.S. bank. Moreover, examiners treated the former banks less strictly than before their economic troubles and less strictly than oil-region banks that suffered similar economic difficulties. These measures, however, provide no evidence that any such leniency by examiners was intentional. Perhaps examiners were surprised by the severity of the New England and California problems, but were less surprised by the severity of problems in the oil region.

It is probably true that the large additions to reserves for loan losses made by banks in New England and California in the early 1990s may have diminished these banks' ability to lend. But our data indicates that those additional reserves at best only made up for an extended period when reserves were too low relative to expected loan losses. It seems unlikely, therefore, that *inappropriate* action by bank examiners exacerbated the effects of the 1990–91 recession in these regions.

REFERENCES

Avery, Robert B., Gerald A. Hanweck, and Myron L. Kwast. "An Analysis of Risk-Based Deposit Insurance for Commercial Banks," in *Proceedings of a Conference on Bank Structure and Competition* (Federal Reserve Bank of Chicago, May 1985), pp. 217–50.

Berger, Allen N., Kathleen Kuester King, and James M. O'Brien. "The Limitations of Market Value Accounting and a More Realistic Alternative," *Journal of Banking and Finance,* vol. 15 (September 1991), pp. 753–83.

Bernanke, Ben S., and Cara S. Lown. "The Credit Crunch," *Brookings Papers on Economic Activity,* 2:1991, pp. 205–39.

Bizer, David S. "Regulatory Discretion and the Credit Crunch," Working Paper. Washington: U.S. Securities and Exchange Commission, April 1993.

Cole, Rebel A., and Jeffrey W. Gunther. "Separating the Likelihood and Timing of Bank Failure," Finance and Economics Discussion Series, no. 93-20. Washington: Board of Governors of the Federal Reserve System, Division of Research and Statistics, June 1993.

Hirschhorn, Eric. "Developing a Proposal for Risk-Related Deposit Insurance," *Banking and Economic Review,* September/October 1986, pp. 3–10.

McLaughlin, Mary M., and Martin H. Wolfson. "The Profitability of Insured Commercial Banks in 1987," *Federal Reserve Bulletin,* vol. 74 (July 1988), pp. 403–18.

Mengle, David L., and John R. Walter. "How Market Value Accounting Would Affect Banks," in *Rebuilding Banking,* Proceedings of the 27th Annual Conference on Bank Structure and Competition (Federal Reserve Bank of Chicago, May 1991), pp. 511–33.

Peek, Joe, and Eric S. Rosengren. "Bank Regulation and the Credit Crunch." Unpublished manuscript. February 1993.

_____. "The Capital Crunch in New England," Federal Reserve Bank of Boston *New England Economic Review,* May/June 1992, pp. 21–31.

Walter, John R. "Loan Loss Reserves," Federal Reserve Bank of Richmond *Economic Review,* vol. 77 (July/August 1991), pp. 20–30.

An Error-Correction Model of the Long-Term Bond Rate

Yash P. Mehra

Most recent studies of long-term interest rates have emphasized term structure relations between long and short rates. They have not, however, looked behind these relations to find the basic economic factors that affect the overall level of interest rates.[1] In this article, I examine empirically the role of economic fundamentals in explaining changes in the long-term U.S. Treasury bond rate.

The economic determinants of the bond rate are identified by building on the loanable funds model used by Sargent (1969), among others.[2] The bond rate equation estimated here, however, differs from the one reported in Sargent in two major respects. First, it uses the federal funds rate rather than the money supply to capture the influence of monetary policy actions on the real component of the bond rate. As is now widely recognized, financial innovations and the deregulation of interest rates have altered the short-run indicator properties of the empirical measures of money. Hence, it is assumed that the impact of monetary policy actions on the real bond rate is better captured by changes in the real funds rate than in the real money supply. Second, it uses cointegration and error-correction methodology, which is better suited to distinguish between the short- and long-run economic determinants of the bond rate than the one used in Sargent and elsewhere.

The views expressed are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] The main reason for this neglect is that the studies in question have been interested primarily in testing the validity of the expectations theory of the term structure of interest rates. One recent exception is the study by Goodfriend (1993), who has attempted to look at fundamentals. In Goodfriend, the long bond rate is viewed as an average of expected future short rates, the latter in turn depending on monetary policy actions and the expected trend rate of inflation. Goodfriend then uses a narrative approach to discuss the interactions between the bond rate and its economic determinants, including monetary policy and expected inflation. He does not, however, formally test for or estimate the impact of these economic determinants on the bond rate.

[2] For example, see Echols and Elliot (1976) and Hoelscher (1986), who have employed variants of this model to study the behavior of the long-term bond rate.

The empirical work presented here suggests several results. First, inflation rather than the deficit appears to be the major long-run economic determinant of the bond rate. The long-run deficit-interest rate link found here in the data is fragile.[3] Second, monetary policy actions measured by the real funds rate have substantial short-run effects on the real bond rate. Third, the bond rate equation estimated here is consistent with the bond rate's actual, long-run behavior from 1971 to 1993. Nevertheless, it fails to explain some large, short-run upswings in the bond rate that have occurred during the subperiod 1979Q1 to 1993Q4. Those upswings in the bond rate are most likely due to short-run swings in its major long-run economic determinant—expected inflation—and hence may be labeled as reflecting inflation scares as in Goodfriend (1993).

The plan of this article is as follows. Section 1 presents the model and the method used in estimating the bond rate equation. Section 2 presents empirical results, and Section 3 contains concluding observations.

1. THE MODEL AND THE METHOD

A Discussion of the Economic Determinants of the Bond Rate: A Variant of the Sargent Model

The economic determinants of the nominal bond rate are identified here by specifying a loanable funds model employed by Sargent (1969), among others. In this model, the nominal interest rate is assumed to be composed of a real component, a component reflecting inflationary expectations, and a component reflecting the influence of monetary policy actions on the real rate. In particular, consider the identity (1) linking real and nominal components:

$$Rn_{(t)} = Re_{(t)} + [Rm_{(t)} - Re_{(t)}] + Rn_{(t)} - Rm_{(t)}, \tag{1}$$

where *Rn* is the nominal interest rate, *Re* is the equilibrium real rate, and *Rm* is the market real rate. The nominal interest rate equation estimated here is based on hypotheses used to explain each of the three terms on the right-hand side of (1).

The first term, *Re*, is the real rate that equates ex ante saving with investment and the government deficit. Assume that savings (S) and investment (I) depend upon economic fundamentals as in (2) and (3):

$$I_{(t)} = g_0 + g_1 \Delta y_{(t)} - g_2 R_{e(t)} \tag{2}$$

$$S_{(t)} = s_0 + s_1 y_{(t)} + s_2 R_{e(t)}, \tag{3}$$

[3] This result is consistent with the Ricardian hypothesis that neither consumption nor interest rates are affected by the stock of government debt or by the deficit. In an extensive survey, Seater (1993) also concludes that the Ricardian hypothesis is approximately consistent with the data.

where y is real income. Equation (2) is an accelerator-investment equation with interest rate effects, while equation (3) is a standard Keynesian savings function. In equilibrium, the government deficit must be covered by an excess of savings over investment. Hence, the equilibrium real rate is the rate that solves equation (4):

$$RDEF_{(t)} = S_{(t)} - I_{(t),} \tag{4}$$

where *RDEF* is the real government deficit. Substituting (2) and (3) into (4) yields the following expression for the equilibrium real rate:

$$R_{e(t)} = \frac{1}{s_2 + g_2}[(g_0 - s_0) + g_1 \Delta y_t - s_1 y_t + RDEF_{(t)}]. \tag{5}$$

The deficit and increases in the rate of growth of real income raise the demand for funds and hence drive up the equilibrium real rate. In contrast, a higher level of output generates a larger volume of savings and hence reduces the equilibrium real rate.

The second term on the right-hand side of (1) is the deviation of the market real rate from the equilibrium real rate. This interest rate gap arises in part from monetary policy actions. The Federal Reserve can affect the real rate by changing the supply of high-powered money. In the loans market, such changes in the supply of money have effects on the demand and supply curves for funds and hence the market real rate as in (6):

$$Rm_{(t)} - Re_{(t)} = -h_i[\Delta rM_{(t)}], \tag{6}$$

where rM is the real supply of money. A rise in real money supply drives the market rate downward with respect to the equilibrium real rate.

The third term on the right-hand side of (1) is the gap between the nominal and real market rates of interest. Such a gap arises as a result of anticipated inflation and is expressed as in (7):

$$Rn_{(t)} - Rm_{(t)} = \beta \dot{p}^e_t, \tag{7}$$

where $\dot{p}^e$ is anticipated inflation. Substituting (5), (6), and (7) into (1) produces (8), which includes the main potential economic determinants of the bond rate suggested in Sargent (1969).

$$Rn_{(t)} = d_0 + d_1 \dot{p}^e_{(t)} + d_2 RDEF_{(t)} - d_3 y_{(t)} - d_4 \Delta rM^s_{(t)} + d_5 \Delta y_{(t)} \tag{8}$$

Equation (8) says that the nominal bond rate depends on anticipated inflation, the deficit, changes in real money supply and income, and the level of income.

An Alternative Econometric Specification

Sargent (1969) estimates equations like (8) for one-year and ten-year bond yields using annual data from 1902 to 1940. The bond rate equations estimated here, however, differ from those reported in Sargent in two major respects. In

Sargent, changes in real money supply capture the impact of monetary policy actions on the equilibrium real rate. As is now widely recognized, financial innovations and the deregulation of interest rates have altered the short-run indicator properties of the empirical measures of money.[4] However, the nominal federal funds rate has been the instrument of monetary policy. Therefore, the impact of monetary policy actions on the real rate is captured by including the real funds rate in the bond rate equation.[5] Secondly, the bond rate equation here is based on cointegration and error-correction methodology, which is better suited to distinguish between the short- and long-run economic determinants of the bond rate than the one used in Sargent and elsewhere.

The nominal bond rate equation estimated here consists of two parts: a long-run part and a short-run part. The long-run part that specifies the potential, long-run determinants of the level of the bond rate is expressed in (9).

$$Rn_{(t)} = a_0 + a_1\dot{p}^e_{(t)} + a_2 RFR_{(t)} + a_3 RDEF_{(t)} - a_4 \ln ry_{(t)} + a_5 \Delta \ln ry_{(t)} + U_{(t)}, \tag{9}$$

where *RFR* is the real federal funds rate, *RDEF* is the real deficit, $\ln ry$ is the logarithm of real income, and U is the disturbance term. Equation (9) describes the long-run responses of the bond rate to anticipated inflation, the real federal funds rate, the real deficit, changes in real income, and the level of real income. The coefficients a_i, $i = 1$ to 5, measure the long-run responses in the sense that they are the sums of coefficients that appear on current and past values of the relevant economic determinants. The term $a_1\dot{p}^e$ in (9) captures the inflation premium in the bond rate, whereas the remaining terms capture the influence of other variables on the equilibrium real component of the bond rate. If the nominal bond rate and anticipated inflation variables are nonstationary but cointegrated as in Engle and Granger (1987), then the other remaining long-run impact coefficients (a_2, a_3, a_4, and a_5 in [9]) may all be zero.

Equation (9) may not do well in explaining short-run movements in the bond rate for a number of reasons. First, it ignores the short-run effects of fundamentals. Some economic factors, including those measuring monetary policy actions, may be important in explaining short-run changes in the bond rate, even though they may have no long-run effects. Second, the long-term bond equation (9) completely ignores short-run dynamics. The presence of expectations and/or adjustment lags in the effects of economic fundamentals on the bond rate may cause the bond rate to differ from the value determined in (9). Hence, in order to explain short-run changes in the bond rate, consider the following error-correction model of the bond rate:

[4] See Hetzel and Mehra (1989) and Feinman and Porter (1992) for evidence on this issue.

[5] Goodfriend (1993) also uses the funds rate to measure the impact of monetary policy actions on the real component of the bond rate.

$$\Delta Rn_{(t)} = c_0 + c_1 \Delta \dot{p}^e_{(t)} + c_2 \Delta RFR_{(t)} + c_3 \Delta RDEF_{(t)} + c_4 \Delta \ln ry_{(t)} + c_5 \Delta^2 \ln ry_{(t)} + \sum_{s=1}^{n} c_{6s} \Delta Rn_{(t-s)} + c_7 U_{(t-1)} + \epsilon_{(t)}, \quad (10)$$

where $U_{(t-1)}$ is the lagged residual from the long-run bond equation (9), Δ^2 is the second-difference operator, and other variables are as defined before. Equation (10) is the short-run bond rate equation, and the coefficients c_i, $i = 1$ to 5, capture the short-run responses of the bond rate to economic determinants suggested here. The coefficients that appear on lagged first differences of the bond rate, c_{6s}, $s = 1$ to n, capture short-run dynamics. The equation is in an error-correction form, indicating that the bond rate will adjust in the short run if the actual bond rate differs from its long-run value determined in (9), i.e., if $U_{(t-1)}$ is different from zero in (10). The coefficient c_7 that appears on the lagged error-correction residual in (10) thus captures the short-run influence of long-run dynamics on the bond rate.

Data and Definition of Variables

The main problem in estimating (9) or (10) is that long-run anticipated inflation is an unobservable variable. The empirical work here first uses actual inflation as a proxy for long-run anticipated inflation. In this case, the coefficient a_1 that appears on actual inflation in the long-run bond equation (9) measures the bond rate's response to anticipated inflation, where the latter is modeled as a distributed lag on current and past inflation rates. Hence, this specification is similar in spirit to the one used in Sargent (1969), who had employed an infinite (geometric) distributed lag as a proxy for inflationary expectations. I, however, also examine results using one-year-ahead inflation rates from the Livingston survey to proxy for long-run anticipated inflation.

The empirical work uses quarterly data from 1955Q1 to 1993Q4.[6] The bond rate is the nominal yield on 30-year U.S. Treasury bonds. Inflation is measured by the behavior of the consumer price index. The real federal funds rate is the nominal federal funds rate minus the actual, annualized quarterly inflation rate. The real deficit variable is included in ratio form as federal government deficits scaled by nominal GDP.[7] Real income is real GDP. Hence, the empirical specifications considered here are given in (11) and (12).

[6] The data on the Livingston survey are provided by the Philadelphia Fed. All other data series are from the Citibank data base.

[7] This specification reflects the assumption that in a growing economy higher deficits result in higher interest rates only if the deficit rises relative to GDP. Hence, the deficit is scaled by GDP. This specification amounts to the restriction that the coefficients a_3 and a_4 in (9) are equal in magnitude but opposite in sign. However, none of the results here qualitatively change if the deficit (*RDEF*) and real GDP ($\ln ry$) enter separately in regressions.

$$R30_{(t)} = a_0 + a_1\dot{p}_{(t)} + a_2 RFR_{(t)} + a_3(DEF/y)_{(t)} + a_4\Delta \ln ry_{(t)} + U_{(t)} \qquad (11)$$

$$\Delta R30_{(t)} = c_0 + c_1\Delta\dot{p}_{(t)} + c_2\Delta RFR_{(t)} + c_3\Delta(DEF/y)_{(t)} + c_4\Delta^2 \ln ry_{(t)} + c_5 U_{(t-1)} + \epsilon_{(t)}, \qquad (12)$$

where $R30$ is the bond rate, $\dot{p}$ is actual inflation, and (DEF/y) is the ratio of deficits to GDP. Equation (11) is the long-run bond rate equation and equation (12) the short-run equation. The alternative specification investigated here replaces $\dot{p}_{(t)}$ in (11) and (12) with $\dot{p}^e_{(t)}$, where $\dot{p}^e$ is the Livingston survey measure of inflationary expectations.

Estimation Issues: The Long-Run Bond Rate Equation

The stationarity properties of the data are important in estimating the long-run bond equation. If empirical measures of economic determinants including the bond rate are all nonstationary variables but cointegrated as in Engle and Granger (1987), then the long-run equation (11) can be estimated by ordinary least squares. Tests of hypotheses on coefficients that appear in (11) can then be carried out by estimating Stock and Watson's (1993) dynamic OLS regressions of the form

$$R30_{(t)} = a_0 + a_1\dot{p}_{(t)} + a_2 RFR_{(t)} + a_3[DEF_{(t)}/y_{(t)}] + a_4\Delta \ln ry_{(t)} + \sum_{s=-k}^{k} a_{4s}\Delta\dot{p}_{(t-s)} + \sum_{s=-k}^{k} a_{5s}\Delta RFR_{(t-s)} + \sum_{s=-k}^{k} a_{6s}\Delta[DEF_{(t-s)}/y_{(t-s)}] + \sum_{s=-k}^{k} a_{7s}\Delta^2 \ln ry_{(t-s)} + \epsilon_{(t)}. \qquad (13)$$

Equation (13) includes, in addition to current levels of economic variables, past, current, and future values of changes in them.

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 tests are performed by estimating the augmented Dickey-Fuller regression of the form

$$X_{(t)} = m_0 + \rho X_{(t-1)} + \sum_{s=1}^{k} m_{1s}\Delta X_{(t-s)} + \epsilon_{(t)}, \qquad (14)$$

where X is the pertinent variable, ϵ is the random disturbance term, and k is the number of lagged first differences of X necessary to make ϵ serially uncorrelated. If $\rho = 1$, X has a unit root. The null hypothesis $\rho = 1$ is tested using

the t-statistic. The lag length (k) used in tests is chosen using the procedure given in Hall (1990), as advocated by Campbell and Perron (1991).[8]

The Dickey-Fuller statistic tests the null hypothesis of unit root against the alternative that X is mean stationary. Recently, some authors including DeJong et al. (1992) have presented evidence that the Dickey-Fuller tests have low power in distinguishing between the null and the alternative. These studies suggest that it would also be useful to perform tests of the null hypothesis of mean stationarity to determine whether the variables are stationary or integrated. Thus, tests of mean stationarity are performed using the procedure advocated by Kwiatkowski, Phillips, Schmidt, and Shin (1992). The test, hereafter denoted as the KPSS test, is implemented by calculating the test statistic

$$\hat{n}_u = \frac{1}{T^2} \sum_{t=1}^{T} S_{(t)}^2 / \hat{\sigma}_k^2,$$

where $S_{(t)} = \sum_{i=1}^{t} e_i, t = 1, 2, \ldots T$, e_t is the residual from the regression of $X_{(t)}$ on an intercept, $\hat{\sigma}_k$ is a consistent estimate of the long-run variance of X, and T is the sample size.[9] The statistic $\hat{n}_u$ has a nonstandard distribution and its critical values have been provided by Kwiatkowski et al. (1992). The null hypothesis of stationarity is rejected if $\hat{n}_u$ is large. Thus, a variable $X_{(t)}$ is considered unit root nonstationary if the null hypothesis that $X_{(t)}$ has a unit root is not rejected by the augmented Dickey-Fuller test and the null hypothesis that it is mean stationary is rejected by the KPSS test.

The test for cointegration used is the one proposed in Johansen and Juselius (1990). The test procedure consists of estimating a VAR model that includes differences as well as levels of nonstationary variables. The matrix of coefficients associated with levels of these variables contains information about the long-run properties of the model. To explain the model, let Z_t be a vector of time series on the bond rate and its economic determinants. Under the hypothesis that the series in Z_t are difference stationary, one can write a VAR model as

$$\Delta Z_t = \Gamma_1 \, \Delta Z_{(t-1)} + \cdots + \Gamma_{(k-1)} \, \Delta Z_{(t-k-1)} + \Pi \, Z_{(t-k)} + \epsilon_{(t)}, \qquad (15)$$

[8] The procedure is to start with some upper bound on k, say k max, chosen a priori (eight quarters here). Estimate the regression (14) with k set at k max. If the last included lag is significant, select $k = k$ max. If not, reduce the order of the estimated autoregression by one until the coefficient on the last included lag (on ΔX in [14]) is significant. If none is significant, select $k = 0$.

[9] The residual e_t is from the regression $X_t = a + e_t$. The variance of X_t is the variance of the residuals from this regression and is estimated, using the Newey and West (1987) method, as

$$\hat{\sigma}_k = \frac{1}{T} \sum_{t=1}^{T} e_t^2 + \frac{2}{T} \sum_{s=1}^{T} b(s,k) \sum_{t=s+1}^{T} e_t e_{t-s},$$

where T is the sample size, the weighing function $b(s,k) = 1 + \frac{s}{1+k}$, and k is the lag truncation parameter. The lag parameter was set at $k = 8$.

where Γ_i, $i = 1, 2, \ldots k-1$, and Π are matrices of coefficients that appear on first differences and levels of the time series in Z_t.

The component ΠZ_{t-k} in (15) gives different linear combinations of levels of the time series in Z_t. Thus, the matrix Π contains information about the long-run properties of the model. When the matrix's rank is zero, equation (15) reduces to a VAR in first differences. In that case, no series in Z_t can be expressed as a linear combination of other remaining series. This result indicates that there does not exist any long-run relationship between the series in the VAR. On the other hand, if the rank of Π is one, then there exists only one linear combination of series in Z_t. That result indicates that there is a unique, long-run relationship between the series.

Two test statistics can be used to evaluate the number of the cointegrating relationships. The trace test examines the rank of Π matrix and the hypothesis that rank $(\Pi) \leq r$, where r represents the number of cointegrating vectors. The maximum eigenvalue statistic tests the null that the number of cointegrating vectors is r, given the alternative of $r+1$ vectors. The critical values of these test statistics have been reported in Johansen and Juselius (1990).

OLS estimates are inconsistent if any right-hand explanatory variable in the long-run bond equation (11) is stationary. In that case, the long-run bond equation (11) can be estimated jointly with the short-run bond equation (12). To do so, solve (11) for $U_{(t-1)}$ and then substitute for $U_{(t-1)}$ into (12) to yield (16).

$$\begin{aligned}\Delta R30_{(t)} = {} & (c_0 - c_5 a_0) + c_1 \Delta \dot{p}_{(t)} + c_2 \Delta RFR_{(t)} + c_3 \Delta (DEF_{(t)}/y_{(t)}) \\ & + c_4 \Delta^2 \ln ry(t) - c_5 R30_{(t-1)} - c_5 a_1 \dot{p}_{(t-1)} - c_5 a_2 RFR_{(t-1)} \\ & - c_5 a_3 DEF_{(t-1)}/y_{(t-1)} - c_5 a_4 \Delta \ln ry_{(t-1)} + \epsilon_t \qquad (16)\end{aligned}$$

Equation (16) is the short-run bond rate equation that includes levels as well as differences of the relevant economic determinants. The long-run coefficients a_i, $i = 1, 2, 3$, can be recovered from the reduced-form estimates of equation (16).[10] The equation can be estimated by ordinary least squares,[11] or by instrumental variables if contemporaneous right-hand variables are correlated with the disturbance term.

[10] The long-run coefficient on inflation (a_1) is the coefficient on $\dot{p}(t-1)$ divided by the coefficient on $R30_{(t-1)}$; the long-run coefficient on deficit (a_3) is the coefficient on $DEF_{(t-1)}/y_{(t-1)}$ divided by the coefficient on $R30_{(t-1)}$; and the long-run coefficient on the real funds rate is the coefficient on $RFR_{(t-1)}$ divided by the coefficient on $R30_{(t-1)}$. The intercept a_0, however, cannot be recovered from these reduced-form estimates.

[11] Since lagged levels of economic determinants appear in (16), ordinary least squares estimates are consistent if some variables on the right-hand side of (16) are in fact stationary.

2. ESTIMATION RESULTS

Unit Root Test Results

Table 1 presents test results for determining whether the variables $R30$, $\dot{p}$, $\dot{p}^e$, and (DEF/y) 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 all these variables. 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 $R30$, $\dot{p}$, $\dot{p}^e$ and (DEF/y), but small for *RFR*. These results thus indicate that $R30$, $\dot{p}$, $\dot{p}^e$, and (DEF/y) have a unit root and are thus nonstationary in levels. The results are inconclusive for the *RFR* variable.

As indicated before, a variable has a unit root if $\rho = 1$ in (14). In order to indicate the extent of uncertainty about the point-estimate of ρ, Table 1 also contains estimates of ρ and their 95 percent confidence intervals. As can be seen, the estimated intervals contain the value $\rho = 1$ for levels of these variables. However, these intervals appear to be quite wide: their lower limits are close to .90 for the series shown. These results indicate that the variables may well be mean stationary. Hence, I also derive results treating all variables as stationary.

Table 1 also presents unit root tests using first differences of $R30$, $\dot{p}$, $\dot{p}^e$, *RFR*, $\ln ry$ and (DEF/y). As can be seen, the t-statistic for the hypothesis $\rho = 1$ is fairly large for all these variables. The point-estimates of ρ also diverge away from unity. These results indicate that first differences of these variables are stationary.

Cointegration Test Results

Treating the bond rate, inflation, the real funds rate, and government deficits as nonstationary variables, Table 2 presents test statistics for determining whether the bond rate is cointegrated with any of these variables.[12] Change in real income ($\Delta \ln ry$) is not considered because it is a stationary variable. Trace and maximum eigenvalue statistics, which test the null hypothesis that there is no cointegrating vector, are large for systems $(R30, \dot{p})$, $(R30, \dot{p}^e)$, $(R30, DEF/y)$, $(R30, \dot{p}, DEF/y)$ and $(R30, \dot{p}^e, DEF/y)$, but are very small for the system $(R30, RFR)$. These results indicate that the bond rate is cointegrated with inflation (actual or expected) and deficits, but not with the real funds rate. That is, the bond rate stochastically co-moves with inflation and the deficit variable, but not with the real funds rate.

[12] 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 JJ procedure is the one that results in the rejection of the restricted model.

Table 1 Tests for Unit Roots and Mean Stationarity

	Panel A Tests for Unit Roots				Panel B Tests for Mean Stationarity
Series X	$\hat{\rho}$	$t_{\hat{\rho}}$	k	**Confidence Interval for ρ**	$\hat{n}_u$
$R30$	.97	−1.65	5	(.93, 1.03)	1.31*
$\dot{p}$	.87	−2.74	7	(.84, 1.01)	.53*
$\dot{p}^e$	.97	−1.78	2	(.92, 1.02)	1.02*
RFR	.85	−2.38	2	(.87, 1.02)	.39
DEF/y	.93	−2.46	1	(.87, 1.02)	1.42*
$\Delta R30$	−.02	−5.47*	8		
$\Delta\dot{p}$	−.70	−5.09*	8		
$\Delta\dot{p}^e$	.38	−6.33*	1		
ΔRFR	−1.50	−5.52*	7		
$\Delta DEF/y$	−.80	−6.18*	8		
$\Delta \ln ry$	.20	−4.83*	7		

* Significant at the 5 percent level.

Notes: $R30$ is the 30-year bond rate; $\dot{p}$ is the annualized quarterly inflation rate measured by the behavior of consumer prices; $\dot{p}^e$ is the Livingston survey measure of one-year-ahead expected inflation; RFR is the real federal funds rate; and DEF/y is the ratio of federal government deficits to nominal GDP. Δ is the first-difference operator. The sample period studied is 1955Q1 to 1993Q4. 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 number of lagged first differences (k) included in these regressions are chosen using the procedure given in Hall (1990), with maximum lag set at eight quarters. The confidence interval for ρ is constructed using the procedure given in Stock (1991).

The test statistics $\hat{n}_u$ in Panel B above 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 .463.

Table 3 presents the dynamic OLS estimates of the cointegrating vector between the bond rate and its long-run determinants, inflation and the deficit. Panel A presents estimates with actual inflation ($\dot{p}$) and Panel B with expected inflation ($\dot{p}^e$). In addition, the cointegrating vector is estimated under the restriction that the bond rate adjusts one for one with inflation in the long run. In regressions estimated without the above-noted full Fisher-effect restriction, the right-hand explanatory variables have their theoretically predicted signs and are statistically significant. Thus, the bond rate is positively correlated with inflation and deficits in the long run. The coefficient that appears on the

Table 2 Cointegration Test Results

System	k[a]	Trace Test	Maximum Eigenvalue Test
$(R30, \dot{p})$	8	23.7*	21.2*
$(R30, \dot{p}^e)$	8	20.6*	17.6*
$(R30, RFR)$	5	12.1	8.9
$(R30, DEF/y)$	5	30.4*	27.5*
$(R30, \dot{p}, DEF/y)$	8	46.8*	35.6*
$(R30, \dot{p}^e, DEF/y)$	8	48.3*	31.9*

[a] The lag length k was selected using the likelihood ratio test procedure described in footnote 12 of the text.

* Significant at the 5 percent level.

Notes: Trace and maximum eigenvalue tests are tests of the null hypothesis that there is no cointegrating vector in the system. For the two-variable system, the 5 percent critical value is 17.8 for the trace statistic and 14.5 for the maximum eigenvalue statistic. Critical values are from Johansen and Juselius (1990). (For the three-variable system, the corresponding 5 percent critical values are 31.2 and 21.3.)

Table 3 Cointegrating Regressions; Dynamic OLS

(Leads, Lags)	Without the Full Fisher-Effect Restriction	With the Full Fisher-Effect Restriction
	Panel A: $(R30, \dot{p}, DEF/y)$	
$(-4, 4)$	$R30_t = 2.0 + .61\dot{p}_t + .73(DEF/y)_t$ (.03) (.03) (.04)	$R30_t = 1.3 + 1.0\dot{p}_t + .30(DEF/y)_t$ (.10) (.03) (.05)
$(-8, 8)$	$R30_t = 2.0 + .60\dot{p}_t + .78(DEF/y)_t$ (.12) (.03) (.08)	$R30_t = 1.3 + 1.0\dot{p}_t - .13(DEF/y)_t$ (.10) (.03) (.03)
	Panel B: $(R30, \dot{p}^e, DEF/y)$	
$(-4, 4)$	$R30_t = 2.6 + .77\dot{p}^e_t + .46(DEF/y)_t$ (.13) (.03) (.03)	$R30_t = 2.1 + 1.0\dot{p}^e_t + .13(DEF/y)_t$ (.11) (.03) (.03)
$(-8, 8)$	$R30_t = 2.6 + .72\dot{p}^e_t + .57(DEF/y)_t$ (.15) (.06) (.13)	$R30_t = 2.8 + 1.0\dot{p}^e_t + .00(DEF/y)_t$ (.14) (.00) (.03)

Notes: All regressions are estimated by the dynamic OLS procedure given in Stock and Watson (1993), using leads and lags of first differences of the relevant right-hand side explanatory variables. Parentheses contain standard errors corrected for the presence of moving average serial correlation. The dynamic OLS regressions also include leads and lags of the real federal funds rate.

inflation variable ranges between .6 and .8 and is less than unity, indicating that the bond rate does not adjust one for one with inflation in the long run. The coefficient that appears on the deficit variable ranges between .4 and .8, indicating that a one percentage point increase in the ratio of deficits to GDP raises the bond rate by 40 to 80 basis points.[13] However, the coefficient that appears on the deficit variable is sensitive to the restriction that there is a full Fisher effect. If the cointegrating regression is reestimated with this restriction, then the deficit variable coefficient becomes small and even turns negative in some cases (see Table 3).

The full Fisher-effect restriction is in fact rejected by the data, indicating that it should not be imposed routinely on the bond regression. Nevertheless, it is a reasonable restriction to consider if one wants to carry out the sensitivity analysis. The finding that the long-run deficit-interest rate link weakens when the restriction is imposed indicates that the deficit may be proxying the information that is already in inflation. The deficit appears to raise the long rate because of its positive effect on anticipated inflation rather than on the real component of the bond rate. Hence, these results imply that inflation is the main, long-run economic determinant of the bond rate.

The Short-Run Bond Rate Equation

Since unit root test results are inconclusive for some series, the short-run bond equation is estimated jointly with the long-run part as in (16), which includes lagged levels of the series. If the variables are stationary in levels, OLS estimates will still be consistent.

Table 4 presents instrumental variable estimates of the bond equation (16).[14] Panel A there reports regressions with actual inflation ($\dot{p}$), and Panel B regressions with expected inflation ($\dot{p}^e$). In addition, I estimate the equation with and without the constraint that the bond rate adjusts one for one with inflation in the long run (compare equations in columns A.1 and B.1 versus A.2 and B.2, Table 4). As can be seen, the coefficients that appear on various economic variables have their theoretically predicted signs and in general are statistically significant. The results there indicate that in the short run the bond rate rises if inflation increases, or if the real federal funds rate rises. Changes

[13] These estimates are close to those reported in Hoelscher (1986). Hoelscher uses the ten-year bond rate and the Livingston survey measure as proxies for long-term expected inflation. He estimates the bond regression from 1953 to 1984, using annual data. The coefficients that appear on his inflation and deficit variables are .84 and .42, respectively. Hoelscher does not, however, examine the sensitivity of results to the restriction that the bond rate adjusts one for one with inflation in the long run.

[14] I use instrumental variable estimates because contemporary values of changes in the funds rate, inflation, and real income variables may be correlated with the disturbance term. For example, the evidence in Mehra (1994) indicates that the Fed has responded to the information in the bond rate about long-run expected inflation. Hence, the change in the funds rate may be contemporaneously correlated with the disturbance term.

Table 4 Error-Correction Bond Rate Regressions

Explanatory Variables	Panel A Regressions Using Actual Inflation Data		Panel B Regressions Using Livingston Survey Inflation Data	
	A.1	A.2	B.1	B.2
constant	.55 (2.1)	−.01 (0.1)	1.80 (2.4)	.59 (4.5)
$R30_{t-1}$	−.29 (4.7)	−.18 (4.6)	−.59 (3.2)	−.30 (6.2)
$\dot{p}_{t-1}$	.20 (4.9)	.18 (4.6)		
$\dot{p}^e_{t-1}$			.43 (4.2)	.30 (6.2)
$(DEF/y)_{t-1}$	.18 (3.2)	.06 (2.2)	.24 (1.8)	.02 (0.8)
RFR_{t-1}	.19 (3.6)	.13 (2.9)	.31 (2.9)	.15 (4.1)
$\Delta\dot{p}_t$	.40 (3.7)	.32 (3.3)		
$\Delta\dot{p}^e_t$			.61 (2.1)	.26 (1.8)
ΔRFR_t	.35 (4.6)	.24 (4.0)	.31 (2.5)	.14 (2.5)
$\Delta \ln rY_t$	−.01 (0.2)	.05 (1.4)	−.10 (1.2)	.03 (1.1)
$\Delta R30_{t-1}$	−.10 (1.1)	−.24 (3.2)	.31 (1.4)	−.01 (0.1)
$\Delta R30_{t-2}$	.16 (1.7)	.09 (1.0)	.10 (0.8)	.03 (0.4)
SER	.451	.434	.709	.504
DW	2.0	1.84	2.0	1.95
Q(36)	35.3	37.6	54.9	46.3
$n(\dot{p}, RFR, DEF/y)$	(.7, .7, .6)	(1.0, .7, .3)		
$n(\dot{p}^e, RFR, DEF/y)$			(.7, .5, .4)	(1.0, .5, .1)

Notes: All regressions are estimated by instrumental variables. The instruments used are a constant, one lagged value of the bond rate, inflation, the real federal funds rate, and the ratio of deficits to GDP and two lagged values of changes in inflation, the real funds rate, real GDP, and the bond rate. Regressions in columns A.2 and B.2 above are estimated under the restriction that coefficients on $R30_{t-1}$ and $\dot{p}_{t-1}$ ($\dot{p}^e_{t-1}$) sum to zero (there is complete Fisher-effect), while those in columns A.1 and B.1 are without this restriction. SER is the standard error of regression, DW is the Durbin-Watson statistic, and Q(36) is the Lung-Box Q-statistic based on 36 autocorrelations of the residuals. $n(x_1, x_2, x_3)$ indicates the long-run (distributed) responses of the 30-year bond rate to x_1, x_2, and x_3, respectively.

in real GDP do not have much of an impact on the bond rate.[15] The coefficients that appear on contemporaneous values of these variables range from .3 to .6 for inflation and from .2 to .4 for the real funds rate. Thus, a one percentage point increase in the rate of inflation raises the bond rate between 30 to 60 basis points, while a similar increase in the real funds rate raises it by 14 to 35 basis points in the short run.[16]

[15] First differences of the deficit variable and second differences of real GDP when included in regressions were generally not significant.

[16] The point-estimates of the short-run, monetary policy impact coefficient found here are close to those found or assumed in some other studies. For example, the empirical work presented in Cook and Hahn (1989) indicates that a one percentage point rise in the funds rate target raises the long rate by 10 to 20 basis points, whereas in Goodfriend (1993) such an increase in the funds rate is assumed to raise the long rate by 25 basis points.

As indicated before, the bond rate's long-run distributed-lag responses to economic determinants here can be recovered from the reduced-form estimates of the short-run bond equation presented in Table 4. As can be seen, the long-run coefficients that appear on these variables range from .7 to 1.0 for inflation, .5 to .7 for the real funds rate, and .1 to .6 for the deficit variable. Moreover, as before, the long-run coefficient on the deficit variable becomes small and is statistically insignificant if the full Fisher-effect restriction is imposed on the data (see equations A.2 and B.2 in Table 4). The long-run coefficient that appears on the real funds rate, however, remains quite large and is statistically significant. This result indicates that (stationary) movements in the real funds rate can have substantial effects on the bond rate in the short run.[17,18]

Predictive Ability of the Bond Rate Equation

I now examine whether bond rate regressions presented in Table 4 can explain the actual behavior of the bond rate. In particular, I examine one-year-ahead dynamic forecasts of the bond rate from 1971Q1 to 1993Q4, using regressions A.2 and B.2 of Table 4. Recall that regression A.2 uses actual inflation as a proxy for long-run inflationary expectations and regression B.2 uses one-year-ahead expected inflation as a proxy. Since the forecast performance of these two regressions is similar, I discuss results only for the former.

Figure 1 charts the quarterly values of the bond rate, actual and predicted. As can be seen, the regression captures fairly well broad movements in the bond rate from 1971Q1 to 1993Q4. The mean prediction error is small, only 6 basis points, and the root mean squared error is .74 percentage points. This regression outperforms a purely eight-order autoregressive model of the bond rate. For the time series model, the mean prediction error is 13 basis points and the root mean squared error is 1.2 percentage points.

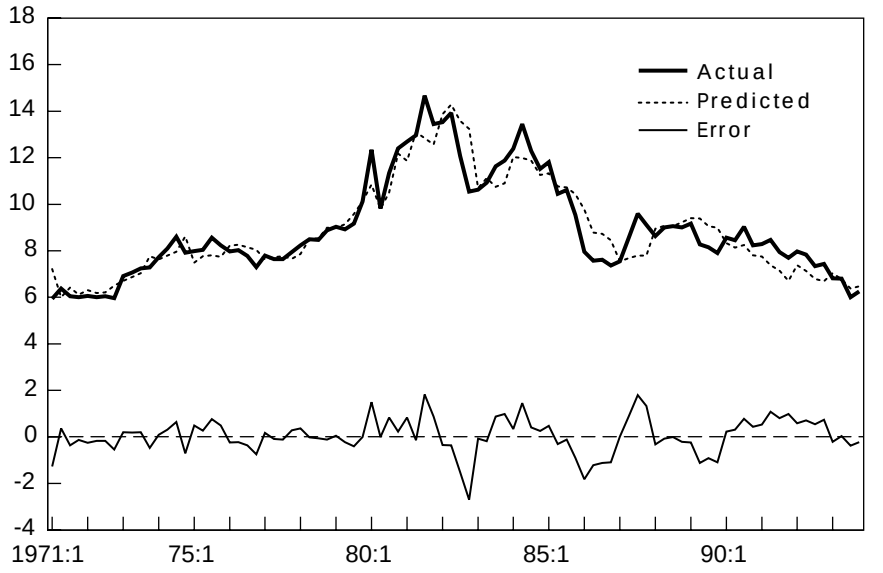
I evaluate further the predictive performance from 1971Q1 to 1993Q4 by estimating regressions of the form (17).

---

[17] If all variables are stationary, then the long-run coefficient that appears on the funds rate in the short-run bond equation measures the sum of coefficients associated with current and past values of changes in the funds rate. Since permanent movements in the funds rate have no permanent effects on the bond rate, this long-run coefficient in fact measures the short-run response of the bond rate to changes in the funds rate.

[18] The long-run coefficient that appears on the funds rate in the bond rate regression may be an upwardly biased estimate of the impact of monetary policy actions on the real component of the bond rate. The main source of this potential bias is the absence of the relevant long-run expected inflation variable in these regressions. If the Fed responds to variables that have information about long-run expected inflation, then the funds rate may be picking up the influence of expected inflation on the bond rate rather than the influence of monetary policy actions on the real component of the bond rate. Some evidence that favors this view emerges in Table 4. As can be seen, the magnitude of the long-run coefficient that appears on the funds rate declines from .7 to .5 if one-year-ahead expected inflation (Livingston) data are substituted for actual inflation in the regression.

Figure 1 Actual and Predicted 30-Year Bond Rate



Note: Predicted values are generated using the regression with actual inflation (regression A.2 in Table 4).

$$A_{(t)} = d_0 + d_1 P_{(t)}, \tag{17}$$

where A is the actual quarterly value of the bond rate and P is the value predicted by the bond rate regression. If $d_0 = 0$ and $d_1 = 1$, then regression forecasts are unbiased. The coefficients d_0 and d_1 take values .3 and .97, respectively, for regression A.2[19] and 1.7 and .8, respectively, for the time series model. The hypothesis $d_0 = 0$, or $d_1 = 1$, is rejected for the time series model, but not for the economic models.[20]

Unpredictable, Short-Run Upward Swings in the Bond Rate: Inflation

A look at Figure 1 indicates that the bond rate regression estimated here fails to predict some large, short-run movements in the bond rate that have occurred during the post-1979 period.[21] Table 5 presents quarterly changes in the bond rate from 1979Q1 to 1994Q2. It also presents changes predicted by the bond

[19] For regression B.2 of Table 4, $d_0 = .13$ and $d_1 = 1.0$.

[20] For regression A.2, the relevant Chi-squared statistics that test $d_0 = 0$ and $d_1 = 1$ are .3 and .2, respectively. The relevant statistics are .03 and 0.0 for regression B.2 of Table 4. For the time series model, the relevant Chi-squared statistics take values 3.9 and 4.1. Each Chi-squared statistic is distributed with one degree of freedom. The 5 percent critical value is 3.84.

[21] Such large, short-term increases in the bond rate did not occur during the pre-1979 period.

Table 5 Actual and Predicted Quarterly Changes in the Bond Rate 1979Q1 to 1994Q2

Year/Qtr.	Actual	Predicted	Error	Year/Qtr.	Actual	Predicted	Error
1979Q1	.15	.07	.07	1987Q1	.18	.20	−.02
1979Q2	−.11	.06	−.17	1987Q2	1.02[a]	.16	.85
1979Q3	.25	.37	−.12	1987Q3	1.02[a]	−.07	1.09
1979Q4	.95	.49	.46	1987Q4	−.47	−.22	−.24
1980Q1	2.22[a]	.55	1.66	1988Q1	−.49	−.25	−.24
1980Q2	−2.53	−.69	−1.84	1988Q2	.37	.25	.12
1980Q3	1.53[a]	.74	.79	1988Q3	.06	−.07	.13
1980Q4	1.06	1.09	−.03	1988Q4	−.05	.22	−.27
1981Q1	.29	−.63	.92	1989Q1	.16	.44	−.27
1981Q2	.27	1.32	−1.06	1989Q2	−.90	.03	−.93
1981Q3	1.71[a]	−.32	2.03	1989Q3	−.12	−.10	−.01
1981Q4	−1.22	−.14	−1.07	1989Q4	−.25	.12	−.37
1982Q1	.08	.44	−.36	1990Q1	.66	.55	.11
1982Q2	.39	.53	−.14	1990Q2	−.10	−.31	.21
1982Q3	−1.85	−.69	−1.15	1990Q3	.57	−.01	.58
1982Q4	−1.53	.01	−1.54	1990Q4	−.79	−.80	.01
1983Q1	.09	.42	−.33	1991Q1	.05	−.67	.72
1983Q2	.30	.48	−.18	1991Q2	.18	−.47	.65
1983Q3	.70[a]	−.30	1.00	1991Q3	−.52	−.50	−.02
1983Q4	.25	.06	.18	1991Q4	−.25	−.64	.39
1984Q1	.50	.13	.36	1992Q1	.27	−.36	.63
1984Q2	1.06[a]	.01	1.05	1992Q2	−.13	−.45	.32
1984Q3	−1.15	−.35	−.79	1992Q3	−.50	−.50	.00
1984Q4	−.77	−.37	−.40	1992Q4	.10	−.17	.27
1985Q1	.29	−.11	.40	1993Q1	−.62	−.60	−.02
1985Q2	−1.36	−.59	−.77	1993Q2	−.01	−.51	.29
1985Q3	.16	.17	−.01	1993Q3	−.81	−.51	−.29
1985Q4	−1.07	−.36	−.70	1993Q4	.25	.32	−.07
1986Q1	−1.58	.50	−2.10	1994Q1	.35	−.43[b]	.78
1986Q2	−.39	.41	.03	1994Q2	.80[a]	.01[b]	.78
1986Q3	.05	−.13	.18				
1986Q4	−.25	−.03	−.21				

Mean Error	−.004
RmSE	.74

[a] This significant increase in the bond rate is not predicted by the bond rate regression A.2 of Table 4 (the prediction error is at least as large as the root mean squared error).

[b] This forecast assumes that during the first and second quarters the ratio of deficits to GDP equals the value observed in 1993Q4.

Notes: The predicted values are generated using the bond rate regression A.2 of Table 4.

regression. If we focus on quarterly increases in the bond rate that are significantly underpredicted by the regression (that is, magnitudes of prediction errors either equal or exceed the root mean squared error), the results then indicate that the bond rate rose 2.2 percentage points in 1980Q1, 1.53 in 1980Q3, 1.71 in 1981Q3, .7 in 1983Q4, 1.1 in 1984Q2, 2.1 in 1987Q2 to 1987Q3, and .8 in 1994Q2 (see Table 5). Except for the latest episode, most of these short-run upswings in the bond rate have been subsequently reversed, so that for the period as a whole the bond rate is well predicted by the regression.

The bond rate equation here attempts to explain changes in the bond rate using actual, not long-run anticipated, values of economic fundamentals. In the long run, actual values of fundamentals may move with anticipated values, but that may not be so in the short run. Hence, if the bond rate in fact responds to anticipated fundamentals, then the bond rate regressions estimated here may not explain very well short-run movements in the bond rate. These considerations suggest one possible explanation of some unpredictable short-run upswings in the bond rate that have occurred since 1979: namely, short-term movements in its anticipated fundamentals. Since, as indicated by cointegration test results, inflation, rather than the deficit or the real funds rate, is the main long-run economic determinant of the bond rate, the short-run increases in the bond rate may then be due to short-run movements in its long-run determinant—anticipated inflation.[22] Thus, the bond rate may rise with anticipated inflation in the short run even as actual inflation remains steady. Such upswings, however, are likely to be reversed if they are not substantiated by the behavior of actual inflation. As can be seen in Table 5, that in fact has been the case.

Following Goodfriend (1993), the periods during which large, unpredictable increases in the bond rate have occurred can be labeled as inflation scares. Goodfriend uses a narrative approach to discuss the interactions among the bond rate, the federal funds rate, and economic determinants such as inflation and real growth. He assumes that inflation is the bond rate's main long-run determinant and that changes in the funds rate have minor short-run effects on the bond rate. Hence, he calls a significant bond rate rise in the absence of an aggressive funds rate tightening an inflation scare. The results from a more formal bond rate equation here are in line with those in Goodfriend (1993).

[22] Short-run changes in anticipated monetary policy actions and deficits cannot explain the big increases in the bond rate either. As noted before, the bond rate is unrelated to short-term changes in the deficit. Furthermore, the magnitudes of future funds rate increases needed to explain the current increases in the bond rate are too big to be consistent with past Fed behavior. In the past, the Fed has moved the funds rate in small increments most of the time.

3. CONCLUDING OBSERVATIONS

Using cointegration and error-correction methodology and building on the loanable funds model of interest rate determination given in Sargent (1969), this article identifies the main long- and short-run economic determinants of the bond rate. In the cointegrating regression, inflation and fiscal deficits appear as two potential long-run economic determinants of the bond rate. That regression indicates that the bond rate is positively correlated with inflation and the deficit and that the bond rate does not adjust one for one with inflation in the long run. However, if that regression is reestimated under the restriction that the bond rate does in fact adjust one for one with inflation, then the long-run deficit-interest rate link found here weakens. Those results imply that the positive effect of the deficit on the real component of the bond rate found here is suspect. Hence, inflation emerges as the main economic determinant of the long rate.

The results here also indicate that changes in the real federal funds rate have substantial short-run effects on the bond rate, even though long-run stochastic movements in the bond rate are unrelated to the real funds rate. In addition, the bond rate rises if inflation accelerates. Surprisingly, current changes in real GDP do not have much of an effect on the bond rate.

The bond rate regressions estimated here are broadly consistent with the actual behavior of the bond rate from 1971 to 1993. However, these regressions fail to predict some large, short-run upswings in the bond rate that have occurred during the subperiod 1979Q1 to 1994Q2. One possible explanation of these results is that actual inflation may be a poor proxy for the long-run expected rate of inflation, the main long-run economic determinant of the bond rate. Hence, the bond rate may rise significantly in the short run if long-run anticipated inflation increases, even though actual inflation may have been steady.

REFERENCES

Campbell, J. Y., and P. Perron. "Pitfalls and Opportunities: What Macroeconomists Should Know About Unit Roots," in O. J. Blanchard and S. Fischer, eds., *NBER Macroeconomics Annual, 1991*. Cambridge, Mass.: MIT Press, 1991, pp. 141–200.

Cook, Timothy, and Thomas Hahn. "The Effect of Changes in the Federal Funds Rate Target on Market Interest Rates in the 1970s," *Journal of Monetary Economics,* vol. 24 (November 1989), pp. 331–51.

DeJong, David N., and John C. Nankervis, N. E. Savin, and Charles H. Whiteman. "Integration Versus Trend Stationarity in Time Series," *Econometrica,* vol. 60 (March 1992), pp. 423–33.

Dickey, D. A., and W. A. Fuller. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association,* vol. 74 (June 1979), pp. 427–31.

Echols, Michael E., and Jan Walter Elliot. "Rational Expectations in a Disequilibrium Model of the Term Structure," *American Economic Review,* vol. 66 (March 1976), pp. 28–44.

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

Feinman, Joshua, and Richard D. Porter. "The Continuing Weakness in M2," Finance and Economic Discussion Paper #209. Washington: Board of Governors of the Federal Reserve System, September 1992.

Fuller, W. A. *Introduction to Statistical Time Series*. New York: Wiley, 1976.

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, A. "Testing for a Unit Root in Time Series with Pretest Data Based Model Selection." Manuscript. North Carolina State University, 1990.

Hetzel, Robert L., and Yash P. Mehra. "The Behavior of Money Demand in the 1980s," *Journal of Money, Credit, and Banking,* vol. 21 (November 1989), pp. 455–63.

Hoelscher, Gregory. "New Evidence on Deficits and Interest Rates," *Journal of Money, Credit, and Banking,* vol. XVII (February 1986), pp. 1–17.

Johansen, Soren, and Katarina Juselius. "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.

Kwiatkowski, Denis, Peter C. B. Phillips, Peter Schmidt, and Yoncheol 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.

Mehra, Yash P. "A Federal Funds Rate Equation," Mimeo, Federal Reserve Bank of Richmond, March 1994.

Newey, Whitney K., and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica,* vol. 55 (May 1987), pp. 703–8.

Sargent, Thomas J. "Commodity Price Expectations and the Interest Rate," *Quarterly Journal of Economics,* vol. 83 (February 1969), pp. 127–40.

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

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

Stock, James H. "Confidence Intervals for the Largest Autoregressive Root in U.S. Macroeconomic Time Series," *Journal of Monetary Economics*, vol. 28 (December 1991), pp. 435–59.

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

Medical Care Price Indexes

Robert F. Graboyes

Health care expenditures have grown from 4.4 percent of the U.S. economy in 1950 to over 13 percent in 1994. At the same time, medical care prices have risen twice as fast as other prices, according to the Consumer Price Index (CPI). That apparent increase in the price of medical care (relative to other goods and services) would explain by itself the additional spending for health care, though some research suggests that the numbers not be taken at face value. The purpose of this article is to give an understanding of how medical care price indexes are created and why some researchers have expressed concerns about how these indexes are interpreted.

The article is organized as follows: Section 1 introduces the notion and purpose of a price index. Section 2 explains what is meant by quality change and focuses on areas such as the changing efficacy of a medical intervention, the introduction of new goods, and the use of generic drugs. An additional subsection outlines several proposals for the difficult task of constructing a valid price index when quality changes. Section 3 explains some index problems not associated with quality change. Section 4 summarizes the concern that today's indexes may overstate medical inflation. Finally, the appendix gives details on some currently published indexes.

1. LOGIC AND CONSTRUCTION OF PRICE INDEXES[1]

A price index measures the average price of a set of goods and services in one period against the average price of the same goods in another period. The central logic is that this basket of goods and services provides an adequate measure of some average purchaser's standard of living or level of satisfaction.

This article also appears in the third edition of *Macroeconomic Data: A User's Guide*, Roy Webb, ed. (Richmond: Federal Reserve Bank of Richmond, 1994). The views expressed are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Wallace and Cullison (1981) and Webb and Willemse (1994) describe in detail procedures used and problems encountered in constructing any macroeconomic price index.

As the price of the basket changes, the index changes proportionally. A 10 percent rise in a medical care price index thus implies a 10 percent increase in the cost of a fixed quantity of medical care for some average purchaser, even though some individual prices will have risen and others will have fallen.

The first task in creating these indexes is to define the limits of medical care: Do we treat cough drops as medicine and include them, or do we call them candy and exclude them? Do we include gymnasium membership dues, since exercise helps prevent illness, or do we count the dues as recreational expenses and exclude them? Once the medical sector or subsector is defined, individual medical price data can be collected. Then these data must be averaged into an index by using some set of arithmetic weights. These weights generally reflect the relative amount spent on each product in some base period. In the CPI, hospital services receive larger weights than aspirin because consumers spend more on hospital services than on aspirin.

2. MEDICAL CARE PRICE INDEXES AND QUALITY CHANGE

Technological progress has changed significantly the quality of medical care in this century, and this is the fundamental complication in producing medical care price indexes. Implicitly, a price index assumes that one's consumption basket does not change over time and a given basket provides a constant level of satisfaction. While these assumptions are never strictly true for any set of commodities, they are especially problematic in medicine. The treatments given in 1944 barely resemble those given in 1994. And the health benefits of a given treatment can change through the years as well.[2]

The productivity of medical care has advanced greatly over this century. Some of the types of technological progress include the following: [1] **Previously untreatable disease becomes treatable:** In recent decades, heart transplants and coronary bypass operations have given years of life, whereas earlier patients would have died. Therapies such as antibiotics, beta blockers, insulin therapy, and kidney dialysis have effected similar improvements. [2] **Previous treatment replaced by new treatment:** Laparoscopic techniques, using fiber optics and tiny incisions, have largely replaced traditional open surgery in many areas. For example, newer techniques for gallbladder removal require only one to two days in the hospital, compared with three to seven days for traditional surgery. The laparoscopic procedure also results in fewer postoperative complications, less pain, and a shorter convalescence. In addition, some patients for whom traditional surgery is too risky can safely undergo the newer technique.[3]

[2] For example, the expected benefit of a heart transplant is much higher in 1994 than it was in 1970, when the operation was still experimental.

[3] Legorretta et al. (1993).

[3] **Cheap prevention of costly diseases:** Vaccines against polio, smallpox, and other diseases have provided relatively inexpensive means to eradicate diseases that, if contracted, would impose tremendous costs. [4] **Decreased resource requirements for an existing treatment:** Electronic monitors allow some conditions to be tracked at home rather than in a hospital bed, thus reducing the need for hospital resources. Some cost reductions have resulted more from a change in medical opinion than in any change in technology; for instance, doctors now recommend shorter hospital stays following childbirth. [5] **Movement up the learning curve:** Since the first coronary bypasses were performed, practice and observation have made surgeons more adept at the procedure, resulting in higher success rates.

In some areas of medicine, however, a given level of medical spending may now provide fewer health benefits than in the past. Defensive medicine—care that does not benefit the patient and whose purpose is to avoid malpractice claims—has become a fixture of American medicine.[4] The health benefits of other procedures are hotly debated—prostate and breast cancer screenings are examples. Heroic end-stage care for the terminally ill is another. A final complication in measuring the quality of medical care is that the population being treated and the illnesses people suffer change over time. It is impossible to neatly compare the productivity of a medical system pre- and post-AIDS, for example.

Quality changes such as these complicate the construction and interpretation of medical care price indexes. Some examples discussed below illustrate difficulties encountered when [1] the efficacy of a good or service changes, [2] new goods are introduced, and [3] old goods are reintroduced under new labels.

Change in Efficacy

Over time, the improved health from using a specific medical commodity often increases (or decreases). This section uses a hypothetical example to demonstrate the analytical difficulties posed by changes in the quality of medical care. Table 1a shows data on a hypothetical economy in which gross domestic product (GDP) consists of two goods: medical procedures (say, an operation) and food. From year 0 to year *t*, nominal GDP (the sum of spending on food and medicine) grows from $9.5 million to $12.2 million. As the expenditures index shows, total purchases have grown 28 percent.

Table 1b uses an alternative measure of medical output. Instead of defining output as the number of procedures performed, this table defines it as the number of lives saved (alternatively, we could use quality of life or some other measure of medical outcome). According to these figures, there has been a

[4] Brostoff (1993) describes a study by Lewin-VHI, Inc., that estimated the costs of defensive medicine to be $36 billion per year.

Table 1a Measuring Aggregate Price and Quantity Changes: Medical Procedures and Food

	Medical Procedures	Food	Indexes (year 0 = 100)
Year 0			
Expenditures	500,000	9,000,000	100
Quantity	1,000	100,000	100
Price	500	90	100
Share of economy	5.3%	94.7%	
Year *t*			
Expenditures	1,200,000	11,000,000	128
Quantity	2,000	110,000	115
Price	600	100	112
Share of economy	9.8%	90.2%	

Interpretation: Index = 128 implies 28 percent growth over the period. Calculation of year *t* indexes:

$$\textit{Expenditure:} \quad 128 \approx 100 \times \frac{1{,}200{,}000 + 11{,}000{,}000}{500{,}000 + 9{,}000{,}000}$$

$$\textit{Price:} \quad 112 \approx 100 \times (5.3\% \times \frac{600}{500} + 94.7\% \times \frac{100}{90})$$

$$\textit{Quantity:} \quad 15 \approx 100 \times \frac{\textit{Expenditure Index}}{\textit{Price Index}}$$

dramatic quality change in the medical procedure. Thirty percent of the patients survive in year *t* (600 out of 2,000), compared with only 10 percent in year 0 (100 out of 1,000). Because of this, the price of one life saved has *dropped* from $5,000 to $2,000, compared with an *increase* in the price per procedure from $500 to $600.

Inflation is measured in Table 1b as 7 percent, compared with Table 1a's rate of 12 percent. Real economic growth is 15 percent in Table 1a and 20 percent in Table 1b. The practical effects of such measurement discrepancies are not trivial. Throughout the economy, wage contracts, government benefits, taxes, and other contractual arrangements tie payments to changes in the general price level. It matters to a company whether its workers should be given a 12 percent or a 7 percent cost-of-living increase.

For most purposes, it would be better to measure growth as in Table 1b rather than as in Table 1a, since it is lives saved and not procedures performed that indicate economic well-being. We can guess, for example, that improvements in X-ray machines and in doctors' abilities to read X-rays have led to a greater efficacy in the use of X-rays. How much sooner, on average, are

Table 1b Measuring Aggregate Price and Quantity Changes: Lives Saved and Food

	Lives Saved	Food	Indexes (year 0 = 100)
Year 0			
Expenditures	500,000	9,000,000	100
Quantity	100	100,000	100
Price	5,000	90	100
Share of economy	5.3%	94.7%	
Year *t*			
Expenditures	1,200,000	11,000,000	128
Quantity	600	110,000	120
Price	2,000	100	107
Share of economy	9.8%	90.2%	

Interpretation: Index = 128 implies 28 percent growth over the period. Calculation of year *t* indexes:

Expenditure: $128 \approx 100 \times \frac{1{,}200{,}000 + 11{,}000{,}000}{500{,}000 + 9{,}000{,}000}$

Price: $107 \approx 100 \times (5.3\% \times \frac{2{,}000}{5{,}000} + 94.7\% \times \frac{100}{90})$

Quantity: $120 \approx 100 \times \frac{\textit{Expenditure Index}}{\textit{Price Index}}$

cases of disease found in 1994 than in 1954 on a per-X-ray basis? How much more does the average X-ray extend or improve life today? Even if we could accurately answer these questions, what would be the dollar value of each improvement? Since the answers are difficult to even approximate, analysts in statistical bureaus with limited budgets usually shrug their shoulders and use the number of X-rays to represent output, rather than using some measure of abatement of disease or extension of life.

The difficulty in distinguishing quality, quantity, and price changes exists for all goods and services. For example, a pound of chicken in 1994 is not the same product as a pound of chicken was in 1924. The taste, consistency, and nutritional characteristics have all changed. Also, the qualities of a computer in 1994 are vastly different from what they were in 1974. At least for these tangible products, one can imagine how quality might be defined. With services, however, the difficulty in defining output makes it especially problematic to measure changes in the quality of that output. In no service industry is the effort more daunting than in medicine. Measuring medical care production in terms of the means (procedures) rather than the ends (good health) is somewhat

akin to measuring vegetable prices in dollars per acre planted rather than dollars per bushel of yield. The former would lead us to mistakenly measure increased yields per acre from added fertilizer as inflation.

New Goods Problem

Another serious indexing problem is that new products and technologies have been introduced rapidly into medicine (and other industries) in the last 50 years. Price index weights, however, are revised only infrequently. As a result, price indexes may miss significant reductions in the cost of living. Gordon (1992) writes that "penicillin entered the CPI in 1951, after it had already experienced a 99 percent decline from its initial price" (p. 9). Berndt, Griliches, and Rosett (1993) examined the new goods problem with respect to the introduction of new pharmaceuticals. They found that the Bureau of Labor Statistics (BLS) tends to give insufficient weight to newer products and that these products tend to experience lower-than-average price increases. Together, these two tendencies would bias the measured price increases upwards.

We can illustrate the mechanics of the new goods problem by departing from medicine for a moment and considering two familiar products from the electronics industry. Suppose a long-term price series used 1940 expenditure weights. There would be no weight for the transistor, and the skyrocketing price of vacuum tubes would appear to contribute to inflation. Of course, vacuum tube prices are up largely because the production volumes have become small. The invention of the transistor has greatly *reduced* the cost of devices that amplify and rectify electronic signals.

While the BLS deals with the new goods problem in several ways, the most common process is called "linking" in which, at some arbitrary point, one good is dropped and the other added. Importantly, the new good is introduced in such a way that this replacement leaves the price level unchanged.

The data in Table 2 provide a hypothetical example of linking. Suppose drug A is replaced over time by drug B, but that for a time both are on the market. The first two columns represent the prices of drug A and drug B in years 1 through 6. The price of drug A is rising due to general inflation and other factors. New products like B frequently will decline in price after introduction because [1] through experience the company entering the market improves its manufacturing techniques, [2] the new company increases its market share and can take advantage of economies of scale, and [3] close substitutes increasingly compete with profitable established products. The next-to-last column represents a price index that, beginning in year 2, reflects changes in the price of drug B. The last column represents an index that reflects drug A prices until year 5 and then switches to drug B. The problem is deciding at what point to drop drug A and to add drug B. This table shows that such a choice may completely change the message sent by the price index. Again, this problem

Table 2 Linking Old and New Goods in a Price Index

Year	(Old) Drug A Price	(New) Drug B Price	Price Index: Drug B Added in Year 2	Price Index: Drug B Added in Year 5
1	100	420	100 (base year)	100 (base year)
2	130	390	$93 \approx 100 \times 390 \div 420$	$130 \approx 100 \times 130 \div 100$
3	140	370	$88 \approx 93 \times 370 \div 390$	$140 \approx 130 \times 140 \div 130$
4	160	340	$81 \approx 88 \times 340 \div 370$	$160 \approx 140 \times 160 \div 140$
5	190	380	$90 \approx 81 \times 380 \div 340$	$179 \approx 160 \times 380 \div 340$
6	240	470	$112 \approx 90 \times 470 \div 380$	$221 \approx 179 \times 470 \div 380$

Notes: If drug B replaces drug A in the price index in year 2, the index shows a smaller price rise than if B replaces A in year 5. This is because the year 5 link misses drug B's price decline in the first few years. Calculations are not exact due to rounding.

may be more serious in medicine than in most other sectors of the economy. The relative costs and benefits of transistors and vacuum tubes can be defined in fairly objective terms, while new drugs are rarely as easy to compare.

Old Goods, New Label Problem—Generic Drugs

A variant on the new goods problem is the case in which an existing good is reintroduced to the market under a new label, as in the case of generic drugs. Following is a hypothetical example illustrating the generic drug problem described by Scherer (1993). Suppose that [1] a name-brand drug X costs one dollar per pill, [2] a biochemically identical generic drug Y is introduced at fifty cents per pill, [3] half the market switches from X to Y. If one treats X and Y as a single drug, then the average price has dropped by 25 percent, since purchasers of the pills are now paying 25 percent less on average than they used to.

In fact, this change will not normally show up in the CPI as a price reduction. First, weights in the CPI market basket are changed infrequently. CPI data are collected on specific brands, like our drug X. Until the weights are revised, the price of brand Y will not enter into the calculation of the CPI. Second, when the generic drug Y is added to the CPI market basket, it will be added into the index as a new product, separate from name-brand drug X. Thus, the addition of Y to the basket will not show up as a decline in price.

Price indexes indicate generally that pharmaceuticals prices have risen at a high rate compared with general inflation or even with other parts of the medical sector. Scherer (1993), Berndt, Griliches, and Rosett (1993), and Griliches and Cockburn (1993) examine this trend and conclude that mismeasurement is partly to blame. This mismeasurement occurs in part because of the way generic drugs are introduced into indexes such as the CPI.

Alternative Approaches to Measuring Medical Care Prices

Researchers have suggested alternative ways of measuring medical output that might yield better estimates of medical prices than do current procedures. Wilensky and Rossiter (1986) describe four ways of measuring medical care output: the procedure (e.g., one day's radiation therapy), the case (e.g., a cancer, from diagnosis to conclusion of treatment), the episode (e.g., a particular period of the illness), and per capita (e.g., the patient's total health care, including the cancer). Procedure-based indexes are the most commonly used today, but alternative indexes have been proposed that would use alternative units of output.

Health Insurance Premiums as Price Proxies

Some researchers have suggested that a good indicator of price increases might be found in the premiums paid on a standard health insurance policy.[5] The logic is that an insurance policy represents a fixed bundle of medical goods and services, and if quality remains constant, the price of the policy will represent the price of that bundle. This idea found some favor in the late 1960s, when, it can be argued, health insurance policies were fairly standardized. Problems with that approach have become apparent, however, as policies have grown less uniform, with broad differences in copayments, deductibles, payout limits, and services provided. Technological and other changes in medicine mean that a policy today provides very different care from an identical policy 30 years ago; thus, quality changes are as big a problem as they are with a procedure-based measure. Also, the real values of policies differ across states, since each state's regulatory practices partially determine the insurance companies' liabilities. Finally, the quantity of medical care demanded by the average policyholder differs across localities.

Costs of Treatment of a Representative Group of Illnesses

Scitovsky (1964) proposed taking a group of illnesses and measuring how the prices of treating those illnesses changed over time.[6] Instead of measuring inputs like hospital beds, operations, and drugs, this approach would take an occurrence of a number of illnesses—say, a case of pneumonia, a brain tumor, and a broken leg—and measure the total costs of treating this set of illnesses. Quality change would still be a problem, though, since the means of treating a particular disease changes over time. This proposal did suggest adjusting the measured treatments for quality, using indicators like infant mortality and age-adjusted death rates per numbers of cases as proxies for quality. Scitovsky was

[5] This idea is discussed in Feldstein (1993), pp. 71–72. Feldstein also refers readers to Reder (1969), p. 98, and Barzel (1969).

[6] This proposal is described in Feldstein (1993), pp. 64–71.

concerned, however, that simple quality adjustments such as these would be inadequate, given the complexity of measuring medical outcomes.

Hedonic Pricing

One method of adjusting for quality that has been used by statistical agencies and academic researchers is hedonic pricing. Hedonic pricing values a good by assuming that the good is really a bundle of characteristics and that there are separate demands for each of these characteristics. In measuring price changes in computers, for example, the Bureau of Economic Analysis uses a model that breaks the computer down into a set of characteristics (e.g., number of calculations per unit of time), and then measures the prices of those characteristics. Recombining these separate prices yields an estimated price for a computer, holding quality constant (see Triplett [1986]). In this approach, quality is merely the sum of a group of quantities.

An example of the hedonic approach applied to medical equipment is Trajtenberg (1990). He compares three price indexes for Computerized Tomographic X-ray devices (CT or CAT scanners): [1] a standard index with no adjustment for quality change; [2] a hedonic index, assuming that a CAT scanner is really a bundle of four characteristics (head vs. body, scan time, resolution, and image reconstruction time); and [3] a welfare-change index based on the same four characteristics, but designed for a very different objective—measuring the consumer's well-being rather than the price of these four characteristics. Over the period 1973 to 1982, the standard index *increases* from 100 to 259.4, the hedonic price index *decreases* from 100 to 27.3, and the welfare-change index decreases from 100 to .07. Thus, one methodology produces a price index 3,700 times higher than does another price index. As Getzen (1992) writes:

> Differences of this magnitude in only a few items would be sufficient to show that rather than being the fastest rising component of the [general price level], the real quality-adjusted price of medical care is falling—a conclusion that would be confirmed by most rational consumers given a choice of 1931 medicine at 1931 prices and the medical technology of 1991 at 1991 prices. (P. 116)

Hedonic pricing may provide a promising approach for some goods. However, the procedure adds markedly to the cost of data, and not all goods and services are good candidates for the procedure.

3. OTHER PROBLEMS

The passages below describe several index problems not associated with quality changes. They include [1] the use of list prices instead of transaction prices, [2] statistical sampling problems, [3] the measuring problems introduced by health insurance policies, and [4] substitution bias.

Transaction Versus List Prices

Price indexes may sometimes use data from list prices rather than actual transaction prices. Ideally, price indexes should include only transaction prices. In many medical care transactions, the discrepancies are large. For example, a hospital bill may state the charge for a procedure as $600, but Medicare may reimburse the hospital only $400. If the hospital receives no additional compensation from the patient or from private insurers, then $400 should be the price of the procedure used in compiling the price index. Unfortunately, it is often the case that list prices are easier to come by than transaction prices, so it is these fictional list prices that are used in the index.

If discounts (or the ratio of list to transaction prices) were constant over time, this problem would not be particularly pernicious. Medical discounts, however, have grown rapidly over recent decades, so the use of list prices appears to have imparted an upward bias to reported increases in medical care prices.[7]

Sampling Problems

In a world of costless data collection, an ideal index of medical care prices would incorporate the price of every single medical transaction that actually takes place, down to every individual box of aspirin sold. Collecting one price for each individual transaction, though, is impractical or impossible, so the producer of a price index must drastically reduce the number of prices collected by sampling. Instead of measuring the price of every single aspirin purchased in America in October, the statistician can more readily measure only the list price for brand X aspirin at five stores each in one hundred localities on October 12. The effects of such sampling are not neutral, and the sample may therefore misrepresent the total aspirin purchases nationally. Analysts have identified several ways in which typical procedures could distort price indexes. For example, list prices may be higher than actual prices paid because of routine store discounts. Much aspirin may be purchased by bulk users such as clinics who pay less than list prices. Brand X may be higher-priced than store brands. The localities and stores selected may be unrepresentative. And Columbus Day may be a poor day to sample prices because many stores will have one-day discounts.

Medical Insurance

Getzen (1992, p. 85) notes that the problem of measuring medical prices is further complicated by medical insurance. Most medical payments in the United

[7] For example, the hospital component of the CPI, which uses list price data, consistently rises faster than does either the HCFA Hospital Transaction Output Price Index or the PPI Hospital Services Index, both of which use transaction price data (see Table 3). Bottiny (1993, p. 32) cites figures showing that from 1984 to 1988, California hospital list prices (charges billed) increased by 11.1 percent annually, while transaction prices increased by 7.0 percent.

States are made through public or private insurance policies. Payments under these policies make it difficult or impossible to separate out the prices paid by specific individuals for specific procedures. Insurance has exacerbated the problem of "cost-shifting." This problem arises when one group of patients is charged more than the full cost of treatments in order to subsidize another group whose charges do not fully cover treatment costs. Health care providers often make up losses on Medicare and Medicaid patients by raising prices to other patients, thus causing some prices to be overstated and others to be understated. If (as with the medical component of the CPI) the sample mostly measures payments by non-Medicare, non-Medicaid patients, then an increase in cost-shifting will impart an upward bias to the index.

Choice of Weights and Substitution Bias

A price index is simply a weighted average of prices, and the weights are generally derived from the mix of items consumed across the economy. The consumption mix, though, changes dramatically over time in response to shifts in relative prices and other factors, and the choice of weights is important. In Table 1b, medical expenditures rise from 5.3 percent of output to 9.8 percent. Based on year 0 weights (5.3 and 94.7 percent), the price level rises from 100 to 107. Based on year *t* weights (9.8 and 90.2 percent), however, the price level would rise from 100 to only 104 (9.8 percent $\times$ 2000/5000 $+$ 90.2 percent $\times$ 100/90). In the U.S., most price indexes use the first method, infrequently changing weights.

A general principle in economics is that as the price of one good rises, consumers tend to shift at the margin out of that good and substitute into other goods whose prices are falling or rising more slowly. In Table 1b, for example, the shift in spending toward medical procedures may result from the decline in the price of one life saved relative to the price of one unit of food purchased. With fixed expenditure weights, these demand shifts will be missed and the price index will give too much weight in later years to the good whose price is rising fastest, a statistical phenomenon known as substitution bias. As a practical matter, substitution bias appears to be fairly small in most price indexes and is dwarfed by quality-measurement problems.

4. SUMMARY: RISING EXPENDITURES VERSUS RISING PRICES

This article has explained why some researchers suspect that the CPI and other indexes systematically overstate (or, possibly, understate) the rise in medical prices, though the case is difficult to quantify with any precision. True medical outputs (the number of lives saved, improvement in patients' quality of life, relief from pain, etc.) are difficult or impossible to measure. For this reason,

statisticians usually substitute quantities of inputs (number of coronary bypasses performed, number of hospital days), treating them statistically and semantically as if they were outputs. To some extent, this problem of disentanglement exists for all goods and services, but the undeniable but difficult-to-quantify progress in health care implies that the problem must be especially troubling in medicine. [See next page for a listing of papers and articles on this subject.]

Quality changes in medical care compound the problem, since a price index implicitly assumes that the quality of the underlying good or service does not change over time. In medicine, evolving technology and treatment regimes have steadily increased the quality of medical care over the past half century. To the extent that price indexes overstate medical inflation, these errors, in turn, will cause price indexes like the CPI to overstate general inflation. And the impact of any such errors may grow in the future because, according to some projections, medical care may grow from the present 13 percent of the national economy to 20 percent by 2010.

Perhaps the largest cost of measurement errors would be inappropriate policy decisions. Much of the present debate on health care reform is premised on the "fact" that medical prices have grown faster than those of most other goods and services. One can imagine that today's health policy debate and proposals would be very different were there a general perception that medical prices were growing slowly.

APPENDIX

A REFERENCE GUIDE TO PUBLISHED INDEXES

A number of price indexes are produced, each based on a segment of national health expenditures (NHE). The Consumer Price Index (CPI) is perhaps the best-known measure of aggregate price changes in the U.S. economy. Similarly, the Medical Care Price Index (MCPI), the medical component of the CPI, is the best-known measure of price changes in the medical sector and is often cited as representing "the" rate of inflation in medical care. It should be noted, however, that the MCPI covers a basket of goods and services that in many ways is unrepresentative of national health expenditures, as is explained later in this section.

Numerous other indexes measure medical care price changes—some in narrower ranges of transactions than those entering the MCPI, and some in broader ranges. The MCPI, however, must be considered the paramount medical price series. Data from the MCPI are used as proxies for prices and weights in producing most other medical care price series. Thus, whatever problems exist in the MCPI filter through into almost all other series. Many series also borrow

The following papers and articles discuss possible sources of biases in medical price series. The majority, though not all, presume that the biases are upward.

Article	Concerns Addressed Include
Armknecht and Ginsburg (1990)	The CPI may understate medical insurance cost increases.
Berndt, Griliches, and Rosett (1993)	The CPI fails to incorporate price decreases associated with generic drugs.
Bottiny (1993)	The CPI may overstate medical inflation because of [1] the exclusion of most government health expenditures (which have risen less than private payments) from the MCPI, [2] substitution bias, and [3] the heavy reliance on list prices rather than transaction prices.
Cleeton, Goepfrich, and Weisbrod (1992)	Lags in introducing new drugs, plus the lack of information on effectiveness and safety of drugs, may bias the CPI either upwards or downwards.
Getzen (1992)	Traditional price indexes are not suitable for use as deflators of health expenditures.
Griliches and Cockburn (1993)	The CPI fails to incorporate price decreases associated with generic drugs.
Kroch (1991)	The CPI fails to adjust for changes in the quality of medical care associated with, for example, hospital room modifications, nurse-to-patient ratios, and introduction of new technologies.
Lebow, Roberts, and Stockton (1992)	The CPI fails to incorporate price decreases associated with new goods.
Madigan (1991)	The CPI fails to adjust for improvements in the quality of medical care.
Newhouse (1988)	The source of medical expenditure increases cannot be determined because the CPI [1] measures the prices of inputs, not outputs, [2] uses list prices, not transaction prices, [3] largely ignores technological change, and [4] uses inappropriate weights.
Scherer (1993)	The CPI has shortcomings in how it absorbs generic drugs, new products, and quality improvements.
Trajtenberg (1990)	Because of quality changes, the CPI may dramatically overstate increases in CAT scanner prices.
Tregarthen (1993)	The CPI relies on list prices and fails to adjust for quality changes.

data from other series produced by the Department of Labor, the Department of Commerce, the Health Care Financing Administration, the American Medical Association, the American Hospital Association, and others. So, all the medical care price series tend to share many of the same methodological problems.

The following section contains comparative information on a number of currently available medical care data series. Entries generally include the following sections:

Coverage: The basket of goods and services whose average price the index measures

Purpose: The reason for producing the index

Years/Periodicity: The years of available data and the periodicity (e.g., annual, quarterly, monthly)

Source: The organization that produces the index

Reported: The publication in which data can be found

References: Articles or books explaining the index

Miscellaneous: Other pertinent information

Historical data on these series are found in Table 3.

Medical Care Price Index (MCPI)—Coverage: A basket of goods and services representing consumers' out-of-pocket medical expenditures—roughly 20 percent of the expenditures included in national health expenditures. Does not include most medical costs paid for by public or private insurance programs. Includes health insurance premiums paid directly by the consumer, but not those paid by employers or governments. **Purpose:** Comprises part of the Consumer Price Index (CPI). The CPI is widely used as a benchmark for adjusting contractual payments, including wage and Social Security payments, for inflation. **Years/Periodicity:** 1936–1946/quarterly; 1947–present/monthly. **Source:** U.S. Department of Labor, Bureau of Labor Statistics. **Reported:** In *Monthly Labor Review.* **References:** *BLS Handbook of Methods* (1992, ch. 19); Getzen (1992); Feldstein (1993). **Miscellaneous:** The absence of payments made by public and private insurance policies is a weakness if one is using the MCPI as a proxy for overall medical inflation; however, the MCPI is not produced with that purpose in mind.

National Health Expenditures (NHE) Deflator—Coverage: All medical care goods and services included in National Health Expenditures, a measure of total medical care spending. **Purpose:** To measure price movements in the entire medical sector. **Years/Periodicity:** Series under development as of November 1994. **Source:** Health Care Financing Administration (HCFA). **Reported:** Available on request from HCFA, Office of the Actuary.

Table 3 Annual Percentage Change in Medical Care Price Indexes

Price Index	Dec-29 to Dec-51	Dec-51 to Dec-65	Dec-65 to Dec-69	Dec-69 to Dec-80	Dec-80 to Dec-93	Years of Data
CPI	3.9%*	1.3%	4.3%	7.5%	4.0%	1935–93
MCPI	2.8%*	3.2%	6.1%	7.9%	7.5%	1935–93
Medical Care Commodities	2.7%*	0.7%	0.3%	4.9%	7.1%	1947–93
Professional Medical Services			5.8%*	7.4%	6.5%	1967–93
Dental Services	3.1%*	2.3%	5.4%	6.9%	6.5%	1935–93
Eye Care					3.8%*	1986–93
Hospital and Related Services				11.7%*	9.0%	1977–93
NHE Deflator	Data series under development as of November 1994					
PCE, Fixed-Weight, Medical Component	2.5%	3.4%	6.6%	7.8%	6.7%	1929–93
PHCE Deflator		2.3%*	5.4%	7.4%	7.1%*	1960–91
MEI					3.7%	1980–93
AHA Hospital		3.2%*	6.4%	8.0%	6.7%	1963–93
HCFA PPS Hospital					4.6%	1980–93
HCFA Hospital Transaction Output Price Index		3.5%*	6.6%	8.2%	6.7%	1960–93
HCFA Nursing Home				8.3%*	4.9%	1972–93
NHA-BEA Nursing Home				8.0%*	5.2%	1972–93
HCFA Home Health				8.3%*	5.7%	1972–93
PPI: Drugs and Pharmaceuticals	−2.8%*	−0.5%	0.0%	5.4%	6.5%	1947–93
PPI: X-Ray/Electromedical				8.5%*	1.7%	1971–93
PPI Hospital Services					4.0%*	1992–93

*Data available for only part of the period (see right-hand column for dates).

Notes: All series are discussed in text, except the five MCPI components. The time periods approximately delineate periods in which medical prices were subject to distinctive influences, as follows: 1935–1951: moderate technological change, most payments made out-of-pocket by patients, Great Depression, World War II; 1951–1965: faster technological change, rapid growth of private medical insurance; 1965–1969: introduction of Medicare and Medicaid; 1969–1980: high general inflation, low economic growth, rapid technological progress; 1980–1993: lower general inflation.

Personal Consumption Expenditures (PCE), Fixed-Weight Index, Medical Component—Coverage: Payments for individuals' medical care—approximately 88 percent of national health expenditures. Includes payments made by individuals and by public and private insurance programs. Does not include expenditures such as medical research and certain construction expenses. **Purpose:** PCE comprises part of the National Income and Product Accounts, and fixed-weight price indexes are produced for an array of NIPA segments. **Years/Periodicity:** 1929–1946/annual; 1947–present/quarterly. **Source:** U.S. Commerce Department, Bureau of Economic Analysis. **Reported:** *Survey of Current Business.* **References:** Getzen (1992), p. 96. **Miscellaneous:** The BEA formerly produced a deflator of the PCE medical component, but these data are no longer distributed.

Personal Health Care Expenditures (PHCE) Deflator—Coverage: Includes public and private spending for direct health and medical services to individuals. Included are expenditures for hospital care, physician services, dental services, other professional services, drugs and other medical nondurables, vision products and other medical durables, and nursing home care. Does not include medical research, construction of medical facilities, public health activities (e.g., disease prevention and control), program administration, and the net cost of private health insurance. **Purpose:** To provide a broad-based measure of medical care inflation that addresses some of the methodological problems inherent in the MCPI—the CPI's narrow expenditure base, for example. **Years/Periodicity:** 1960–1991/annual. **Source:** Health Care Financing Administration (HCFA). **Reported:** *Health Care Financing Review.* **References:** Letsch (1993).

HCFA Medicare Economic Index (MEI)—Coverage: Inputs to physician office services (roughly 25 percent of national health expenditures), plus an adjustment for economy-wide productivity growth. Inputs include physician earnings, nonphysician earnings, office expenses, medical materials and supplies, professional liability costs, medical equipment, and some other goods and services. **Purpose:** Used in annual updates of Medicare's physician fee schedule. The Secretary of Health and Human Services considers the MEI in recommending a new schedule to Congress. If Congress takes no action, the MEI is used in calculating an automatically updated schedule. **Years/Periodicity:** 1980–present/quarterly; ten-year forecasts. **Source:** Health Care Financing Administration (HCFA). **Reported:** *Federal Register.* **References:** For a fuller description of the data sources and of the Medicare Economic Index in general, see Office of the Federal Register (1992, 1993) and Freeland, Chulis, Arnett, and Brown (1991). **Miscellaneous:** By congressional intent, the MEI is backward-looking rather than forward-looking because Congress believed that increases in Medicare reimbursements should follow, rather than lead, inflation.

AHA Hospital Market Basket Index—Coverage: Hospital expenditures—roughly 40 percent of national health expenditures in 1991. It is an input

price index for hospitals, measuring the changes in prices of hospital inputs—the goods and services hospitals buy. **Purpose:** To serve as a guideline in contract negotiations between hospitals and their contractors. Deflates hospital expenditures over time in order to produce measures of real hospital spending growth. **Years/Periodicity:** 1963–present/monthly. **Source:** The American Hospital Association (AHA). **Reported:** Quarterly in AHA's *Economic Trends*. **References:** This index and the HCFA Hospital Market Basket Index are compared in Dyer and Li (1990). **Miscellaneous:** Uses fixed-expenditure weights that do not vary over time.

HCFA Prospective Payment System (PPS) Hospital Input Price Index—Coverage: Hospital expenditures—roughly 40 percent of national health expenditures in 1991. Input price index for hospitals. Measures the changes in prices of the goods and services hospitals buy as inputs into their production of goods and services. Used in the Medicare PPS update formula to adjust hospital reimbursements for year-to-year inflation. **Purpose:** To provide a regulatory baseline for adjusting the schedule of fees paid to hospitals under Medicare and Medicaid. **Years/Periodicity:** 1986–present/quarterly. Backcast data also have been produced for 1980–1986. **Source:** Health Care Financing Administration (HCFA). **Reported:** *Federal Register.* **References:** Office of the Federal Register (1990); Freeland, Anderson, and Schendler (1979); Freeland, Chulis, Brown et al. (1991), Freeland and Maple (1992). The HCFA and AHA indexes are compared in Dyer and Li (1990). **Miscellaneous:** Uses fixed-quantity weights, where the quantities are fixed from a base year but relative importance shares change over time as prices change.

HCFA Hospital Transaction Output Price Index—Coverage: Estimates the price of hospital outputs rather than inputs. To do so, the index uses list price data to estimate transaction price data. **Purpose:** Seeks to measure the rate of growth in transaction prices (rather than list prices) for hospital goods and services. Because of increasing volume discounts for large purchasers, list prices may overstate the actual growth in costs. **Years/Periodicity:** 1960–1993/annual. **Source:** Health Care Financing Administration (HCFA). **Reported:** Not formally reported, but available through HCFA. **References:** Fisher (Spring 1992, Fall 1992). **Miscellaneous:** There are two versions of this index. One uses patient revenues, while the other uses total revenues, of which patient revenues are only a part.

HCFA Regulation Skilled Nursing Home Input Price Index—Coverage: A market basket of the most commonly used nursing home inputs—approximately 8 percent of national health expenditures. **Purpose:** To reimburse skilled nursing facilities' inpatient routine service costs under Medicare. **Years/Periodicity:** 1972–present/quarterly. **Source:** Health Care Financing Administration (HCFA). **Reported:** Biannually in the *Federal Register*. **References:** Office of the Federal Register (October 7, 1992).

National Health Accounts–Bureau of Economic Analysis (NHA–BEA) Nursing Home Input Price Index with Capital Costs—Coverage: Inputs, including capital, for the production of nursing home services. **Purpose:** To estimate and project growth in nursing home prices while holding constant content of per-diem services, productivity, and profit margins. **Years/Periodicity:** 1972–present/quarterly. **Source:** Health Care Financing Administration (HCFA). **Reported:** Available on request from HCFA, Office of the Actuary.

HCFA Regulation Home Health Agency Input Price Index—Coverage: Goods and services used in producing home health care services—just over 1 percent of national health expenditures. **Purpose:** To determine reimbursement limits under Medicare. **Years/Periodicity:** 1972–present/quarterly. **Source:** Health Care Financing Administration (HCFA). **Reported:** Periodically in the *Federal Register*. **References:** *Federal Register* (July 7, 1992).

Producer Price Index (PPI): individual medical components—Coverage: Medical goods sold by producers, including both intermediate and final goods. The PPI covers goods used as inputs to medical care, though there is no aggregate index of medical producer prices. Two of the most important categories are drugs and pharmaceuticals and X-ray and electromedical machinery. **Purpose:** To construct the overall PPI. **Years/Periodicity:** Drugs and pharmaceuticals: 1947–present/monthly; X-ray and electromedical machinery: 1971–present/monthly. **Source:** U.S. Department of Labor, Bureau of Labor Statistics. **Reported:** *Producer Price Indexes* monthly publication of data. **References:** Various PPI releases from the BLS. **Miscellaneous:** Traditionally, the PPI has covered only goods, so much of the medical care industry has been excluded. However, several areas of medical services have recently been added to the PPI's coverage (see PPI—Hospitals, below).

Producer Price Index (PPI) Price Indexes for Hospitals—Coverage: These indexes for various classes of hospitals (general, psychiatric, etc.) are based on output data—the revenues paid to hospitals for an average hospital stay or outpatient treatment. **Purpose:** In 1993, the BLS began producing indexes of hospital prices as part of a long-range plan to incorporate service industries into the PPI. **Years/Periodicity:** 1993–present/monthly. **Source:** U.S. Department of Labor, Bureau of Labor Statistics. **Reported:** *Producer Price Indexes* monthly data publication. **References:** U.S. Department of Labor (1993), p. 5. **Miscellaneous:** Similar indexes have been or will be introduced in 1994 for physician services, medical laboratories, and nursing care facilities.

REFERENCES

Armknecht, Paul A., and Daniel H. Ginsburg. "Improvements in Measuring Price Changes in Consumer Services: Past, Present, and Future," in Zvi Griliches, ed., *Output Measurement in the Service Sectors, NBER Studies in Income and Wealth Volume 56*. Chicago: University of Chicago Press, 1992.

Barzel, Yoram. "Productivity and the Price of Medical Services," *Journal of Political Economy,* vol. 77 (November–December 1969), pp. 1014–27.

Berndt, Ernst R., Zvi Griliches, and Joshua G. Rosett. "Auditing the Producer Price Index: Micro Evidence from Prescription Pharmaceutical Preparations," *Journal of Business and Economic Statistics,* vol. 11 (July 1993), pp. 251–64.

Bottiny, Walt. "Is Medical Care Consumer Price Inflation Overstated?" *DRI/McGraw-Hill Cost and Price Review,* Second Quarter 1993, pp. 31–33.

Brostoff, Steven. "Eliminate Defensive Medicine, Save $36: Study," *National Underwriter,* vol. 97 (February 8, 1993), p. 5ff.

Cleeton, David L., Valy T. Goepfrich, and Burton A. Weisbrod. "What Does the Consumer Price Index for Prescription Drugs Really Measure?" *Health Care Financing Review,* vol. 13 (Spring 1992), pp. 45–51.

Dyer, Carmela, and Weiwei Li. "A Comparison of the AHA and HCFA Market Basket Indices." American Hospital Association Policy Brief #90-01. March 23, 1990.

Feldstein, Paul J. *Health Care Economics,* 4th ed. Albany: Delmar Publishers, 1993.

Fisher, Charles R. "Hospital and Medicare Financial Performance Under PPS, 1985–90," *Health Care Financing Review,* vol. 14 (Fall 1992), pp. 171–83.

_____. "Trends in Total Hospital Financial Performance Under the Prospective Payment System," *Health Care Financing Review,* vol. 13 (Spring 1992), pp. 1–16.

Freeland, Mark, and Brenda T. Maple. "U.S. Health Input Price Index Developed by HCFA," *Health Expenditure Analysis Letters,* vol. 1 (Summer 1992). Philadelphia: Temple University School of Business and Management.

Freeland, Mark S., Gerard Anderson, and Carol Ellen Schendler. "National Hospital Input Price Index," *Health Care Financing Review,* vol. 1 (Summer 1979), pp. 37–61.

Freeland, Mark S., George S. Chulis, Ross H. Arnett, III, and Aaron P. Brown. "Measuring Input Prices for Physicians: The Revised Medicare Economic Index," *Health Care Financing Review,* vol. 12 (Summer 1991), pp. 61–73.

Freeland, Mark S., George S. Chulis, Aaron P. Brown, David Skellan, Brenda T. Maple, Naphtale Singer, Jeffrey Lemieux, and Ross H. Arnett, III. "Measuring Hospital Input Price Increases: The Rebased Hospital Market Basket," *Health Care Financing Review,* vol. 12 (Spring 1991), pp. 1–13.

Getzen, Thomas E. "Medical Care Price Indexes: Theory, Construction, & Empirical Analysis of the US Series 1927–1990," *Advances in Health Economics and Health Services Research,* vol. 13 (1992), pp. 83–128.

Gordon, Robert J. "Measuring the Aggregate Price Level: Implications for Economic Performance and Policy," Working Paper 3969. Cambridge, Mass.: National Bureau of Economic Research, January 1992.

Griliches, Zvi, and Iain Cockburn. "Generics and New Goods in Pharmaceutical Price Indexes," Working Paper 4272. Cambridge, Mass.: National Bureau of Economic Research, February 1993.

Kroch, Eugene. "Tracking Inflation in the Service Sector," Federal Reserve Bank of New York *Quarterly Review,* vol. 16 (Summer 1991), pp. 30–35.

Lebow, David E., John M. Roberts, and David J. Stockton. "Economic Performance Under Price Stability," Working Paper 125. Washington: Board of Governors of the Federal Reserve System, April 1992.

Legorretta, Antonio P., Jeffrey H. Silber, George Costantino, Richard W. Kobylinski, and Steven L. Zatz. "Increased Cholecystectomy Rate After the Introduction of Laparoscopic Cholecystectomy," *JAMA,* vol. 270 (September 22/29, 1993), pp. 429–32.

Letsch, Suzanne W. "DataWatch: National Health Care Spending in 1991," *Health Affairs,* Spring 1993, pp. 105–10.

Madigan, Kathleen. "How Reliable Is the Consumer Price Index?" *Business Week,* April 29, 1991, pp. 70–71.

Newhouse, Joseph P. "Measuring Medical Prices and Understanding Their Effects," *Rand Paper P-7448,* 1988.

Office of the Federal Register, National Archives and Records Administration. *Federal Register,* vol. 58 (December 2, 1993), pp. 63864–65. Washington: U.S. Government Printing Office.

_____. *Federal Register,* vol. 57 (November 25, 1992), pp. 55896–55913. Washington: U.S. Government Printing Office.

_____. *Federal Register,* vol. 57 (October 7, 1992), pp. 46178–79. Washington: U.S. Government Printing Office.

_____. *Federal Register,* vol. 57 (July 7, 1992), pp. 29412–13. Washington: U.S. Government Printing Office.

_____. *Federal Register,* vol. 55 (September 4, 1990), pp. 35990–36080. Washington: U.S. Government Printing Office.

Reder, Melvin. "Some Problems in the Measurement of Productivity in the Medical Care Industry," in V. Fuchs, ed., *Production and Productivity in the Service Industries*. New York: Columbia University Press, 1969.

Scherer, Frederick M. "Pricing, Profits, and Technological Progress in the Pharmaceutical Industry," *Journal of Economic Perspectives,* vol. 7 (Summer 1993), pp. 97–115.

Scitovsky, Anne. "An Index of the Cost of Medical Care—A Proposed New Approach," in Solomon J. Axelrod, ed., *The Economics of Health and Medical Care*. Ann Arbor: Bureau of Public Health Economics, University of Michigan, 1964.

Trajtenberg, M. "Economic Analysis of Product Innovation: The Case of CT Scanners," *Harvard Economic Studies,* vol. 160. Cambridge, Mass.: Harvard University Press, 1990.

Tregarthen, Suzanne. "Statistics Overstate Health Care Costs," *The Wall Street Journal,* August 18, 1993, editorial page.

Triplett, Jack E. "The Economic Interpretation of Hedonic Methods," *Survey of Current Business,* vol. 66 (January 1986), pp. 36–40.

United States Department of Labor, Bureau of Labor Statistics. *Producer Price Indexes: Data for January 1993.* Washington: Government Printing Office, 1993.

_____. *BLS Handbook of Methods*. Washington: Government Printing Office, 1992.

Wallace, William H., and William E. Cullison. *Measuring Price Changes: A Study of the Price Indexes,* 4th ed. Richmond: Federal Reserve Bank of Richmond, 1981.

Webb, Roy H., and Rob Willemse. "Macroeconomic Price Indexes," in Roy H. Webb, ed., *Macroeconomic Data: A User's Guide,* 3d ed. Richmond: Federal Reserve Bank of Richmond, 1994.

Weisbrod, Burton A. "Productivity and Incentives in the Medical Care Sector," *Scandanavian Journal of Economics,* vol. 94 (1992), pp. S131–S150.

Wilensky, Gail R., and Louis F. Rossiter. "Alternative Units of Payment for Physician Services: An Overview of the Issues," *Medical Care Review,* vol. 43 (Spring 1986), pp. 133–56.