Limited Commitment and Central Bank Lending

Marvin Goodfriend and Jeffrey M. Lacker

Central bank lending is widely regarded as a vital part of the public safety net supporting the stability of the banking system and financial markets. An independent central bank can provide liquidity to financial institutions on very short notice.[1] Indeed, central bank lending has been a prominent part of regulatory assistance to troubled financial institutions in recent years. The idea of a central bank as lender of last resort, however, has been around at least since Walter Bagehot wrote about it over 100 years ago.[2]

For most of that time it was taken for granted that central bank lending had benefits with little or no cost. In the past decade, that view has been challenged. For instance, in the United States the Federal Deposit Insurance Corporation Improvement Act (FDICIA) of 1991 recognized that Federal Reserve lending to undercapitalized banks has the potential to impose higher resolution costs on the Federal Deposit Insurance Corporation (FDIC). More recently, the idea that lending by the International Monetary Fund has led to increased risk-taking in international financial markets is being taken seriously by financial market

This article was prepared for the Second Joint Central Bank Research Conference on Risk Measurement and Systemic Risk at the Bank of Japan, Tokyo, November 16-17, 1998. The authors are grateful for the comments of Urs Birchler and Doug Diamond on an earlier draft and the comments of Tom Humphrey, Bob Hetzel, John Walter, and John Weinberg on this version. The article also benefitted from presentations at the Konstanz Seminar on Monetary Theory and Policy, Konstanz, Germany; the Center for Financial Studies Conference on Systemic Risk and Lender of Last Resort Facilities, Frankfort, Germany; and the Federal Reserve Bank of Cleveland's Payments System Workshop. The authors' views do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Because a central bank can create money, it has the option of financing lending with an increase in the money supply. We would call such lending a combination of monetary policy and credit policy. When we speak of central bank lending in this article, however, we confine ourselves to pure credit policy. Pure central bank credit policy finances loans with proceeds from the sale of securities (Goodfriend and King 1988).

[2] See Humphrey and Keleher (1984).

participants and policymakers alike.[3] In the United States, financial economists have acknowledged "moral hazard" to be a problem for government financial guarantees ever since the savings and loan crisis of the 1980s.

In this article we look at central bank lending in light of the concerns about moral hazard. Our aim is a practical one: we present principles to help guide central bank lending. Our approach builds on the observation that central bank lending is a publicly provided line of credit. Commercial lines of credit and central bank lending are similar in that both provide substantial funding on very short notice.

Line-of-credit products are complex. We use recent advances in the theory of financial contracts to interpret the structure of loan commitments. By dissecting the incentive implications of the contractual obligations and rights involved in credit lines, we appreciate the tensions present in line-of-credit relationships. In particular, we see how contract terms control the ex post incentives of the borrower and the lender under limited commitment to assure that the line-of-credit product is efficient. We then employ our understanding of these issues to benchmark and inform our analysis of central bank lending.

The nature of the problem is this: A line-of-credit product is designed to meet the current obligations of a firm when it is judged to be illiquid though solvent. Inevitably, then, a loan commitment shifts potential losses from short- to longer-term claimants. For instance, a commercial bank's line of credit to an ordinary business has the potential to shift losses to the borrowing firm's long-term bondholders and residual claimants. Analogously, a central bank's line of credit has the potential to shift losses from uninsured creditors to the deposit insurance fund or general taxpayers. Likewise, lending by the International Monetary Fund to finance a country's balance-of-payments deficit has the potential to shift losses from short-term creditors of that country to the country's taxpayers.[4]

Private line-of-credit agreements, together with the firm's capital structure, balance the liquidation costs of a conservative lending policy against the moral hazard associated with more liberal lending. Covenant provisions in line-of-credit agreements give private lenders the ability and the incentive to constrain credit to insolvent firms when appropriate. In contrast, central banks appear to lack explicit institutional mechanisms to credibly precommit to limit lending. An excessively liberal central bank line of credit makes short-term capital more inclined to move in the direction of favorable yield differentials irrespective

[3] Strictly speaking the International Monetary Fund is not a central bank since it does not have the power to create money. Nevertheless, it is financially a relatively independent governmental organization, and it does make large loans on relatively short notice to countries in financial distress (Masson and Mussa 1995).

[4] Some dilution of long-term claimants is desirable, however, to avoid socially inefficient liquidation (Diamond 1993).

of the risk involved, with the idea that the credit line could finance a quick withdrawal.

The inability to commit to limit lending is the principal weakness of central bank lending policy. The problem is that central banks responsible for the stability of the financial system are inclined to lend whenever not lending could plausibly trigger a systemic crisis. That inclination encourages both domestic and international "hot money" investments—short-term investments that implicitly rely on central bank liquidity support for repayment in the event of a crisis—and thereby increases the scope for violent reversals and flights to safety whenever the market begins to doubt central bank lending intentions. We are agnostic about whether there is a welfare-enhancing role for central bank lending. The critical policy problem is how to limit central bank lending to socially appropriate circumstances.

The article proceeds as follows. Section 1 contains a description of the structure and mechanics of private lines of credit. In Section 2, central bank lending is characterized as a line of credit and the line-of-credit analogy is exploited to identify the nature and source of the undesirable consequences of lending by central banks. In Section 3, we consider how well some actual and possible components of central bank lending policy cope with the problem of limited commitment. We conclude that no simple institutional mechanisms could confidently precommit a central bank to limit its lending. Reasoning by analogy to the historical reduction of inflation, we argue that the only way for a central bank to credibly limit lending is for it to build up a reputation over time for lending restraint. Exploiting the inflation analogy further, we describe a sequence of events that we think will be necessary for a central bank to successfully acquire such a reputation.

1. THE ECONOMICS OF PRIVATE LINES OF CREDIT

The parallel between central bank lending and private lending under lines of credit is illuminating for the similarities and the differences that emerge (Goodfriend and King 1988). Both involve lending large amounts on short notice. However, private credit lines are explicit contractual commitments, while a central bank's commitment to lend is a matter of policy choice. In this section we review the economics of private lines of credit. We will focus in particular on what determines the contingencies under which private banks deny credit.

The Line of Credit Product

Lines of credit (loan commitments) specify a maximum amount that can be borrowed and a formula that determines the interest rate on advances, or "take-downs." Borrowing rates are usually set as a fixed markup over a reference rate such as the LIBOR or the lending bank's prime rate. Borrowers pay

an up-front fee when the line of credit is initiated, as well as an annual "commitment fee" proportional to either the undrawn portion or the entire amount of the commitment (Crane 1973, Schockley 1995). Line-of-credit lending is generally secured by collateral, although the largest and most creditworthy borrowers can obtain unsecured loan commitments. Some loan commitments provide "back-up" support for commercial paper issued by the firm; the loan is drawn down in the event that the firm cannot roll over its maturing paper. In this case the line of credit provides a bank guarantee for the liquidity of the commercial paper issued by the firm, assuring holders of an orderly exit in adverse circumstances (Calomiris 1998).

Loan commitment agreements contain covenants that place restrictions on the borrower's future financial condition. If the borrower violates one of the covenants, the lender has the right (though not the obligation) to terminate the agreement and demand immediate repayment. Some covenants utilize specific financial indicators—minimum net worth, minimum working capital, or maximum leverage ratio, for example. Other covenants restrict the disposition of assets or the issuance of other debt.

Loan commitment agreements also generally contain a clause that allows the bank to declare a default in the event of any "materially adverse change in the financial condition of the borrower." This ambiguously worded clause provides a backstop to the other formal covenants, allowing the lender to terminate lending when the borrower's financial condition deteriorates, even if the specific covenants are technically satisfied. At the same time, a borrower that is in good financial health can be assured that the bank is still obligated to lend.

Because the markup does not vary with subsequent changes in the borrower's creditworthiness, the line of credit represents an implicit insurance arrangement—a credit risk derivative. The implicit ex post insurance payout in a given state of the world is the present value of the difference between the contractual markup and the risk premium appropriate to that borrower in that state of the world. The contract does not provide full insurance, however, because the bank can limit large payouts by invoking covenants and denying credit. This partial insurance is valuable to borrowers as a way of smoothing the cost of contingent funding across various states of the world. Without a line of credit, the firm would pay a high risk premium if it needed funds when creditworthiness had deteriorated. With a line of credit, the firm pays ex ante fees and agrees to the possibility that credit is denied in some states in order to assure ex post access to funds at a lower risk premium. The ex ante fees compensate the bank for the implicit insurance provided.

Lines of credit tend to be provided by financial intermediaries, in general, and banks, in particular. By diversifying over a large number of risks that are to some degree independent, banks can offer insurance-like products at low cost. Bank loan officers specialize in evaluating creditworthiness, and are ideally

suited to monitor the borrower's condition over the life of the commitment. Such information gathering, built up through repeated interactions with the borrower, is crucial in evaluating later requests by the borrower to take down credit. In addition, bank monitoring activities save costs for other creditors. Historically, lending and related credit evaluation activities often have been combined with the issue of demand deposits (Goodfriend 1991, Nakamura 1993). Because of these advantages, banking institutions have traditionally dominated the line-of-credit business.

Agency Problems

Modern theory explains financial contracts as the result of ex ante negotiation among contracting parties in the context of competition from alternative borrowers and lenders. Contractual provisions help control agency problems—adverse incentives that may arise due to asymmetric information during the course of a contractual relationship. Bargaining is presumed to lead to contractual arrangements that are efficient in the sense that no other feasible contracts would make one party better off without making some other party worse off. Competition ensures that no contracting party is worse off than it would be if it contracted with another party instead.[5]

When banks lend to commercial firms, the critical agency problem is managerial moral hazard. Many managerial actions are difficult or impossible to specify as explicit conditions of the contract, either because they are not easily verifiable by the lender or a court, or because their complexity makes them too costly to include. Continuing to operate the business often yields private benefits to the manager-borrower, known as "control rents," which are impossible to transfer to outsiders. The manager may have significant human capital tied to the existing organization and operation, the value of which might be lost or diminished in a closure or liquidation. Also, the manager may enjoy perquisites from controlling the cash flow of the firm. More fundamentally, inducing the manager to take actions that benefit the firm might require giving the manager a pecuniary interest in the firm's profits. Borrowers and lenders may in some circumstances have conflicting interests over such actions. When the net worth of the firm is low, the manager's interest in the continuation of the firm strongly resembles an option; the manager would reap much of the upside gain in the business, while the costs of a deterioration would affect mainly the creditors. The manager can have a distorted incentive to make "all-or-nothing" gambles on excessively risky prospects.

If left unchecked, the moral hazard problem at a firm tends to grow over time. Losses erode net worth to the point where risk incentives shift. The firm begins to seek out investments with large potential payoffs, hoping to gamble

[5] See Harris and Raviv (1991, 1992) for surveys of the financial contracting literature.

its way back to health. The cost of such investments is below-normal rates of return under conditions in which the large payoffs are not realized. As a result, net worth is most likely to erode further, exacerbating the moral hazard problem. Each round of losses further strengthens risk-taking incentives.

Moral hazard can involve more than just the borrower. Other creditors will adopt a strategy that depends on the behavior of the firm's line-of-credit provider. If a lender pulls a line of credit that backs up a commercial paper program in a situation in which the borrower does not have the funds to roll over maturing claims, the firm defaults and investors may take a loss. The rate of return on the commercial paper will therefore reflect market expectations about the future funding behavior of the lender. Overly lax lending policy will show up as an inappropriately small risk premium on the firm's commercial paper or as an overly generous willingness to lend on the part of private investors. This issue is crucial for firms with illiquid assets that wish to issue liquid liabilities, because their creditors will be particularly concerned about prospects for future liquidity. A lender who is confident of the solvency of the firm will be willing to lend, while a lender who believes that the firm is insolvent will likely withdraw funds.

At the time the lending contract is negotiated, the contracting parties will anticipate the agency problems that could arise. Financial contracts deal with agency problems in two ways. First, contractual conditions explicitly constrain a manager's decisions. Such constraints show up in lending agreements as loan covenants, which we discuss in detail below. Second, contractual provisions affect the contingencies which force a change in control that removes the manager of the firm from a decision-making role. Liquidation is a leading example; the firm's tangible assets are sold and the proceeds are distributed to creditors. A "reorganization" supervised by a bankruptcy court is another type of change in control; management is often removed, but even when it remains in place its decisions are sharply constrained while the firm is under court-sponsored supervision.

Changes in control serve three purposes in the context of the agency problems that afflict lending arrangements. First, removing existing management prevents further value-wasting actions. Second, separating management from the quasi-rents associated with controlling the firm acts as a pecuniary punishment that helps provide ex ante incentives to manage the firm faithfully. Third, control changes facilitate restructuring the firm's liabilities in order to realign them with changed circumstances and allow repayment of creditors that wish to terminate their relationship with the firm.

Liquidation will be efficient ex post if it maximizes the total value of the firm. Inefficient liquidation—selling the firm's assets for less than the value of the firm as a going concern—reduces the total expected value of the firm when it occurs, and thus reduces the ex ante expected value of the firm. Ex ante both parties will prefer provisions that reduce the likelihood of inefficient

ex post liquidation. On the other hand, managerial control rents are extinguished when the firm is liquidated. The loss of these rents is a social cost of liquidation. Since control rents can only accrue to the managers, lenders will not take them into account in deciding when to liquidate. The cost of transferring control rights to lenders is that they will want to liquidate too often—when liquidation value exceeds the value as an ongoing concern, excluding control rents. Efficient liquidation rules balance the benefit of control changes against the cost of inefficient liquidation (Diamond 1993).

Credible Commitments

The circumstances under which control changes take place are determined by contractual terms (as well as the implicit background rules embodied in the relevant legal codes) that determine the assignment of property rights under various contingencies. The borrower and the lender will have an incentive ex ante to design contractual provisions so that ex post decisions about liquidation and the allocation of control rights are efficient, in the sense that they maximize the expected ongoing value of the concern as a whole, subject to the constraints imposed by the agency problems they face.[6] Loan covenants and collateral provisions play a central role in structuring the ex post incentives to effect control changes under line-of-credit arrangements.

Loan Covenants

Under the conditions defined in the covenants, the lender has the right to withdraw funding. If the borrower cannot obtain funding elsewhere, as is likely (see discussion below), the lender can essentially force reorganization or liquidation. Absent violation of the covenants, the borrower retains control of the firm. Loan covenants thus can be viewed as a means for conditionally transferring control of the reorganization/liquidation decision to the lender. Covenants also control other forms of ex post moral hazard directly by limiting the manager's right to take on new risks, change lines of business, assume new indebtedness, and so on (Aghion and Bolton 1992, Berlin and Mester 1992).

Loan covenants can be quite strong. In practice, however, the violation of a loan covenant is merely an occasion for renegotiation between lender and borrower. The lender can waive the violation or use the ability to declare (technical) default as leverage to obtain more favorable monetary terms or more stringent covenant conditions (a partial control transfer). Renegotiation allows outcomes to vary with ex post contingencies in ways that would be difficult to provide for ahead of time in a formal contract (Huberman and Kahn 1988, Kahn and Huberman 1989). Strict covenant restrictions can be adopted, with the expectation that in some circumstances they will be waived or loosened by

[6] Not all control changes are instigated by lenders; they can also take place at the initiative of the firm's governing board, presumably representing the interests of shareholders.

the lender. Although the borrower and the lender cannot precommit to refrain from renegotiating, the loan agreement can influence outcomes by ensuring that the allocation of property rights depends on future circumstances.

It makes sense, from an ex ante point of view, for the allocation of bargaining rights implied by loan covenants to depend on the riskiness of increased lending. When covenants are violated, managerial moral hazard is likely to be more pronounced. If further lending is to take place, the lender must do as well as if it withdrew the credit line and forced reorganization or liquidation. In this case covenants put the lender in a position to insist on a higher markup or more collateral to compensate for the heightened risk of continued lending. If the lender cannot be satisfied—if no such terms or collateral exist—then further lending is, presumably, ex post inefficient or infeasible, and the borrower is insolvent. When covenants are fully satisfied, managerial moral hazard is likely to be muted and so the lender does not need the right to prevent further lending. The bargaining power rests with the borrower, who is quite likely to be solvent in this case. Lending takes place at the borrower's request at the pre-agreed rate. The ex post self-interest of lenders, the ability to renegotiate, and the presence of relatively strict loan covenants provide a contractual mechanism that credibly commits the lender to limit lending when appropriate.

If given the choice ex post, the lender would never want to extend new lending to an insolvent firm. A firm is insolvent when the present discounted value of future cash flows falls short of the real current value of liabilities. Without a positive gap between future receipts and future obligations, the present value of anticipated future repayment streams cannot possibly cover the value of additional loans. Lending in such circumstances would represent subsidization, and a profit-maximizing lender has no reason to subsidize customers under competitive conditions.[7]

Collateral

The secured lender's ability to seize collateral for nonpayment is an important contractual right. A lien on an asset that is essential to the borrower's operations can provide the lender with another means of forcing the borrower's liquidation. In addition, collateral reduces the lender's risk by providing compensation when the borrower cannot pay the obligation in cash, therefore allowing a lower risk markup. Collateral also sharpens the borrower's incentive to repay, which helps relax borrowing constraints by allowing larger credible repayment obligations (Lacker 1998). Moreover, in bankruptcy, secured debt has a priority claim on the pledged assets. Collateral thus prevents dilution of the lender's position.

[7] The control rents enjoyed by the manager should, strictly speaking, be counted as part of the total value of the firm as a going concern, but since (by definition) these rents cannot be pledged to outsiders, they are irrelevant to financing decisions.

The lender's ability to take new assets as collateral later in the lending relationship helps overcome the classic underinvestment problem associated with debt overhang (Stulz and Johnson 1985). When the value of the firm is below the nominal value of outstanding debt, part of the return to any investment accrues to current debtholders; the real value of their debt increases. By pledging collateral, the borrower and the new lender can appropriate and share between them much of the gains from the new investment. Junior lenders can prohibit financing new projects with secured debt by including a "negative pledge clause" that prohibits pledging collateral to other lenders. Many junior creditors do not do so, however, since a negative pledge clause has the potential to prevent value-enhancing investments. For many publicly issued bonds, the firm retains the right to finance new projects with secured debt. Note that the presence or absence of a negative pledge clause for junior debt is a matter of contract. Note also that the lender's decision to take additional collateral is subject to ex post rationality constraints; it must be in the lender's self-interest to do so.

It is important to recognize that collateralized lending is not perfectly safe. The value realized by seizing and disposing of collateral is uncertain, and in some circumstances can fall short of the nominal obligation it backs. This feature is no accident, since borrowers have a greater incentive to default and surrender collateral when its value has fallen below the value of the debt. Why would lenders agree to terms under which they may take a loss on collateral? As previously noted, the key role of collateralized debt is to enhance the repayment incentive of the borrower. Collateral that is worth more to the borrower than to the lender, perhaps because of the transactions costs associated with liquidating the collateral, can provide adequate repayment incentives even though the lender suffers a loss when the borrower defaults and transfers the collateral (Lacker 1998). Moreover, collateralization alters ex post bargaining positions in any renegotiation by the borrower and the lender.

Monitoring

As mentioned above, line-of-credit lending is accompanied by costly information gathering. Banks assess the borrower's credit risk prior to the contractual commitment in order to set contract prices appropriately and to screen inappropriate risks. After the lending commitment has been signed, ongoing monitoring takes place, partly in the form of periodic financial statements required by covenant, and partly through informal contacts. Note that any arbitrary information gathering can, in principle, be negotiated as part of the commitment agreement. For example, many agreements stipulate that the lender receive audited financial reports. In other cases, particularly for small firms, the burden of audited statements is judged too costly and unaudited reports are accepted instead. When the borrower and the lender negotiate the monitoring features of

the contract, they presumably balance the marginal value of gathering additional information against the expected incremental joint cost.

Lenders have a strong incentive to gather information on an ongoing basis in order to be able to assess the solvency of the borrower as accurately as possible. Periodic monitoring thus helps prepare the lender to make critical decisions when the borrower experiences financial distress (Rajan and Winton 1995). What is learned about the characteristics of the firm's cash flow can help the lender interpret payment problems and more accurately assess the value of the firm as a going concern. Such information will be useful when the lender decides whether to extend or deny credit in response to covenant violations. In comparison, a lender with no prior lending relationship with the borrower will be at a distinct informational disadvantage.

Information gathering gives rise to "relationship lending" in which ties between lenders and borrowers are typically long lasting (Berger and Udell 1995, Petersen and Rajan 1994, Petersen and Rajan 1995, and Sharpe 1990).[8] This effect is particularly acute in times of distress, when outsiders are unable to acquire information fast enough to assist the firm on the same terms. The informational hurdles facing alternative lenders make the current lender's decision to grant or deny credit all the more crucial. When the informational advantage of a lending relationship enables a firm to obtain funds at a low enough cost to continue operating, and that same firm would have been unable to obtain funds cheaply enough without that relationship, we can say that the firm is illiquid though solvent. Withdrawing credit in this setting can effectively force reorganization or liquidation.

Safeguards for the Borrower

From the borrower's point of view, the important feature of loan covenants is that they define the limits of the lender's power to abrogate the agreement and demand accelerated payment. If the covenants are not violated, the lender is compelled to lend. As the lending relationship matures over time, the quasi-rents associated with the lender's informational advantage over competing lenders will grow. If the lender had blanket authority to demand repayment, the lender would be tempted to extort concessions from even a financially healthy borrower. All the quasi-rents from the relationship would inevitably accrue to the lender. To safeguard the borrower against such opportunistic behavior, the line-of-credit agreement stipulates that the lender is compelled to lend at a pre-agreed risk premium, absent any violation of the covenant conditions.

To summarize, then, line-of-credit agreements are crafted to address anticipated moral hazard problems that may arise if the borrower later gets into

[8] Relationship lending can also arise outside of formal line-of-credit lending.

trouble. In the presence of loan covenants and collateral provisions, a lender's profit motive allows it to credibly commit to making appropriate decisions to withdraw credit and induce closure or reorganization. Costly periodic monitoring enhances the lender's ability to gauge the borrower's situation.

2. CENTRAL BANK LENDING AS A LINE OF CREDIT

In this section we describe the similarities and differences between central bank lending and lending under private loan commitments. We consider central bank lending practices against the benchmark of private lending mechanisms, without prejudging the usefulness of public line-of-credit lending.[9] The critical difference is that the profit motive provides private line-of-credit lenders with ample incentive to limit lending ex post in the event of borrower adversity. The comparable incentive for central banks is relatively weak. Indeed, the commitment problem facing a central bank is the opposite of that facing a private lender; a lender needs to commit to lend in situations in which it might *not want* to lend, while a central bank needs to forego lending when it might *want* to lend.

Central Bank Lending

At first glance, central bank lending would appear to be quite different from private line-of-credit lending. Central banks do not generally negotiate contractual terms with individual borrowers. Instead, they are given statutory authority to lend to broad classes of institutions. Central banks are publicly chartered institutions and, unlike private lenders, profit maximization is not their primary objective.

Despite these apparent differences, central bank lending functions in fundamentally the same way as a private line of credit—by providing guaranteed access to borrowed funds at a predetermined rate. The rate at which central banks lend is generally posted in advance rather than negotiated ex post with each individual borrower. Thus central bank lending rates do not appear to vary much with the borrower's ex post creditworthiness. At times, distressed borrowers turn to the central bank because terms offered by private lenders would be exorbitant, either in the cost of explicit financing or because the terms would require surrender of control. Access to central bank credit therefore appears to provide implicit insurance to those that qualify. One difference between the pricing of central bank credit and private lines of credit is that central banks generally do not charge explicit ex ante fees for the service, although one could

[9] See Goodhart (1988) and Schwartz (1992) for alternative views on the desirability of central bank lending.

argue that the central bank commitment is bundled together with an array of regulatory burdens (and privileges).[10]

In its classic rationale, central bank lending is intended to help illiquid but solvent financial institutions meet their maturing short-term obligations. In the extreme case, central bank lending might fund a run on demand deposits. Note that this function closely parallels the role of bank lines of credit in backing up commercial paper programs. The facility is designed to help a firm cope with an emergency "run"—an inability to roll over its credits. As noted above, a decision to withdraw credit can trigger default on the commercial paper and closure or reorganization of a firm.

Compare private and central bank lending with respect to the mechanism that links credit withdrawal and closure. A private lender denies credit, causing a default, which leads creditors to seek remedies by seizing assets. The borrower files for bankruptcy to obtain protection from creditors so that a division of the losses can be negotiated without destroying firm value. A central bank that denies credit to a bank forces the hand of the chartering agency or the deposit insurance fund. The central bank's critical role in bank closure brings it face-to-face with the government agencies that have direct responsibility for closing banks.

Agency Problems

A vast array of bank management decisions involves risk-return trade-offs. Attitudes toward risk are to some degree distorted at any leveraged entity, because some decisions affect the value of debtholders' claims. Banks are among the most highly leveraged of institutions. At well-capitalized banks, the value of future control-rents is an asset that acts as an implicit performance bond that offsets risk-taking incentives. When net worth falls, however, the value of the implicit bond vanishes and incentives flip toward risk-taking—little is left to lose. It is widely recognized that the management of a poorly capitalized bank has incentives to take on excessive risks in an attempt to gamble its way out of trouble. When supervisory restraint is lax—as during the U.S. savings and loan crisis, or in the recent emerging markets banking crises—moral hazard steadily grows as the losses pile up (Calomiris 1998).

Private banks make explicit case-by-case decisions to grant lines of credit. In contrast, central bank lending commitments are not usually made on an individual basis. Often legislative and regulatory policies delimit the set of institutions that have access to central bank credit. Sometimes the set of

[10] See Kwast and Passmore (1997) for evidence on the net subsidy provided by the financial safety net in the United States.

institutions with access is quite large.[11] The key difference is that private institutions are able to condition the commitment on an examination of the prospective borrower's financial health and then tailor the contractual terms to the individual borrower. In contrast, access to central bank credit is granted to broad categories of institutions. Also, the terms of central bank lending do not reflect the competitive discipline of arm's-length bargaining.

Central bank supervision of institutions with access to central bank credit is a direct counterpart to the ongoing monitoring performed by banks. Supervisory reports, like the periodic financial statements provided to line-of-credit lenders, keep authorities apprised of changes in the creditworthiness of the prospective borrower. Even for central banks without a direct supervisory role, access to such information performs the same function. Supervisory information is generally far more detailed than the reporting required of private line-of-credit customers. As noted earlier, private contracts can, in principle, mandate stricter disclosure, but there are impediments to doing so. In the United States, provisions of bankruptcy law discourage lenders from becoming so intimate with the management of the firm as to be deemed an "insider" (Baird 1993).

Like private line-of-credit lending, central bank lending is generally collateralized. Specific assets can be documented and evaluated in advance, drawing on the central bank's supervisory knowledge. In addition, the security interests of central banks are generally favored in bank failure resolutions. This fact tends to make central bank lending relatively safe, although, as noted above, collateralized lending is not risk-free in general.

When central banks lend to government-insured institutions, collateral plays a crucial role in the loan's effect on the insurance fund in the event of a failure. Collateralized lending dilutes junior claimants, which in the case of an insured bank includes depositors. The insurance fund stands in for the depositors in the event of closure, however, so central bank lending effectively dilutes the deposit insurance fund. For example, in the United States, the FDIC assumes the failed bank's indebtedness to the Federal Reserve and in exchange retains the pledged assets. When the Fed lends to allow a failing bank to pay maturing short-term obligations the insurance fund retains the collateral, but the maturing short-term obligations have been replaced by a fixed obligation to the Fed. If the short-term claimants whose funds were withdrawn are insured depositors, the operation has merely replaced one fixed obligation for another. It is a different matter, however, if some short-term claimants are uninsured. The

[11] In the United States, for example, all depository institutions that are subject to reserve requirements are eligible to borrow at the Federal Reserve's discount window. In addition, Section 13 of the Federal Reserve Act allows the Board of Governors to authorize the Reserve Banks "in unusual and exigent circumstances" to extend credit to any individual, partnership, or corporation, provided the Reserve Bank obtains evidence that such entity "is unable to secure adequate credit accommodations from other banking institutions."

short-term claimants would have shared in the losses with the FDIC had the central bank not lent.[12] Instead, the insurance fund inherits a bank in which an uninsured claim held by the private sector is replaced by a fixed senior claim held by the Federal Reserve. In the process, closure is delayed and private uninsured creditors are spared.

The Commitment Problem

With private lines of credit, lender profit maximization provides an incentive to advance credit only when it is ex post efficient to do so. The environment surrounding central bank lending is quite different. A central bank has a legislated responsibility for the stability of the financial system as a whole: it could be blamed for any negative consequences of not lending. A central bank that precipitates the demise of one or more financial institutions may be subject to direct action through the legal system or indirect action through the legislature. It is impossible to prove the counterfactual, i.e., that not lending and letting a troubled firm fail would not seriously disrupt markets. Furthermore, it is difficult for outsiders to question, after the fact, a central bank's judgment on such matters. For all of these reasons, central banks are inclined to lend whenever financial stability is at all threatened.

Central banks are careful to protect their loans by taking collateral. In fact, some central banks lend only on terms that virtually guarantee repayment in full. In the United States, for example, discount window loans are virtually always collateralized, assuring priority in closure (Hackley 1973). Moreover, the FDIC generally assumes the debt that the borrowing bank owes the Fed in exchange for the collateral, relieving the Fed of the risk of falling collateral value. This arrangement allows the Reserve Banks to avoid loan losses but has the effect of shifting losses to the deposit insurance agency (Marino and Bennett 1999).

Implicitly restricting central bank lending to be risk-free by taking collateral is a "bright line" policy that is easy to verify ex post. Such a policy is one way to limit central bank involvement in the allocation of credit and to restrict the scope for subsidization. Limits to the central bank's involvement in credit allocation can help buttress the central bank's independence and bolster the fiscal discipline of the deposit insurance fund (Goodfriend 1994). One might think that such a bright-line no-loss policy would sharpen the central bank's incentives, bringing them more closely in line with those of a private line-of-credit provider. By itself, however, taking collateral is not enough, because the central bank then has no pecuniary reason not to lend.

[12] This presumes the current depositor preference regime. In the absence of a depositor preference law, the short-term claimants would have been junior to the FDIC's claim. See Birchler (forthcoming) and Marino and Bennett (1999) for discussion of depositor preference law. Marino and Bennett also discuss the role of Federal Reserve lending in delaying closure of failed banks.

Lending by the central bank creates a potentially severe moral hazard problem. Markets expect the central bank to provide the bank with the funds to allow the exit of uninsured liquid claimholders. Thus, lending by central banks facilitates a reallocation of wealth among the creditors of a failing bank that the deposit insurance fund has neither the capability nor the legal authority to perform by itself. Private lending to a failing firm is subject to the safeguards of bankruptcy law. This includes the fraudulent conveyance provision, which under certain conditions allows the court to unwind transactions, including loan agreements, that occurred immediately prior to bankruptcy if such agreements disadvantaged the bankrupt firm's estate. Collateralized central bank lending accompanied by indemnification from the deposit insurance fund is subject to no such formal discipline, only the vagaries of the political system.[13]

The financial stability mandate can create pressure to expand the scope of central bank lending to nonbank financial institutions. Nonbank financial intermediaries are capable of amassing sizable financial market positions. The liquidation of these positions could be seen as a threat to the stability of asset prices and the solvency of many other financial institutions, including insured banks. A central bank with no formal authority to lend outside a narrowly defined set of institutions is, of course, well positioned to resist influence. Otherwise, we might see a tendency to expand the range of institutions receiving central bank line-of-credit assistance.[14]

We conclude that the incentives for a central bank to limit lending are relatively weak. As a result, we should expect to see a tendency for central banks to overextend lending, creating moral hazard problems among institutions deemed likely to qualify for central bank credit. Moreover, the rate of incidence of financial distress that calls for central bank lending should tend to increase over time as market participants come to understand the range of the central bank's actual (implicit) commitment to lend and adjust expectations accordingly.

3. COPING WITH THE COMMITMENT PROBLEM

To summarize the argument so far, we have seen how commercial banks efficiently and profitably structure contracts to support private lines of credit. They do so because (1) their own money is at stake, (2) they can choose their borrower relationships, (3) the conditions include the right to monitor the value of assets on an ex ante (ongoing) basis to distinguish illiquid from insolvent borrowers in the event of a request for funds, (4) loan covenants give the lender the right to withdraw credit when the borrower's financial condition

[13] For an account of Federal Reserve lending to depository institutions, see U. S. Congress (1991). See also Marino and Bennett (1999).

[14] For an account of Federal Reserve lending to nonbanks, see Garcia (1990).

has deteriorated, and (5) competition and profit maximization induce private providers to balance the risks of accommodating a request for funds against the costs of not lending. To be competitive, the terms of the line-of-credit product must not exploit borrowers; and to be profitable, the credit line must provide a risk-adjusted return comparable to products offered by other banks.

Central banks provide lines of credit under such different circumstances that we cannot presume they will make lending decisions appropriately. First, financial losses are not borne by the central bank but by the Treasury, and, ultimately, taxpayers. Second, a central bank cannot offer "take-it-or-leave-it" conditions because it is responsible for protecting financial markets as a whole and may not be able to refuse to lend to an institution whose failure might threaten the system. Third, for the reason mentioned above, a central bank might feel pressure to lend to an institution that it does not examine thoroughly, or at all. Fourth, a central bank is not disciplined by competition or profit maximization.

At any point in time, then, a central bank will be more inclined to lend whenever not lending could threaten the entire financial system. Such incentives ensure that the central bank carries out its legislative mandate to stabilize financial markets. The problem is that the inclination to lend creates in the public's mind an expectation that a financial institution in a protected class can count on credit assistance from the central bank in certain adverse future circumstances. Private lenders will take advantage of central bank assistance by monitoring less and accepting greater credit risks when lending to implicitly protected firms. Further, borrowing firms in the protected class will take advantage, too, by taking on increasingly risky assets. Over time, the central bank will be inclined to expand the class of firms perceived to be protected and the extent of protection.

The fundamental problem is to find a way to credibly commit to limit lending.[15] It is a difficult problem and there are no easy solutions. In what follows we consider the practical effectiveness of five broad approaches to the commitment problem.

Good Offices Only

In lieu of establishing a practical means of committing a central bank to refrain from lending except in deserving circumstances, we could imagine legislation precluding a central bank from extending its own credit under any circumstances. This possibility is worth considering because a central bank could still play a useful and effective role in facilitating private credit transactions or those of other national or international agencies. A central bank has three institutional

[15] Some question the need for any discount window lending at all. See Goodfriend and King (1988) and Schwartz (1992). Adherents of this view can interpret our analysis as an exploration of the means by which a central bank might limit its lending in practice.

strengths in this regard. First, its financial independence and independence from the budget process makes it impartial with respect to financial matters, unlike most other government agencies, or, for that matter, firms in the private sector. Second, a central bank has a large staff with practical experience in economics, supervision and regulation, payments system operations, and financial law. Third, in the course of carrying out their normal duties, high central bank officials develop personal relationships with their counterparts in the private sector.

Thus, a central bank could offer its "good offices" to help private creditors negotiate a troubled financial firm's recapitalization. The central bank might have knowledge of the troubled firm through existing supervisory relationships. Also it might be in a position to "certify" the solvency of the firm to others, essentially facilitating "due diligence" efforts. Even in the absence of ex ante central bank knowledge of the institution, the central bank might inspect the portfolio for others, acting as a trusted third party. Furthermore, in negotiations among members of a potential lending consortium, the central bank might play the role of neutral arbitrator.

In principle, the extension of good offices need not involve pressure or sweeteners from the central bank. In practice, however, as long as a central bank retains supervisory and regulatory powers, one could not be sure whether private parties to the agreement were influenced implicitly by a concern about punishment should they not sign on to a deal. In effect, then, a deal could have been facilitated by implicitly directed credit allocation because of the central bank's involvement. The parties could also believe that regulatory authorities, including the central bank, would forbear if the institutions that lent became troubled themselves. Of course, a deal could very well involve a considerable transfer of equity from the original owners to the new owners of the troubled firm. If a central bank presides over a deal more favorable to the original owners than they would have received without its help, moral hazard has increased.

One way to ensure that no implicit pressure or sweeteners are involved when a central bank uses its good offices would be to take the central bank out of bank supervision and regulation. But then the central bank would lose the professional and personal connections that make it a good facilitator in the first place. The upshot is that even limiting a central bank's role to one of facilitator tends to create in the public's mind the possibility of assistance of one kind or another.

Lending Hurdles

Recognizing that there are circumstances when central bank lending would be desirable in order to protect the financial system, we consider various hurdles designed to limit the central bank's inclination to lend except in extreme circumstances and to limit its own exposure if it does lend. We deal with these

issues in reverse order. First, we consider the taking of collateral. After that, we consider the effectiveness of hurdles that a central bank might be made to clear before it is authorized to lend in the first place.

Collateral

Some central banks lend only on good collateral to fully protect their funds in the event that the borrower cannot repay. The taking of good collateral certainly protects the financial integrity of central banks themselves. As discussed above, however, collateralized lending does not limit the exposure of the insurance fund and taxpayers.

Its lending well protected, a central bank would have little incentive to precipitate a borrower's insolvency by refusing to lend. When a central bank supervises a borrowing bank, it is in a good position to evaluate the illiquid portions of a portfolio for purposes of collateral and can keep a bank operating for some time. In effect, central bank lending provides uninsured creditors of a troubled bank with free insurance (which encourages uninsured creditors to invest at shorter maturities) and delays the time when a troubled bank would default to one of its creditors and trigger its closing and reorganization. Assets that could have remained in the bank, if it had been closed sooner, are pledged to the central bank and are unavailable to help the deposit insurance fund and the taxpayers pay off insured deposits. Full collateralization of central bank lending conceals the fact that such lending exposes the insurance fund and the taxpayer to a risk of loss.

Early Intervention

One option for better protecting the deposit insurance fund and the taxpayer is to require bank regulators to close a failing bank when its book value equity capital falls to, say, 2 percent rather than to the point of book insolvency. A deterioration of book capital could trigger progressively heavier regulatory restrictions. Such restrictions might prohibit additional central bank lending at some point, unless the highest officials in the government grant written permission to lend.[16]

The problem with this hurdle is that it is based on book rather than market value capital. When depository institutions have assets that are in large part illiquid non-traded loans, they could become insolvent on a market value basis well before they are declared insolvent on a book value basis. For example, consider the Bank of New England which was declared insolvent in January 1991. Soon after, the FDIC released estimates that the deposit insurance claim would cost the taxpayer around $2 billion. Why didn't the regulators act sooner?

[16] The "prompt corrective action" provisions of the FDICIA encourage early closure and help to restrict central bank lending in this way.

The Bank of New England's problems began when the mortgage loans it made in the mid-1980s turned bad. Real estate proved unable to earn a sufficient return to cover the loan payments. The bank, however, still had to pay competitive interest on deposits. So the bank had to divert to depositors a portion of the return on assets that had been going to equity holders. The cut in dividends caused the stock price to fall precipitously, and the bank could not meet the competitive deposit rate payments by reducing dividends alone. The bank had to sell off securities, pledge assets to the Federal Reserve's discount window, and obtain Treasury deposits in order to fund withdrawals of uninsured deposits and pay interest to the remaining depositors. The negative cash flow eventually reduced the book value net worth enough for regulators to seize the bank.

In this case it may be said that regulators were too slow in writing down the value of loans. It is well to remember, however, that there are often good reasons to be cautious. The market value of a loan is the present discounted value of future cash flows. Although current cash flows may be small, there is usually room for disagreement among analysts concerning future cash flows. Therefore, any write-down by a regulator is subjective and subject to challenge ex post by high government officials or by the bank in question itself. As a result, hurdles based on measured capital deficiencies that are designed to protect the deposit insurance fund and the taxpayer against losses due to excessive central bank lending might not work very well in practice.

Constructive Ambiguity

The above argument suggests that one cannot count on simple mechanistic hurdles to limit a central bank's inclination to lend. The problem is that financial markets know that there are circumstances in which a central bank would not refuse to lend to troubled institutions. Thus, owners of institutions that are big enough or central enough to the payments system or to financial markets more generally have an incentive to increase their risk exposure in just those circumstances. Owners know that they keep the upside returns if things go well, but share any losses more broadly, i.e., with the central bank, an insurance fund, or the taxpayer, if things go badly.

This sort of logic puts a central bank in a box. A central banker's willingness to support the financial system in times of potential crisis (to maintain the confidence necessary to facilitate the functioning of financial markets and the economy more broadly) actually causes risks in the system to grow. For this reason, a central bank might be inclined to keep markets guessing about the exact circumstances in which it would be willing to lend. By creating uncertainty in the minds of potential borrowers, such ambiguity might be thought to be constructive because it causes potential borrowers to take on less risk. Constructive ambiguity, under this interpretation, attempts to reduce market

participants' perception of the probability of central bank lending while reserving the central bank's option to lend when systemic concerns seem to require it.

Some ambiguity is unavoidable in any attempt to state the precise contingencies in which a central bank might lend. The true policy would depend on information available to the central bank at a future date, some of which might be private information about specific firms known only to the central bank. A policy that needs to be based on private unpublishable information would not be verifiable and so could not be made completely free of uncertainty and ambiguity. Moreover, lending policies that depend on future circumstances in complicated ways might be difficult to state with clarity in advance.

That said, one might ask whether a central bank might want to deliberately increase the uncertainty surrounding its lending intentions. At one level, ambiguity can be enhanced by not attempting to sharpen or clarify the broad principles of central bank lending in internal discussions or external speeches of high central bank officials. Over time, however, markets will learn the central bank's actual lending policy. If the central bank does not follow through with actions that ratify the announced ambiguity, its rhetoric will ultimately be disregarded. Market expectations will converge on the central bank's actual policy. To be sustainable, therefore, a policy of constructive ambiguity has to be demonstrated in a central bank's lending actions themselves.

In order to increase ambiguity, a central bank would have to add extraneous variability to its lending policy—it would have to play a "mixed strategy" in game-theoretic terms. In effect, a central bank would have to couple each lending decision with a spin of a roulette wheel that would randomly point to "follow through" or "not follow through." The central bank would need to be willing to abide by the wheel. That is, with some probability the central bank would lend when its better judgment said the situation did not call for it; and with some probability the central bank would have to follow the wheel and not lend when it would otherwise wish to do so.

Randomization can be economically useful. For example, tax authorities audit randomly, with audit probabilities that vary with some basic features of the return. Randomization balances the beneficial incentive effects on taxpayer behavior against the expected resource cost of the audits. Tax authorities are able to implement mixed strategies credibly because they have learned over time that failing to audit eventually leads to increased tax evasion.

The problem with adding variability to central bank lending policy is that the central bank would have trouble sticking to it, for the same reason that central banks tend to overextend lending to begin with. An announced policy of constructive ambiguity does nothing to alter the ex post incentives that cause central banks to lend in the first place. In any particular instance the central bank would want to ignore the spin of the wheel.

Constructive ambiguity in the absence of an ability to precommit may actually increase the drift toward expansion. The greater the perceived probability of lending by the central bank in various circumstances, the greater the risk-taking incentive for eligible institutions. Whenever the central bank is seen to lend in a situation in which it had not lent before, perceived probabilities will be revised upward, inducing greater risk-taking.[17]

Extended Supervisory and Regulatory Reach

A central bank could consider extending its supervisory and regulatory authority, or the authority of other government agencies, to all institutions to which it might possibly wish to lend. In principle, such authority would enable the central bank to limit risk-taking directly. A central bank might extend its regulatory authority to financial institutions, banking or otherwise, big enough or central enough to threaten the financial system if they failed.

There are many problems with attempting to control risks by extending regulatory authority. First, regulatory reach does not extend across international borders. An attempt to regulate financial firms too heavily may cause them to locate in those countries willing to impose little regulation in order to attract the business. Second, an attempt to extend regulation within a country causes new institutional forms to develop to escape regulation. Third, the proliferation of new financial instruments associated with derivatives enables institutions to synthesize financial positions in many ways. Sophisticated financial engineering has made circumventing regulatory restrictions much easier. It has become very difficult for regulators to monitor and regulate transactions, i.e., balance sheet and off-balance-sheet positions of a firm. This development prompted the movement from direct supervision of balance sheet items toward a supervisory philosophy focused on institutions' risk management and control processes.

If central banks extend supervisory and regulatory authority to a broader array of financial institutions, they risk a positive feedback effect on central bank lending policy. Supervisory involvement in a financial sector can "taint" government authorities with implicit responsibility for the health of institutions in that sector, heightening the perception that the central bank is willing to lend to them in the event of liquidity problems. A central bank might find it costly to disappoint such expectations. In other words, extending the breadth of supervision and regulation could induce a commensurate extension of the perceived central bank lending commitment.

Supervision and regulation has its place as part of a line-of-credit package, but it is oversold as a means of controlling risk-taking by firms that could

[17] Note that for the tax authority, the fraction of returns that are audited is published and may have far more impact on perceived audit probabilities than an individual audit. In contrast, because the frequency of central bank lending is much lower, individual instances may have a far greater effect on market expectations of future lending.

potentially benefit from having access to central bank lending on favorable terms.

Reputation Building

In our view, none of the above institutional mechanisms can credibly commit a central bank to limit its lending or prevent increased risk-taking induced by a central bank's inability to limit its lending commitment. However, we believe that a central bank could credibly commit to limit its lending by building a reputation for doing so. Given the pressures that a central bank faces, there might seem to be little hope that it could ever build a reputation for lending restraint. It is difficult to imagine how a central bank would begin to do so. Yet, we think that the experience by which central banks around the world have built a reputation for maintaining low inflation provides a road map for how they might credibly commit to limit lending.

Building a Reputation for Low Inflation[18]

In the 1960s, the inflation that accompanied stimulative monetary policy was tolerated as a necessary evil in the United States because it seemed consistent with a stable Phillips curve tradeoff between unemployment and inflation. In retrospect, however, we see that workers and firms came to anticipate deliberately expansionary monetary policy. Workers learned to take advantage of tight labor markets to make higher wage demands, and firms took advantage of tight product markets to pass along higher costs in higher prices. Increasingly aggressive wage and price behavior tended to neutralize the favorable employment effects of expansionary monetary policy, and the Federal Reserve became evermore expansionary in pursuit of low unemployment.

In the 1970s, disaffection with inflationary policy arose as the Phillips curve correlation broke down and both inflation and unemployment moved higher. In the late 1960s, the Fed began periodically to try to brake the acceleration of inflation with tight monetary policy, well aware that such policy actions caused unemployment to rise. The resulting stop/go monetary policy characterized the period from the mid-'60s until the early 1980s. Finally, the great disinflation introduced a period in which the Federal Reserve gradually acquired credibility for low inflation.

Two developments paved the way for the great disinflation. First was the progress that economists made in understanding the causes of inflation. This professional understanding reinforced the Fed's confidence that monetary policy could bring inflation down. Second, two decades of nonmonetary approaches

[18] This account is drawn from Goodfriend (1997).

to controlling inflation—for example, wage/price guidelines and controls, fiscal budget policy, and credit controls—had been tried and had failed.

By the time Paul Volcker became Federal Reserve Chairman in 1979, inflationary policy was widely recognized to have costs with no offsetting benefits. Previous experience with stop/go policy made clear that bringing inflation down would be costly too. Indeed, the inflation was not broken until a sustained tightening of monetary policy that began in 1981 created a serious recession that tested the Federal Reserve's determination and the public's support. With widespread public support, the Federal Reserve has maintained low inflation for almost two decades. Macroeconomic performance has been good compared to that of the inflationary period, and only one mild recession has occurred thus far—in 1990 to 1991.

Building a Reputation for Limited Lending

The analogy to the historical reduction of inflation provides a road map for a central bank that seeks to acquire a reputation for lending restraint. We might imagine the following sequence of events. Initially, the central bank and the public alike recognize only the short-term benefits of central bank lending. Central banks are inclined to extend emergency credit assistance to any institution whose possible failure could present even the most remote risk of disruption to the financial system. The liberal lending policy encourages potential beneficiary firms to take on more risks. Greater risk-taking, in turn, creates more frequent crises and causes the central bank to extend the scope of its lending even further. Policymakers and the public see the frequency and magnitude of financial crises grow even as the willingness of the central bank to lend increases.

Gradually, under this scenario, an understanding might emerge among policymakers and the public that excessively liberal central bank lending is counterproductive. The view would be supported by economists' improved understanding of the causes of increasing risk in the financial system and its relation to excessive central bank lending. As central bankers come to feel overextended, they might be more inclined to incur the risk of short-run disruptions in financial markets by disappointing expectations and by not lending as freely as before. The central bank might backtrack on its initial attempts to disappoint lending expectations. Eventually, the public might decide that the increased financial crises were, in part, due to excessively liberal central bank lending. The public would want the central bank to become more restrictive, even at the cost of precipitating a financial disruption by refusing to lend in a particular crisis. Ultimately, with the public's support and a consistent willingness to risk the consequences, a central bank would acquire a reputation for more limited lending. Financial firms might then take on less risk, and financial market crises might become less common.

One might wonder where we are in this process today. The parallel with monetary policy is again instructive. During the 20 years of great inflation there were four major episodes (1966, 1968, 1973-74, 1979-82) in which the Federal Reserve tightened monetary policy to restrain inflation with adverse consequences for employment. It was not until the savings and loan crisis of the mid-1980s that the public became aware of the greater risk-taking engendered by the government financial safety net, e.g., deposit insurance and central bank lending. To date, there are no instances in which a financial crisis has followed a refusal by the Federal Reserve to extend emergency credit assistance. Granted, provisions of the FDICIA of 1991 impose some constraints on Federal Reserve lending to failing institutions: lending to undercapitalized depository institutions is limited, except in circumstances involving "systemic risk" (requiring high-level certification), and the Fed is exposed to minor losses. These provisions, however, hardly constrain discount window lending; for example, it appears that Fed lending to Continental Illinois in 1984 would have met the requirements of the 1991 Act.

There is little evidence yet that the general public in the United States favors a significantly more restrictive lending policy for the central bank. One might regard the Bank of England's handling of the Barings closure as an instance of a move toward a more restrictive lending policy. But the parallel with monetary policy suggests that episodes of increasing severity may be necessary before central banks definitively alter course in the direction of lending restraint.

4. CONCLUSION

We have presented some guiding principles for central bank lending. Central bank lending should be regarded as a line of credit, and should be expected to exhibit the tensions inherent in private line-of-credit products. The most serious problem is managerial moral hazard, the borrower's incentive to take on more risk after arranging a credit line. We discussed in some detail contractual provisions (loan covenants, collateral, and monitoring) designed to control moral hazard. The key point is that contractual provisions enable profit-maximizing lenders to credibly commit to withdraw credit and induce the closure or reorganization of a borrowing firm when appropriate.

The contractual mechanisms utilized by private line-of-credit providers are less effective for a central bank whose primary mission—to maintain financial system stability—can override its obligation to protect public funds and undercut its ability to limit lending. We considered in some detail five broad approaches to a central bank's commitment problem: offering good offices only, intervening early and taking collateral, adopting a strategy of constructive ambiguity, extending supervisory and regulatory reach, and building a reputation. Our analysis suggested that the first four institutional approaches cannot be

counted on to overcome the fundamental forces causing a central bank to lend.

On the other hand, we believe that it should be possible for a central bank to build a reputation for limiting its lending commitment, just as central banks around the world acquired credibility for low inflation. In fact, we view the forces operating on central bank lending policy as analogous to those influencing the path of inflation. Liberal lending policy initially raises expectations of lending. There is more frequent lending, increased moral hazard, and greater financial instability. Gradually, policymakers and the public become willing to disappoint lending expectations. The economy then experiences a temporary period of heightened financial instability associated with increasingly restrictive lending, which is followed by less financial instability and little central bank lending. It would appear that we are still at the initial stages of what could be a lengthy process.

We are agnostic about whether central bank lending is beneficial. We put off consideration of that difficult question until central bank lending is more restrained, just as the debate on the desirability of low or zero inflation in the steady state was deferred until inflation was brought down sufficiently. Currently, the critical policy question is how to reverse perceptions that central banks are increasingly willing to lend, which increases risk-taking and the likelihood that central banks will feel compelled to lend. Just as monetary policymakers looked for opportunities to disinflate, we think that financial economists and central bankers should look for opportunities to restrain central bank lending.

REFERENCES

Aghion, Philippe, and Patrick Bolton. "An Incomplete Contracts Approach to Financial Contracting," *Review of Economic Studies,* vol. 59 (1992), pp. 473–94.

Baird, Douglas G. *The Elements of Bankruptcy.* Westbury, N.Y.: The Foundation Press, 1993.

Berger, Allen N., and Gregory F. Udell. "Relationship Lending and Lines of Credit in Small Business Finance," *Journal of Business,* vol. 68 (July 1995), pp. 351–81.

Berlin, Mitchell, and Loretta Mester. "Debt Covenants and Renegotiation," *Journal of Financial Intermediation,* vol. 2 (June 1992), pp. 95–133.

Birchler, Urs. "Bankruptcy Priority for Bank Deposits: a Contract Theoretic Explanation," *Review of Financial Studies* (forthcoming).

Calomiris, Charles W. "The IMF as Imprudent Lender of Last Resort," *Cato Journal,* vol. 17 (Winter 1998), pp. 275–94.

Crane, Dwight B. *Managing Credit Lines and Commitments.* Chicago: Trustees of the Banking Research Fund, Association of Reserve City Bankers, 1973.

Diamond, Douglas W. "Seniority and Maturity of Debt Contracts," *Journal of Financial Economics,* vol. 33 (1993), pp. 341–68.

Garcia, G. "The Lender of Last Resort: Recent Development and Nontraditional Examples." Manuscript. Washington: The Committee on Banking, Housing, and Urban Affairs of the United States Senate, December, 1990.

Goodfriend, Marvin, and Robert G. King. "Financial Deregulation, Monetary Policy, and Central Banking," Federal Reserve Bank of Richmond *Economic Review,* vol. 74 (May/June 1988), pp. 3–22.

Goodfriend, Marvin. "Monetary Policy Comes of Age: A 20th Century Odyssey," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 83 (Winter 1997), pp. 1–22.

_____. "Why We Need an Accord for Federal Reserve Credit Policy," *Journal of Money, Credit, and Banking,* vol. 26 (August 1994, Part 2), pp. 572–80.

_____. "Money, Credit, Banking and Payments System Policy," Federal Reserve Bank of Richmond *Economic Review,* vol. 77 (January/February 1991), pp. 7–23.

Goodhart, Charles. *The Evolution of Central Banks.* Cambridge: MIT Press, 1988.

Hackley, Howard H. *Lending Functions of the Federal Reserve Banks: A History.* Washington: Board of Governors of the Federal Reserve System, 1973.

Harris, Milton, and Artur Raviv. "Financial contracting theory," in Jean-Jacques Laffont, ed., *Advances in Economic Theory, Sixth World Congress,* II. Cambridge, Cambridge University Press, 1992.

_____, and _____. "The Theory of Capital Structure," *Journal of Finance,* vol. 46 (March 1991), pp. 297–355.

Huberman, Gur, and Charles Kahn. "Limited Contract Enforcement and Strategic Negotiation," *American Economic Review,* vol. 78 (June 1988), pp. 471–84.

Humphrey, Thomas M., and Robert E. Keleher. "The Lender of Last Resort: A Historical Perspective," *The Cato Journal,* vol. 4 (Spring/Summer 1984), pp. 275–318.

Kahn, Charles, and Gur Huberman. "Default, Foreclosure, and Strategic Renegotiation," *Law and Contemporary Problems,* vol. 52 (Winter 1989), pp. 49–61.

Kwast, Myron L., and S. Wayne Passmore. "The Subsidy Provided by the Federal Safety Net: Theory, Measurement and Containment," Working Paper 1997–58, Finance and Economics Discussion Series. Washington: Board of Governors of the Federal Reserve System, Divisions of Research and Statistics and Monetary Affairs, December 1997.

Lacker, Jeffrey M. "Collateralized Debt as the Optimal Contract," Working Paper 98-4. Federal Reserve Bank of Richmond, May 1998.

Marino, James A., and Rosalind L. Bennett. "The Consequences of National Depositor Preference," *FDIC Banking Review,* vol. 12 (1999), pp. 19–37.

Masson, Paul R., and Michael Mussa. "The Role of the IMF," Working Paper 50, Pamphlet Series. Washington, D.C.: International Monetary Fund, 1995.

Nakamura, Leonard I. "Recent Research in Commercial Banking: Information and Lending," *Financial Markets, Institutions & Instruments,* vol. 2 (December 1993), pp. 73–88.

Petersen, Mitchell A., and Raghurajan G. Rajan. "The Effect of Credit Market Competition on Lending Relationships," *Quarterly Journal of Economics,* vol. 110 (1995), pp. 407–43.

_____, and _____. "The Benefits of Firm-Creditor Relationships: Evidence from Small Business Data," *Journal of Finance,* vol. 49 (March 1994), pp. 3–47.

Rajan, Raghurajan G., and Andrew Winton. "Covenants and Collateral as Incentives to Monitor," *Journal of Finance,* vol. 50 (September 1995), pp. 1113–46.

Schockley, Richard L. "Bank Loan Commitments and Corporate Leverage," *Journal of Financial Intermediation,* vol. 4 (July 1995), pp. 272–301.

Schwartz, Anna J. "The Misuse of the Fed's Discount Window," Federal Reserve Bank of St. Louis *Review,* vol. 74 (September/October 1992), pp. 58–69.

Sharpe, Steven A. "Asymmetric Information, Bank Lending, and Implicit Contracts: A Stylized Model of Customer Relationships," *Journal of Finance,* vol. 45 (September 1990), pp. 1069–85.

Stulz, Rene M., and Herb Johnson. "An Analysis of Secured Debt," *Journal of Finance,* vol. 14 (1985), pp. 501–21.

U.S. Congress, House of Representatives, Committee on Banking, Finance, and Urban Affairs. *An Analysis of Federal Reserve Discount Window Loans to Failed Institutions.* 102 Cong. 2 Sess. Washington: Government Printing Office, 1991.

Sticky Prices, Marginal Cost, and the Behavior of Inflation

Alexander L. Wolman

A principal goal of economic modeling is to improve the formulation of economic policy. Macroeconomic models with imperfect competition and sticky prices set in a dynamic optimizing framework have gained wide popularity in recent years for examining issues involving monetary policy. For example, Rotemberg and Woodford (1999b) and McCallum and Nelson (1999) examine the behavior of model economies under a variety of monetary policy rules; Ireland (1995) examines the optimal way to disinflate; and Benhabib, Schmitt-Grohe, and Uribe (forthcoming) and Wolman (1998) study the monetary policy implications of the zero bound on nominal interest rates.[1] Nevertheless, serious questions remain as to whether these models accurately describe the U.S. economy, and therefore as to how one should interpret the results of this research.

One criticism of optimizing sticky-price models is that the relationship between output and inflation they generate is inconsistent with the behavior of these variables in the United States.[2] However, recent research by Sbordone (1998) and Galí and Gertler (1999) has breathed new life into these models by shifting attention away from the relationship between output and inflation and toward one between marginal cost and inflation—the latter being a more fundamental relationship in the models. If firms have some market power, as under imperfect competition, the behavior of their marginal cost of production is an important determinant of how they set prices. In turn, the overall price

The author thanks Mike Dotsey, Andreas Hornstein, Tom Humphrey, Bob King, Wenli Li, and Pierre Sarte for helpful comments and discussions. This article does not necessarily represent the views of the Federal Reserve Bank of Richmond or any branch of the Federal Reserve System.

[1] These are but a few of the many papers using such models. For a survey, see Taylor (1999).

[2] See Fuhrer and Moore (1995).

level and inflation rate are determined by aggregating individual firms' pricing decisions. There is then a clear relationship between the behavior of individual firms' marginal cost and the behavior of inflation. Sbordone (1998) and Galí and Gertler (1999) use this relationship to estimate and evaluate optimizing sticky-price models.[3] They find that such models can accurately replicate the observed behavior of inflation.

In this article, we work through the details of a sticky-price model, making explicit the relationship between marginal cost and inflation just described. We then offer a criticism of the specific form of price stickiness used by Sbordone and Galí and Gertler; essentially, they let prices be implausibly sticky. Plausible forms of price stickiness generate fundamentally different inflation dynamics and hence will be more difficult to reconcile with the behavior of marginal cost and inflation in the United States. However, the methodology introduced by Sbordone (1998) and Galí and Gertler (1999) remains a promising approach for evaluating sticky-price models. We suggest two ways in which this research agenda can continue progressing.

We concentrate on partial equilibrium analysis. The analysis takes as given the average inflation rate and the behavior of demand and real marginal cost. A complete general equilibrium version of our sticky-price framework would include descriptions of factor markets, consumer behavior, and monetary policy. In a general equilibrium, marginal cost and inflation would be endogenous; conditional on private behavior, policy would determine the behavior of inflation. Nonetheless, even in a general equilibrium, one would observe the relationship between marginal cost and inflation that is the focus of this article.

1. FROM INDIVIDUAL FIRMS' PRICING TO AGGREGATE INFLATION

Two central components comprise most of the recent optimizing sticky-price models: (1) monopolistic competition among a large number of firms producing differentiated products and (2) limited opportunities for price adjustment by individual firms. Monopolistic competition makes it feasible for some firms not to adjust their price in a given period; under perfect competition, only firms that charged the lowest price would sell anything. Limited price adjustment means that real and nominal variables interact; output and real marginal cost—both real variables—affect individual firms' pricing decisions, which in turn affect the price level and inflation.

[3] There is a separate literature dating back at least to the 1960s and continuing today that relates the behavior of inflation to marginal cost in reduced-form econometric models. See, for example, Eckstein and Wyss (1972).

Monopolistic Competition

The first component is monopolistic competition. The monopolistic competition framework most common in recent models is that of Dixit and Stiglitz (1977). The large number of firms mentioned above is represented mathematically by a continuum, and the firms are indexed by $z \in (0,1)$. Assume that these firms' differentiated products can be aggregated into a single good, interpreted as final output. If $y_t(z)$ is the amount produced by firm z, final output is

$$y_t = \left(\int_0^1 y_t(z)^{(\varepsilon-1)/\varepsilon} dz \right)^{\varepsilon/(\varepsilon-1)}. \tag{1}$$

With this aggregator function and market structure, demand for the good produced by firm z is given by

$$y_t(z) = \left(\frac{P_t(z)}{P_t} \right)^{-\varepsilon} y_t, \tag{2}$$

where $P_t(z)$ is the nominal price of good z, and P_t and the price of one unit of y_t. According to (2), demand for good z has a constant elasticity of $-\varepsilon$ with respect to the relative price of good z, and given the relative price, demand is proportional to the index of final output (y_t). The Appendix contains a detailed derivation of the demand function (2) and shows that the price index (P_t) is

$$P_t = \left(\int_0^1 P_t(z)^{1-\varepsilon} \right)^{\frac{1}{1-\varepsilon}}. \tag{3}$$

The price index has the property that an increase in the price of one of the goods has a positive but not necessarily one-for-one effect on the index. If that good's nominal price is lower (higher) than the price index, an increase in its price raises the price index more (less) than one-for-one, because the good has a relatively high (low) expenditure share.

Limited Price Adjustment

Limited opportunities for price adjustment constitute the second important component of our representative model. We assume that any firm $z \in (0,1)$ faces an exogenous probability of adjusting its price in period t and that the probability may depend on when the firm last adjusted its price. The probability of adjusting is non-decreasing in the number of periods since the last adjustment, and we denote by J the maximum number of periods a firm's price can be fixed.[4] The key notation describing limited price adjustment will be a vector $\underset{\sim}{\alpha}$; the j^{th} element of $\underset{\sim}{\alpha}$, called α_j, is the probability that a firm adjusts its price in period t, conditional on its previous adjustment having occurred in period $t-j$.

[4] Looking ahead, one of the specifications we will focus on has $J = \infty$.

From the vector $\underset{\sim}{\alpha}$ we derive the fractions of firms in period t charging prices set in periods $t-j$, which we denote by ω_j. To do this, note that

$$\omega_j = (1-\alpha_j)\omega_{j-1}, \text{ for } j = 1, 2, ..., J-1,$$
$$\text{and} \tag{4}$$
$$\omega_0 = 1 - \sum_{k=1}^{J-1} \omega_k.$$

This system of linear equations can be solved for ω_j as a function of $\underset{\sim}{\alpha}$. The most common pricing specifications in the literature are those first described by Taylor (1980) and Calvo (1983). Taylor's specification is one of uniformly staggered price setting: every firm sets its price for J periods, and at any point in time a fraction $1/J$ of firms charge a price set j periods ago. The $(J-1)$-element vector of adjustment probabilities for the Taylor model is $\underset{\sim}{\alpha} = [0, ..., 0]$, and the J-element vector of fractions of firms is $\underset{\sim}{\omega} = [1/J, 1/J, ..., 1/J]$. In contrast, Calvo's specification involves uncertainty about when firms can adjust their price. No matter when a firm last adjusted its price, it faces a probability α of adjusting. Thus, the infinite vector of adjustment probabilities is $\underset{\sim}{\alpha} = [\alpha, \alpha, ...]$, and the infinite vector of fractions of firms is $\omega_j = \alpha(1-\alpha)^j$, $j = 0, 1, ...$ For the specification we will advocate in Section 3, contrary to Taylor and Calvo, the adjustment fractions are strictly increasing in j and J is finite.

For any pattern of price adjustment, as defined by the α_j or ω_j, the price index (3) can be simplified to reflect the fact that all firms that set their price in the same period will choose the same price.[5] Let $P_{0,t}$ denote the price chosen by adjusting firms in period t. Then the price index can be written as

$$P_t = \left(\sum_{j=0}^{J-1} \omega_j \cdot (P_{0,t-j})^{1-\varepsilon} \right)^{\frac{1}{1-\varepsilon}}. \tag{5}$$

The next step is to show how $P_{0,t}$ is determined.

Optimal Pricing Decisions

In those periods when a firm is able to adjust its price, the price that it chooses will be affected by the pattern of future adjustment opportunities it expects, that is, by $\underset{\sim}{\alpha}$. To determine the optimal price for an adjusting firm, we must first state the firm's profit-maximization problem. If $\pi_{j,t}$ denotes the nominal profits in period t of a firm that charges a price set in period $t-j$, then the

[5] If there are firm-specific state variables other than price, then all adjusting firms will generally not choose the same price. This would be the case, for example, if firms faced costs of adjusting their labor input. Such a model would be more difficult to analyze, as the number of different types of firms one would need to track would grow without bound over time.

expected present discounted value of profits that the firm is concerned with when it adjusts its price is

$$\Pi_t = E_t \sum_{j=0}^{J-1} \Delta_{j,t+j} \left(\omega_j/\omega_0\right) \pi_{j,t+j}, \tag{6}$$

where E_t denotes expectation that is conditional on information available when the period t pricing decision is made, and $\Delta_{j,t+j}$ is the discount factor appropriate for discounting nominal profits from period $t+j$ back to period t.[6] The factor (ω_j/ω_0) is the probability that a firm that adjusts its price in period t will still be charging that price in period $t+j$.[7] Although the summation stops with period $t+J-1$, the firm of course cares about its profits further in the future than period $t+J-1$. However, its choice of a price in period t has no bearing on profits beyond period $t+J-1$, because by then a new price will be chosen.[8] From (6), the firm's optimal price sets expected discounted marginal profits to zero:

$$E_t \sum_{j=0}^{J-1} \Delta_{j,t+j} (\omega_j/\omega_0) \frac{\partial \pi_{j,t+j}}{\partial P_{0,t}} = 0. \tag{7}$$

If the firm could adjust its price every period, then ω_j would be zero for all j greater than zero; the optimal price would make marginal profits zero within every period. Price stickiness means that marginal profits are generally nonzero within a period, but the discounted sum of marginal profits is zero.

Profits in a given period are the difference between revenue and costs. For a firm in period $t+j$ that charges a price it set in period t, we will denote the demand it faces and its costs of production by $y_{j,t+j}$ and $TC_{j,t+j}$, respectively. Its profits can then be expressed as

$$\pi_{j,t+j} = P_{0,t} y_{j,t+j} - TC_{j,t+j}. \tag{8}$$

Substituting from the demand function (2) yields

$$\pi_{j,t+j} = P_{0,t} \left(\frac{P_{0,t}}{P_{t+j}}\right)^{-\varepsilon} y_{t+j} - TC_{j,t+j}. \tag{9}$$

Total revenue—the first term in (9)—is simply the product of the price the firm charges and the demand it faces. Total costs will generally depend on

[6] Below, we will assume that the discount factor is given by the product of nominal interest rates:

$$\Delta_{j,t+j} = \left((1+R_t)^{-1}(1+R_{t+1})^{-1} \cdots (1+R_{t+j-1})^{-1}\right)$$

for $j > 0$, and $\Delta_{0,t} = 1$.

[7] This factor can also be written $\prod_{k=0}^{j}(1-\alpha_k)$, where $\alpha_0 \equiv 0$.

[8] We are implicitly assuming there are no other linkages between profits in the current period and the firms' decisions in prior periods. This assumption may not be innocuous. For example, it rules out dependence of the firm's costs in period t on its production in a prior period.

factor prices, factor utilization, and the level of technology. For now we leave unspecified the determinants of costs. Below we will describe assumptions that imply marginal cost can be easily measured. Differentiating (9) yields an expression for marginal profits:

$$\frac{\partial \pi_{j,t+j}}{\partial P_{0,t}} = (1-\varepsilon)\left(\frac{P_{0,t}}{P_{t+j}}\right)^{-\varepsilon} \cdot y_{t+j} - \frac{\partial TC_{j,t+j}}{\partial P_{0,t}}. \tag{10}$$

The first term in (10) is marginal revenue with respect to price. Because there is a constant elasticity of demand greater than unity, marginal revenue *with respect to price* is always negative; *lowering* its price will always increase a firm's revenue. The second term is marginal cost with respect to price. It is convenient to express the firm's marginal cost with respect to quantity produced rather than with respect to price; therefore, use the fact that $y_{j,t+j} = \left(\frac{P_{0,t}}{P_{t+j}}\right)^{-\varepsilon} \cdot y_{t+j}$ to write (10) as

$$\frac{\partial \pi_{j,t+j}}{\partial P_{0,t}} = (1-\varepsilon)\left(\frac{P_{0,t}}{P_{t+j}}\right)^{-\varepsilon} \cdot y_{t+j}$$

$$+ \varepsilon \left(\frac{P_{0,t}}{P_{t+j}}\right)^{-1-\varepsilon} \cdot y_{t+j}\left\{\frac{\partial TC_{j,t+j}}{\partial y_{j,t+j}} \cdot \frac{1}{P_{t+j}}\right\}. \tag{11}$$

The object in brackets will be referred to as *real marginal cost;* it is the firm's marginal production cost denominated in the final good. The other factors in the second term represent the effect of a change in the price charged on the quantity of goods demanded from the firm. Since real marginal cost plays a major role in what follows, we denote that variable by the shorthand expression

$$\psi_{j,t+j} \equiv \frac{\partial TC_{j,t+j}}{\partial y_{j,t+j}} \cdot \frac{1}{P_{t+j}}. \tag{12}$$

Following up on the above discussion of total costs, real marginal cost will generally depend on variables such as the real wage. Measuring marginal cost directly is generally not a simple matter.

To derive an explicit expression for an adjusting firm's optimal price, first substitute the derivation of marginal profits (11) into the first-order condition (7):

$$E_t \sum_{j=0}^{J-1} \Delta_{j,t+j}(\omega_j/\omega_0)$$

$$\left[(1-\varepsilon)\left(\frac{P_{0,t}}{P_{t+j}}\right)^{-\varepsilon} y_{t+j} + \varepsilon \left(\frac{P_{0,t}}{P_{t+j}}\right)^{-1-\varepsilon} y_{t+j}\psi_{j,t+j}\right] = 0. \tag{13}$$

Next, multiply (13) by $P_{0,t}^{1+\varepsilon} P_t^{-\varepsilon}$ and rearrange to get

$$P_{0,t} = P_t \left(\frac{\varepsilon}{\varepsilon - 1} \right) \frac{E_t \sum_{j=0}^{J-1} \Delta_{j,t+j} (\omega_j / \omega_0) \left(\frac{P_t}{P_{t+j}} \right)^{-1-\varepsilon} y_{t+j} \psi_{j,t+j}}{E_t \sum_{j=0}^{J-1} \Delta_{j,t+j} (\omega_j / \omega_0) \left(\frac{P_t}{P_{t+j}} \right)^{-\varepsilon} y_{t+j}}. \tag{14}$$

If the price level and marginal cost are constant, then (14) yields the constant markup that is familiar from static monopolistic competition models: $P_{0,t} = P_t \left(\frac{\varepsilon}{\varepsilon - 1} \right) \psi$. This is the markup (or relative price) that maximizes one-period profits. If the price level or marginal cost are not constant, then neither is the relative price that maximizes one-period profits. Therefore a firm whose nominal price may be fixed for more than one period chooses a nominal price that it expects will sacrifice the fewest discounted profits over the life of the price.

Inflation

If aggregate demand (y_t), real marginal cost (ψ_t), and nominal interest rates (equivalently the discount factors $\Delta_{j,t+j}$) are taken as given, then the pair of equations (5) and (14) jointly describe the behavior of the aggregate price level and the price chosen by individual firms. Thus, if we knew the processes governing aggregate demand, real marginal cost, and nominal interest rates, then we could use (5) and (14) to determine the behavior of the price level and hence inflation. In general it is tedious to obtain an explicit expression for inflation. However, it is easy to compute the behavior of inflation. A simple pricing specification will suffice to illustrate the method by which one can compute the behavior of inflation. In analyzing this special case, we will linearize the equations for the price index and for optimal pricing around a steady state with constant inflation rate μ. Linear approximations are also used in the empirical work by Sbordone (1998) and Galí and Gertler (1999).

The special case is a model where no firm sets its price for more than two periods, so that $\underset{\sim}{\alpha} = [\alpha_1]$ and $\underset{\sim}{\omega} = [1/(2-\alpha_1), (1-\alpha_1)/(2-\alpha_1)]$. In this case the price index is

$$P_t = \left(\frac{1}{2-\alpha_1} \cdot P_{0,t}^{1-\varepsilon} + \frac{1-\alpha_1}{2-\alpha_1} \cdot P_{0,t-1}^{1-\varepsilon} \right)^{1/(1-\varepsilon)},$$

and the optimal pricing equation is

$$P_{0,t} = P_t \left(\frac{\varepsilon}{\varepsilon - 1} \right) \frac{y_t \psi_{0,t} + (1-\alpha_1) E_t [\Delta_{1,t+1} (P_t / P_{t+1})^{-1-\varepsilon} y_{t+1} \psi_{1,t+1}]}{y_t + (1-\alpha_1) E_t [\Delta_{1,t+1} (P_t / P_{t+1})^{-\varepsilon} y_{t+1}]}.$$

By linearizing these equations around a steady state with gross inflation equal to μ, we will get a system of expectational difference equations. Before linearizing, rewrite the equations in terms of detrended nominal variables:

$$\tilde{P}_t^{1-\varepsilon} = \left(\left(\frac{1}{2-\alpha_1}\right)\cdot \tilde{P}_{0,t}^{1-\varepsilon} + \left(\frac{1-\alpha_1}{2-\alpha_1}\right)\cdot \mu^{\varepsilon-1}\tilde{P}_{0,t-1}^{1-\varepsilon}\right), \tag{15}$$

and

$$\tilde{P}_{0,t} = \tilde{P}_t\left(\frac{\varepsilon}{\varepsilon-1}\right)$$

$$\cdot \frac{y_t\psi_{0,t} + (1-\alpha_1)E_t[\Delta_{1,t+1}\mu^{1+\varepsilon}(\tilde{P}_t/\tilde{P}_{t+1})^{-1-\varepsilon}y_{t+1}\psi_{1,t+1}]}{y_t + (1-\alpha_1)E_t[\Delta_{1,t+1}\mu^{\varepsilon}(\tilde{P}_t/\tilde{P}_{t+1})^{-\varepsilon}y_{t+1}]}. \tag{16}$$

In (15), (16), and henceforth, the variables $\tilde{P}_{0,t}$ and $\tilde{P}_t$ should be interpreted as deviations from a trend that is growing at rate μ; that is, $\tilde{P}_{0,t} = P_{0,t}/\mu^t$ and $\tilde{P}_t = P_t/\mu^t$.

Linearizing the price index (15) yields

$$\left(\frac{P_0}{P}\right)^{\varepsilon-1}\hat{P}_t = \left(\frac{1}{2-\alpha_1}\right)\cdot \hat{P}_{0,t} + \left(\frac{1-\alpha_1}{2-\alpha_1}\right)\cdot \mu^{\varepsilon-1}\hat{P}_{0,t-1}. \tag{17}$$

Here (P_0/P) denotes the ratio of the price set by an adjusting firm to the aggregate price level in a steady state where the price level is growing at rate μ, and $\hat{P}_t$ and $\hat{P}_{0,t}$ are logarithmic deviations from the steady-state values of $\tilde{P}_t$ and $\tilde{P}_{0,t}$, respectively. The steady-state ratio (P_0/P) can easily be determined from (15) as $(P_0/P) = \left(\frac{1+(1-\alpha_1)\mu^{\varepsilon-1}}{2-\alpha_1}\right)^{1/(\varepsilon-1)}$. According to (17), if inflation is high enough or the probability of adjustment is low enough, then a given change in prices set in the previous period has a larger effect on this period's price index than does the same change in prices set this period. The reason is that, with inflation eroding relative prices, the relative prices of goods set in the previous period is low, and hence the quantity of those goods purchased is high. Furthermore, with relatively elastic demand, the share of expenditure on the low-priced goods will be higher than the share of expenditure on the high-priced goods, meaning that goods with a price set in the previous period carry greater weight in the steady-state price index.[9]

Linearizing the equation for optimal price of an adjusting firm (16) yields

$$\left(\frac{P_0}{P}\right)\hat{P}_{0,t} = a_1\hat{P}_t + (a_2-a_3)(E_t\hat{P}_{t+1} - \hat{P}_t) + b_0\hat{\psi}_{0,t} + b_1E_t\hat{\psi}_{1,t+1} + x_t, \tag{18}$$

where

$$a_1 \equiv \left(\frac{\varepsilon}{\varepsilon-1}\right)\left(\frac{\psi_0 + (1-\alpha_1)\Delta_1\mu^{1+\varepsilon}\psi_1}{1+(1-\alpha_1)\Delta_1\mu^{\varepsilon}}\right),$$

[9] If there is a high adjustment probability, then this expenditure share is low, and the opposite result holds.

$$a_2 \equiv \left(\frac{\varepsilon}{\varepsilon-1}\right)\left(\frac{(1-\alpha_1)\Delta_1\mu^{1+\varepsilon}\psi_1}{1+(1-\alpha_1)\Delta_1\mu^{\varepsilon}}\right)(1+\varepsilon),$$

$$a_3 \equiv \left(\frac{\varepsilon}{\varepsilon-1}\right)\left(\frac{\psi_0+(1-\alpha_1)\Delta_1\mu^{1+\varepsilon}\psi_1}{1+(1-\alpha_1)\Delta_1\mu^{\varepsilon}}\right)\left(\frac{(1-\alpha_1)\Delta_1\mu^{\varepsilon}}{1+(1-\alpha_1)\Delta_1\mu^{\varepsilon}}\right)\varepsilon,$$

$$b_0 \equiv \left(\frac{\varepsilon}{\varepsilon-1}\right)\left(\frac{\psi_0}{1+(1-\alpha_1)\Delta_1\mu^{\varepsilon}}\right),$$

$$b_1 \equiv \left(\frac{\varepsilon}{\varepsilon-1}\right)\left(\frac{(1-\alpha_1)\Delta_1\mu^{1+\varepsilon}\psi_1}{1+(1-\alpha_1)\Delta_1\mu^{\varepsilon}}\right),$$

and

$$x_t \equiv \left(\frac{\varepsilon}{\varepsilon-1}\right)\left(\frac{1}{1+(1-\alpha_1)\Delta_1\mu^{\varepsilon}}\right)\{\psi_0\hat{y}_t+(1-\alpha_1)\Delta_1\mu^{1+\varepsilon}\psi_1$$
$$\cdot E_t(\hat{\Delta}_{1,t+1}+\hat{y}_{t+1})-\left(\frac{\psi_0+(1-\alpha_1)\Delta_1\mu^{1+\varepsilon}\psi_1}{1+(1-\alpha_1)\Delta_1\mu^{\varepsilon}}\right)[\hat{y}_t+(1-\alpha_1)\Delta_1\mu^{\varepsilon}$$
$$\cdot E_t(\hat{\Delta}_{1,t+1}+\hat{y}_{t+1})]\}.$$

If α_1 is low enough and μ is high enough, then $(a_2 - a_3)$ will be positive, in which case (18) says that the price set by an adjusting firm is increasing in the price level and increasing in next period's expected inflation. A firm raises its price as the price level rises because it has an optimal level for its *relative* price (note that [16] can be written with $P_{0,t}/P_t$ on the left-hand side). Expected inflation next period raises a firm's desired price because it means that any price set in the current period will erode in relative terms; firms compensate for the erosion by setting a higher price when they can adjust. The coefficients b_0 and b_1 are positive, which means that the price chosen by adjusting firms responds positively to marginal cost in the current period and to expected future marginal cost. Finally, the variable x_t represents the effects on a firm's optimal price of current and future aggregate demand and the nominal discount factor. We have lumped these factors into the variable x_t in order to focus attention on the fact that (17) and (18) jointly determine the behavior of the price level and adjusting firms' optimal prices, *conditional* on the behavior of real marginal cost and the variables in x_t.

Pursuing now the joint determination of $\hat{P}_{0,t}$ and $\hat{P}_t$, we write (17) and (18) as a system of linear expectational difference equations in the variables $\hat{P}_t$ and $\hat{P}_{0,t-1}$:

$$\begin{bmatrix} 0 & 1/(2-\alpha_1) \\ a_3 - a_2 & P_0/P \end{bmatrix} \begin{bmatrix} E_t \hat{P}_{t+1} \\ \hat{P}_{0,t} \end{bmatrix} =$$

$$\begin{bmatrix} (P_0/P)^{\varepsilon-1} & -\left(\frac{1-\alpha_1}{2-\alpha_1}\right)\mu^{\varepsilon-1} \\ a_1 - a_2 + a_3 & 0 \end{bmatrix} \begin{bmatrix} \hat{P}_t \\ \hat{P}_{0,t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ x_t + b_0 \hat{\psi}_{0,t} + b_1 E_t \hat{\psi}_{1,t+1} \end{bmatrix}. \quad (19)$$

The methods described by Blanchard and Kahn (1980) allow one to solve for the behavior of $\hat{P}_t$ given a known process for $\hat{x}_t$ and $\hat{\psi}_{i,t}$.[10] For general specifications of price stickiness—that is, $\underset{\sim}{\alpha}$—the system corresponding to (19) is more complicated. There are additional expected future values of inflation, demand, marginal cost, and discount factors in the analogue to (18), and there are additional past values of optimal prices in the analogue to (17). However, the method for deriving and then solving the system of difference equations is almost identical. For any specification of $\underset{\sim}{\alpha}$ then, the solution to the analogue to (19) describes how the behavior of real marginal cost and x_t translates into the behavior of inflation. This relationship is the basis for the empirical work to be discussed next.

2. TAKING THE MODEL TO DATA

Our microeconomic-based sticky-price model determines the behavior of the aggregate price level, and hence the inflation rate, in partial equilibrium.[11] From an empirical perspective, this relationship is important because it enables researchers to work with aggregate variables like inflation rather than individual variables like the prices of particular goods. Sbordone (1998) and Galí and Gertler (1999) apply this result in a new and interesting way: they estimate the parameters $\underset{\sim}{\alpha}$, and then test whether the estimated model successfully accounts for actual inflation.

Suppose that all of the models' parameters were known and that data on real marginal cost for different types of firms ($\psi_{j,t}$), aggregate output, and nominal interest rates (the components of x_t), were all available. Then we could use (19) to simulate the behavior of inflation. To simulate, solve the model so that the price level is expressed as a function of the exogenous variables (x_t and

[10] Note that an initial condition for $P_{0,t-1}$ is also needed; we will assume the steady-state initial condition. The impulse response functions to be presented below are produced using algorithms developed by King and Watson (1998); these algorithms make it easy to solve the more complicated systems that result from more complicated forms of price stickiness.

[11] Recall from above that in a general equilibrium model there would also be a description of monetary policy. While monetary policy would determine the behavior of inflation, inflation behavior would still have to be consistent with individual firms' pricing. And firms would still take the behavior of inflation as given.

$\hat{\psi}_{j,t}$). Then use the observed sequences of exogenous variables to build up a simulated price-level series, from which it is easy to create a simulated inflation series. Of course the parameters are not known, but they can be estimated so that the simulated behavior of inflation is closest to what we observe. Roughly speaking, this is what Sbordone and Galí and Gertler do.[12]

Galí and Gertler make two key assumptions. The first assumption is that price stickiness is given by the Calvo specification, so that only one parameter is related to price stickiness (recall that the Calvo specification is $\alpha_j = \alpha$ for $j = 1, 2, ...\infty$). The second assumption is that all firms produce using identical Cobb-Douglas technologies and the labor market is competitive over the whole economy.[13] We will take up the pricing specification later. Here we explain the importance of the second assumption.

In order for the empirical approach described above to be feasible, the researcher must have access to data on marginal cost. But unlike GDP or inflation, marginal cost is not a data series measured by a government statistical agency. Measurement is lacking for a good reason: the appropriate measure of marginal cost depends on characteristics of the economy which are only imperfectly understood. These characteristics include the competitiveness of factor markets and the extent of adjustment costs firms face in hiring new workers and installing new capital.[14] The assumptions described above surmount this problem, as they imply that the appropriate measure of real marginal cost is labor's share of output—unit labor costs. Estimates of these series are widely available, and hence estimation of α is feasible.

To see how labor's share can reflect real marginal cost, let Y_t be output, let L_t be labor, and let K_t be capital; then with Cobb-Douglas technology, $Y_t = A_t K_t^{\alpha} L_t^{1-\alpha}$. Nominal marginal cost $\left(\partial TC/\partial Y\right)$ can be decomposed as $\partial TC/\partial Y = \left(\partial TC/\partial L\right) \div \left(\partial Y/\partial L\right)$. In a competitive labor market, $\partial TC/\partial L$ is simply the nominal wage (W_t), and $\partial Y/\partial L$ is of course the marginal product of labor—$(1-\alpha) A_t (K_t/L_t)^{\alpha}$ for the Cobb-Douglas case. Therefore, real marginal cost is $\psi_t = (W_t/P_t) \div \left[(1-\alpha) A_t (K_t/L_t)^{\alpha}\right]$. If we let w_t denote the real wage ($w_t = W_t/P_t$), then real marginal cost can be expressed as $\psi_t = w_t L_t \div \left[(1-\alpha) Y_t\right]$. Real marginal cost, then, is proportional to labor's share of output, and variations in labor's share provide a measure of variations in real marginal cost.

[12] These authors each use different estimation methods. However, it is fair to summarize both of those methods as ones that choose the model's parameters in order to best fit observed inflation.

[13] These authors also linearize around a zero inflation steady state, which simplifies things further. Sbordone allows marginal cost to vary according to when firms last adjusted their price, whereas Galí and Gertler do not. As such, Sbordone's analysis is more general than what we describe.

[14] Similar problems are involved in the measurement of output. Arguably, however, the problems are more severe for marginal cost.

Galí and Gertler (1999) find that, with labor's share as a proxy for real marginal cost, a Calvo pricing model explains post–1960 U.S. inflation quite well. Their estimate of α is roughly 0.2, implying that firms keep their prices fixed for about five quarters on average. This result is striking. It runs counter to the claims of Fuhrer and Moore (1995) and others that forward-looking sticky-price models are inconsistent with the behavior of U.S. inflation. Galí and Gertler reconcile these results by emphasizing that previous work explained inflation through the behavior of output. As is clear from (19), however, the key variable for explaining inflation is real marginal cost rather than output.[15] Thus Galí and Gertler argue that the main empirical difficulty is not in explaining inflation behavior with a forward-looking sticky-price model, but in reconciling the behavior of output with the behavior of real marginal cost.

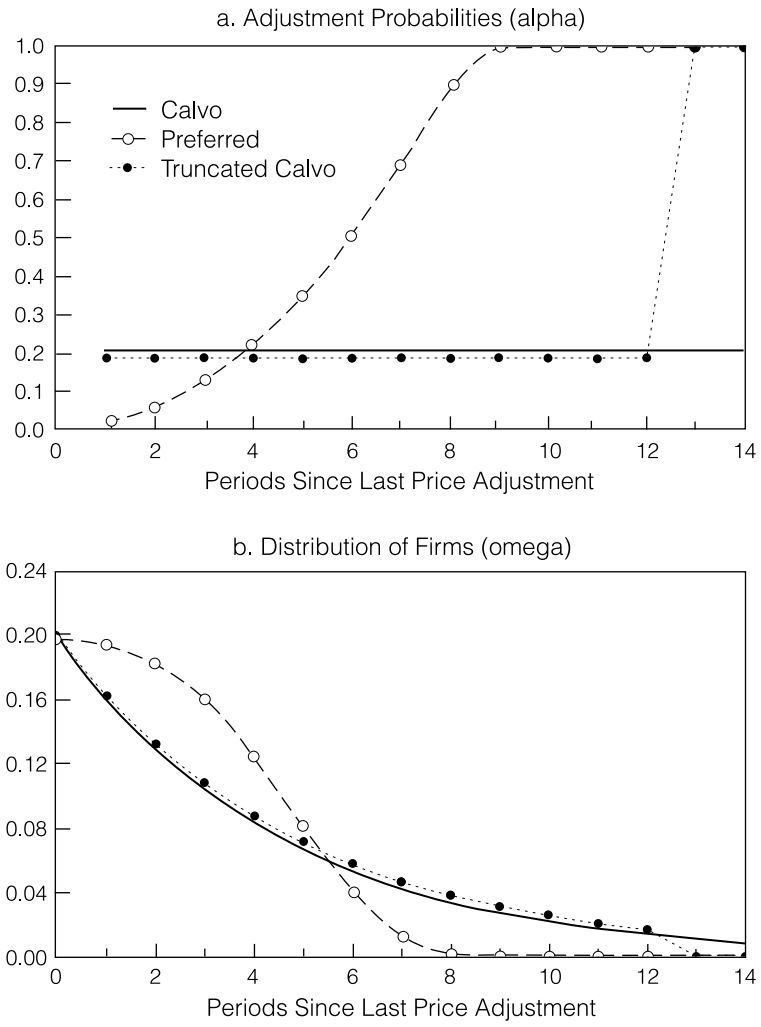
3. INFLATION DYNAMICS ARE SENSITIVE TO THE PRICING STRUCTURE

One might think that because the Calvo model fits inflation data well, it must be an appropriate model. However, another aspect of the data is fundamentally at odds with the Calvo model. Unfortunately, it appears difficult to eliminate this discrepancy without changing the implications for inflation dynamics.

Recall that the Calvo specification posits a common price-adjustment probability (α) for all firms. From (4), we see that the distribution of fractions of firms is then given by $\omega_j = \alpha(1-\alpha)^j$, for $j = 1, 2, ...$ That is, a positive fraction of firms charges a price set arbitrarily many periods in the past. This is clearly a counterfactual implication. However, the Calvo specification allows for a characterization of inflation dynamics even simpler than (16), and it may be worth paying the price of an infinite distribution in order to gain this simplification. Supporting this view is the fact that the fractions of firms become arbitrarily small as the number of periods increases; for example, if $\alpha = 0.2$, less than 0.02 percent of firms charge a price set more than ten years in the past. With numbers that small, it is difficult to believe that the Calvo specification could produce dynamics qualitatively different than those associated with a more plausible specification generating the same average duration of a fixed price. Recent work by Kiley (1998), however, suggests that two such models would produce qualitatively different dynamics. To investigate the implications for inflation of different specifications of price stickiness, we first estimate a univariate autoregression for labor's share (used here to represent real marginal cost). We then compare the sticky-price model's impulse response functions of inflation to a shock to labor's share for the different specifications.

[15] When average inflation ($\mu - 1$) is nearly zero, the coefficients on demand and nominal interest rates in (19) will be small.

Figure 1 Three Pricing Specifications



The three pricing specifications we analyze are illustrated in Figure 1. Panel a shows the patterns of adjustment probabilities (α^s), and panel b shows the distributions of fractions of firms (ω^s). The solid lines represent a Calvo specification close to that estimated by Galí and Gertler.[16] The dashed lines are

[16] Galí and Gertler's specification is slightly different, as they allow for a fraction of firms to be "rule-of-thumb" price setters. However, these firms turn out to be unimportant for their results.

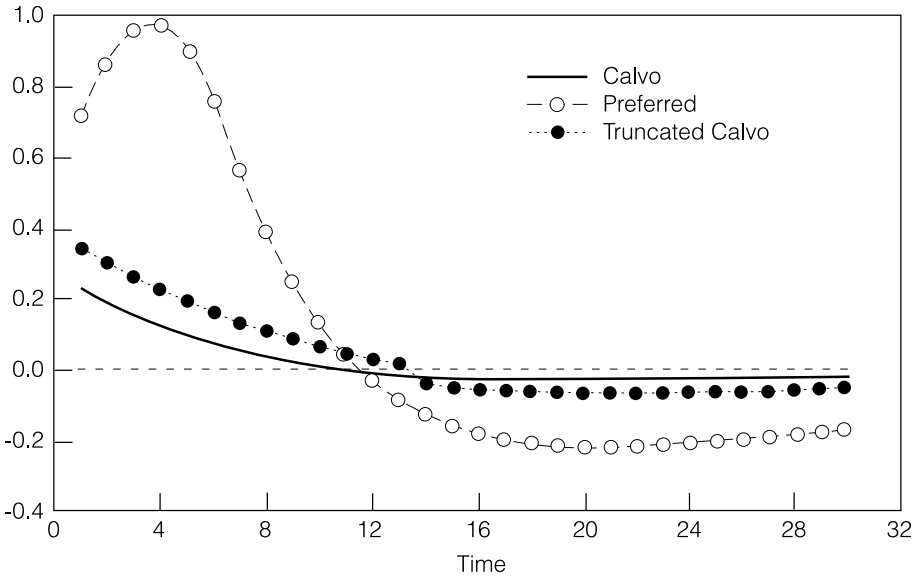
arguably a more reasonable specification: no firms charge a price set more than eight quarters ago, but the average duration of a fixed price is five quarters, just as for the solid-line Calvo specification. The dashed line, which we will refer to as our preferred case, has further appeal in that it is the kind of price-adjustment pattern generated if firms face a distribution of fixed costs of price adjustment, as in Dotsey, King, and Wolman (1999). Finally, the dotted line is an intermediate case: as in the Calvo case, firms face a constant adjustment probability for the first 12 quarters that they are charging a price, but the adjustment probabilities jump to one after the twelfth quarter.

Figure 2 shows the response of inflation to a marginal cost shock (as proxied for by labor's share) under these three specifications of price stickiness. In studying these pictures, it is important to keep in mind that the Calvo specification (solid line) has been shown to be consistent with the behavior of inflation in the United States when labor's share is used to represent marginal cost. For that case, the response of inflation to a marginal cost shock is relatively small, but fairly persistent. In contrast, for the preferred specification, where the adjustment probabilities are smoothly increasing and the distribution of firms does not extend beyond eight quarters, inflation responds much more strongly to the marginal cost shock; the magnitudes of the increase and subsequent decrease in inflation are roughly three times as large as the corresponding magnitudes for the Calvo case. Although the intermediate case gives results closer to Calvo, still the impact effect of the marginal cost shock on inflation is nearly 50 percent greater in the intermediate case than it is for pure Calvo pricing.

Figure 1 can help us to understand the dramatic difference between inflation behavior under the Calvo and preferred specifications. Even though in both cases roughly the same fraction of firms adjusts their price in a given period, for the Calvo case a higher fraction of the adjusting firms are themselves recent adjusters (in panel a, α_j is relatively high for low j in the Calvo case). Since recent adjusters have already responded to a recent shock, their effect on the price level is small. By contrast, in the preferred case most of the adjusting firms have last adjusted several periods ago. Thus, in periods immediately following a shock, the model registers significant additional adjustment to that shock. For an adjusting firm, this means that it responds more strongly to a shock in the preferred case; to not do so would mean that its relative price would move too far from the desired level in ensuing periods.

That the impulse response functions differ sharply for the Calvo and our preferred case has a direct bearing on whether Galí and Gertler's empirical results are sensitive to the assumption of Calvo pricing. Recall they found that the dynamics of U.S. inflation could be closely replicated by a model of Calvo pricing, where the main "forcing variable" for inflation was labor's share, which proxies for real marginal cost. The impulse response function of inflation to marginal cost is one way of summarizing the model's dynamics. Because the

Figure 2 Impulse Response Functions of Inflation to a Marginal Cost Shock



Calvo model matches inflation dynamics, its impulse response function is the "correct one" for matching the behavior of inflation. The preferred case gives such a different impulse response function that it could not also match inflation behavior when driven by the same marginal cost process; inflation would have to be much more volatile than observed in the data, or real marginal cost would have to be much smoother. We conclude, then, in support of Kiley's (1998) finding that the Calvo model is an extreme special case, not just a convenient simplification. Modifying the form of price stickiness so that (1) the probability of price adjustment is a smoothly increasing function of time since last adjustment, and (2) no firm keeps its price fixed more than eight quarters leads to dynamics fundamentally different than those of the Calvo model, even if one holds constant the average length of time a price is held fixed. With such a change, it will no longer be possible to match inflation dynamics with labor's share proxying for real marginal cost.

4. SHOULD WE GIVE UP ON STICKY-PRICE MODELS?

One interpretation of our critique is that models with imperfect competition and sticky prices are poor descriptions of the data and as such should be abandoned.

We prefer a constructive interpretation, which focuses on the assumptions that guarantee labor's share would be a good approximation to real marginal cost. The first interpretation assumes that labor's share does represent real marginal cost. In this case, if Calvo pricing is the only form of price stickiness consistent with inflation dynamics, but is unacceptable for reasons discussed above, then we should give up on this entire class of sticky-price models. On the other hand, if labor's share does not represent real marginal cost, then a more plausible pricing specification might be consistent with data on inflation and marginal cost, correctly measured.

To justify using labor's share as a stand-in for real marginal cost, we assume that all firms produce using identical Cobb-Douglas technologies and that there is an economywide competitive labor market. These assumptions clearly represent an oversimplification. Possibly by constructing a richer marginal cost structure, one could reconcile a plausible sticky-price model with the behavior of inflation. Sbordone (1998) has already analyzed a simple generalization for marginal cost. She assumes the presence of a competitive labor market but allows factor ratios and hence marginal cost to vary depending on when a firm adjusted its price. Sbordone also maintains the assumption of Calvo-style pricing, so it is unclear whether that particular generalization of the marginal cost structure can generate realistic inflation dynamics when combined with a realistic pricing specification.

Once one is willing to relax the assumptions about factor markets and technology, a wide range of behavior for marginal cost is possible; typically, real marginal cost will not simply correspond to labor's share. Rotemberg and Woodford (1999a) work through several formulations: non-Cobb-Douglas technology, overhead labor, overtime pay, labor adjustment costs, labor hoarding, and variable capital utilization. Incorporating these features means that to explain marginal cost one would need not only labor's share but also such variables as output, labor input, the marginal wage, current and expected future growth of labor input, the fraction of labor input which is idle, and hours per worker. Rotemberg and Woodford cite several papers that have pursued these ideas in an attempt to learn about real marginal cost. Our interpretation of Sbordone's and Galí and Gertler's work suggests that a next step would be to study whether more refined estimates of marginal cost can help reconcile a plausible sticky-price specification with the behavior of inflation.

Another worthwhile endeavor would be to use direct evidence on price stickiness to choose α, and then use the relationship between marginal cost and inflation to estimate the behavior of marginal cost. In other words, instead of using independent evidence on marginal cost to estimate the form of price stickiness, one would be using independent evidence on pricing to estimate the behavior of real marginal cost.

5. CONCLUSIONS

Current optimizing sticky-price models imply a tight relationship between real marginal cost and inflation. We have worked through the steps in this relationship in detail: expressions for the price index (5) and for a price-setting firm's optimal price (14) imply a linear system (19) that approximates the behavior of the price level (and inflation) primarily as a function of real marginal cost.

In interesting recent empirical work, Sbordone (1998) and Galí and Gertler (1999) use the relationship between marginal cost and inflation to estimate sticky-price models and evaluate how well these models explain actual inflation. Their results are positive in that they find sticky-price models are able to explain U.S. inflation quite well. However, this empirical work relies on the Calvo pricing specification, where firms face a positive probability of having their price fixed for an arbitrarily long time. This pricing specification is clearly implausible. We have shown—building on work by Kiley (1998)—that if labor's share is used to proxy for real marginal cost, a more plausible pricing specification generates inflation dynamics inconsistent with the data.

Continued progress in empirical evaluation of sticky-price models will require intensive study of the factors determining real marginal cost. With more refined estimates of real marginal cost, it may be possible to reconcile a plausible sticky-price specification with data on inflation. Conversely, to the extent that we are confident a particular pricing specification is correct, the link between real marginal cost and inflation should allow us to come up with independent estimates of real marginal cost. Such estimates will help us learn about other aspects of the economy's structure, such as the form of technology, the competitiveness of factor markets, and the extent of adjustment costs in hiring labor and installing capital. Ultimately, this knowledge will facilitate constructing more accurate general equilibrium models, which can then be used for the kind of policy analysis mentioned at the outset.

APPENDIX

Derivation of Demand Function and Price Index

To derive the demand function (2), solve the following problem: minimize the cost of purchasing a given level of final output $\bar{y}$ by choosing appropriate levels of $y(z)$, $z \in (0,1)$:

$$\mathcal{L} = \left[\int_0^1 P(z) \cdot y(z) dz\right] + P\left[\bar{y} - \left(\int_0^1 y(z)^{\frac{\varepsilon-1}{\varepsilon}} dz\right)^{\frac{\varepsilon}{\varepsilon-1}}\right],$$

where $P(z)$ denotes the nominal price of good z. The Lagrange multiplier on the quantity constraint is the price level P, because the multiplier has the interpretation of the marginal cost of an additional unit of final output, and that is precisely the price index. The first-order conditions for this problem are

$$P(\tilde{z}) = P\left(\int_0^1 y(z)^{\frac{\varepsilon-1}{\varepsilon}} dz\right)^{\frac{1}{\varepsilon-1}} y(\tilde{z})^{\frac{-1}{\varepsilon}}$$

$$\tilde{z} \in (0,1).$$

Using the definition of y in (1), these conditions simplify to

$$y(\tilde{z}) = \left(\frac{P(\tilde{z})}{P}\right)^{-\varepsilon} y \tag{20}$$

$$\tilde{z} \in (0,1); \tag{21}$$

demand for a firm's product is increasing in the level of aggregate demand (y) and decreasing in the relative price the firm charges.

Now we show how the price index is calculated as a function of the prices $P(z)$. Substitute (20) into (1):

$$y = \left\{\int_0^1 \left[\left(\frac{P(\tilde{z})}{P}\right)^{-\varepsilon} y\right]^{(\varepsilon-1)/\varepsilon} dz\right\}^{\frac{\varepsilon}{\varepsilon-1}},$$

which implies

$$1 = \left\{\int_0^1 \left(\frac{P(\tilde{z})}{P}\right)^{(1-\varepsilon)}\right\}^{\frac{\varepsilon}{\varepsilon-1}}$$

and thus

$$P = \left\{\int_0^1 P(\tilde{z})^{1-\varepsilon} d\tilde{z}\right\}^{1/(1-\varepsilon)}. \tag{22}$$

REFERENCES

Benhabib, Jess, Stephanie Schmitt-Grohe, and Martin Uribe. "The Perils of Taylor Rules," *Journal of Economic Theory*, forthcoming.

Blanchard, Olivier J., and Charles M. Kahn. "The Solution of Linear Difference Models under Rational Expectations," *Econometrica*, vol. 48 (July 1980), pp. 1305–11.

Calvo, Guillermo A. "Staggered Prices in a Utility-Maximizing Framework," *Journal of Monetary Economics*, vol. 12 (September 1983), pp. 383–98.

Dixit, Avinask K., and Joseph E. Stiglitz. "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, vol. 67 (June 1977), pp. 297–308.

Dotsey, Michael, Robert G. King, and Alexander L. Wolman. "State Dependent Pricing and the General Equilibrium Dynamics of Money and Output," *Quarterly Journal of Economics*, vol. 114 (May 1999), pp. 655–90.

Eckstein, Otto, and David Wyss. "Industry Price Equations," in Otto Eckstein, ed., *The Econometrics of Price Determination Conference.* Washington, Federal Reserve Board, 1972, pp. 133–65.

Fuhrer, Jeff, and George R. Moore. "Inflation Persistence," *Quarterly Journal of Economics*, vol. 110 (February 1995), pp. 127–59.

Galí, Jordi, and Mark Gertler. "Inflation Dynamics: A Structural Econometric Analysis," *Journal of Monetary Economics*, vol. 44 (October 1999), pp. 195-222.

Ireland, Peter N. "Optimal Disinflationary Paths," *Journal of Economic Dynamics and Control*, vol. 19 (November 1995), pp. 1429–48.

Kiley, Michael T. "Partial Adjustment and Staggered Price Setting." Manuscript. Federal Reserve Board, 1998.

King, Robert G., and Mark W. Watson. "System Reduction and Solution Algorithms for Singular Linear Difference Systems under Rational Expectations." Manuscript. The University of Virginia, 1998.

McCallum, Bennett T., and Edward Nelson. "Performance of Operational Policy Rules in an Estimated Semiclassical Structural Model," in John B. Taylor, ed., *Monetary Policy Rules.* Chicago: University of Chicago Press, 1999, pp.15–45.

Rotemberg, Julio J., and Michael Woodford. "The Cyclical Behavior of Prices and Costs," in John B. Taylor and Michael Woodford, eds., *Handbook of Macroeconomics*. Amsterdam: North-Holland, 1999a.

_____. "Interest Rate Rules in an Estimated Sticky Price Model," in John B. Taylor, ed., *Monetary Policy Rules.* Chicago: University of Chicago Press, 1999b, pp. 57–119.

Sbordone, Argia M. "Prices and Unit Labor Costs: A New Test of Price Stickiness." Manuscript. Rutgers University, 1998.

Taylor, John B. "Staggered Price and Wage Setting in Macroeconomics," in John B. Taylor and Michael Woodford, eds., *Handbook of Macroeconomics*. Amsterdam: North-Holland, 1999.

_____. "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, vol. 88 (February 1980), pp. 1–24.

Wolman, Alexander L. "Staggered Price Setting and the Zero Bound on Nominal Interest Rates," Federal Reserve Bank of Richmond *Economic Quarterly*, vol. 84 (Fall 1998), pp. 1–24.

Means of Payment, the Unbanked, and EFT '99

Edward S. Prescott and Daniel D. Tatar

The Debt Collection Improvement Act of 1996 mandated that all federal payments except tax refunds were to be made by electronic transfer by January 2, 1999. Such payments consist mainly of government benefits such as Social Security or Supplemental Security Income but also include other payments, such as those to vendors.[1] The goal of the mandate was to save the government money by having payments switched from paper checks to less expensive electronic transfers.

The government's move toward electronic means of payment comes at a seemingly opportune time. Recent developments in telecommunication and computer technologies have greatly reduced the cost of electronic communication. A growing number of consumers regularly make purchases and pay bills electronically. Despite this trend, however, there is an important impediment to the government's move: Nearly 15 percent of U.S. households, most of which are low-income, do not own checking accounts.[2]

The authors would like to thank John Caskey, Sheila Crowley, Jeanne Hogarth, Elaine Mandaleris, Ellen Stevens, Michael Stegman, David Stoesz, and the referees Ray Owens, John Walter, and Roy Webb for helpful comments. The focus groups discussed in the appendix were organized and run jointly by the first author, Sheila Crowley, Ellen Stevens, and David Stoesz. The idea of using focus groups to study how low-income households use financial services came from David Saunders, who unfortunately passed away before we began the interviews. The views expressed in this article are those of the authors and not necessarily the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Requirements to make payments by electronic transfer are also found in the Welfare Reform Act of 1996. This Act required that welfare benefits, the costs of which are shared between the states and the federal government, be paid electronically by the year 2002.

[2] We will use the term *low-income* to refer to people who are generally less financially secure. Though this label is too broad for the population we study—for example, even students and wealthy people can have low incomes—we follow this convention because the label is commonly used in this manner.

This obstacle delayed the implementation of EFT '99, the electronic funds transfer portion of the Act. In particular, the Department of the Treasury discarded early plans that required *all* government beneficiaries to receive their payments electronically because the requirement would have imposed a hardship on those without accounts. Instead, the Treasury instituted a strategy of encouraging government beneficiaries to receive payments voluntarily by electronic means. At the center of this strategy was the creation of an inexpensive type of bank account called the Electronic Transfer Account (ETA) through which beneficiaries could receive their payments electronically.

We have two specific objectives in this article. The first is to understand why low-income households choose certain means of payment. In particular, we want to understand why so many people in this group do not own checking accounts. The second objective is to use these findings to assess EFT '99. Understanding why many people do not own checking accounts will provide insight into whether ETAs are likely to be adopted.

Throughout our article we refer both to quantitative and qualitative sources of information. In particular, we report on the results of two focus group interviews and use this information to elaborate on the quantitative findings. We believe that field research is an important method for gathering information about low-income households' need for and use of financial services. Our hope is that, by example, this article illustrates the value of these research methods. (See the Appendix for detailed information about our field research and the two focus groups.)

We start by reporting information on the "unbanked," that is, people without bank accounts.[3] We describe who they are and study their tradeoffs between owning and not owning a checking account. We find that many of the unbanked have inexpensive alternatives to account ownership for their payment services. The majority of the unbanked are cashing their checks for free, ironically, at banks and other institutions such as grocery stores. Few of the unbanked use the much-maligned check-cashing outlets as a regular source for cashing checks. We also argue that, for low-income individuals, owning a checking account can be more expensive than is commonly believed. In particular, we speculate that checking accounts are expensive in part because of the implicit credit extension they contain. Moreover, we find that a small fraction of people forgo bank accounts because their creditors can seize their bank account balances to satisfy debts. For these reasons, we conclude that forgoing the use of a checking account is a rational decision for many of the unbanked. Furthermore, our analysis suggests that ETAs will not be widely adopted by the unbanked.

[3] Throughout this article, *bank accounts* will refer to accounts held at banks, thrifts, and credit unions.

1. BACKGROUND INFORMATION ON THE UNBANKED

There is a surprisingly large percentage of the population that does not own an account at a depository institution. According to the Federal Reserve's triennial Survey of Consumer Finances in 1995, 13 percent of households (roughly 13 million of them) had no bank accounts of any kind and 15 percent did not own checking accounts.[4] These numbers have fluctuated somewhat over time. In the 1977 survey, 9 percent of households did not own bank accounts while in the 1989 survey, the number rose to 15 percent (Kennickell, Starr-McCluer, and Sunden 1997).[5]

Likewise, many recipients of government benefits are unbanked. According to Hawke (1997), the number is at least 10 million. Many of these beneficiaries receive their benefits from the Social Security Administration (SSA) and the Supplemental Security Income (SSI) programs. We do not know the breakdown of the unbanked by each of these programs, but we do know the number of payments each agency makes by check; presumably these two numbers are positively related. For example, the SSA program distributes benefits to 44 million people: Over the six-month period from October 1, 1998, to March 31, 1999, it made 270 million payments, 25 percent of which were by check. The SSI program distributes benefits to 6 million people and, over the same period, it made 40 million payments, 54 percent of which were by check. It is worth noting that the SSI program mainly distributes benefits to low-income people, which, as we will see, is the demographic group most likely not to own an account.

Sources of Information

In general, little information has previously been published about the unbanked. The Fed's Survey of Consumer Finances, which collects detailed information on financial asset holdings of U.S. families every three years, is useful for determining the characteristics of the unbanked because it collects demographic information and data about checking account ownership.[6] It does not collect information, however, about how the unbanked make and receive payments.

For details on payment methods, we sought answers from three specialized surveys and from fieldwork. The first survey was one conducted by John Caskey, as reported in Caskey (1997). His telephone survey asked 900 people

[4] Other data sources give estimates of the unbanked that range from 8 percent to 20 percent. See Hogarth and O'Donnell (1999) for a summary of these sources.

[5] As this article went to press, the 1998 numbers were released. The survey found that the number of households without checking accounts had dropped to 13.2 percent (Kennickell, Starr-McCluer, and Surette 2000).

[6] For more information on the Survey of Consumer Finances and for findings from the 1995 survey see Kennickell, Starr-McCluer, and Sunden (1997).

with incomes less than $25,000 about their use of the "alternative financial sector," e.g., check-cashing outlets, pawnshops, and consumer finance companies. The survey was conducted in only three locations (Atlanta, Oklahoma City, and a group of five smaller Pennsylvania cities), so it is not clear how representative its results are. Nonetheless, the results are valuable for our purposes because the survey was designed to answer questions similar to ours.

The second specialized survey was conducted by Booz, Allen & Hamilton and Shugoll Research (1997). The Treasury commissioned this group to obtain information about the banking patterns of government beneficiaries. Like the Caskey survey, it measures variables that are of interest to us, though some caution should be used in interpreting its statistics. The survey oversamples the smaller government programs, undersamples the larger SSA and SSI programs, and does not adjust the reported results for these sampling rates. Furthermore, this survey was administered in two parts, a telephone survey followed by a mail survey of people whose telephone numbers were unavailable. The mail survey is particularly significant for our purposes because low-income people, who comprise most of the unbanked, are less likely to own a phone. Because the results differed so often between the two types of surveys, we report the results from each separately.

The third survey, also prepared for the Treasury, is Dove (1999), which studied the banking patterns of government beneficiaries. This survey was administered by mail and received 385 responses from individuals without bank accounts.

Responses from two focus groups constitute our final source of information. Focus group participants, drawn from two Richmond area low-income housing developments, were asked questions about their use of financial services, including payment services.[7]

Although information from sources such as focus groups are qualitative and not easily quantified, it can be useful in several ways. First, good qualitative research is a foundation for more formal quantitative research. Evidence from focus groups and other qualitative sources provides an important guide for developing more formal instruments. As we will see, several findings from the two focus groups were quantitatively important factors in the specialized surveys. Second, qualitative research allows investigators to gather more detail and probe further into issues than does quantitative research. In this article, we use the focus group responses to provide additional insight into answers cited by the quantitative surveys. We view this evidence as illustrative, suggestive, and indicative of directions that future research should explore.

[7] More information on how the focus groups were conducted is contained in the Appendix.

Table 1 Demographic Characteristics of the Unbanked

Characteristic	Percent
Overall	12.6
Race/Ethnicity	
Hispanic	29.7
African American	36.9
White	7.4
Other	10.7
Gender	
Male	9.7
Female	19.9
Age	
• 24	28.1
25-34	16.1
35-44	12.3
45-59	10.8
60-64	12.5
, 65	8.1
Average Education	10.8
Income	
• $9,999	38.4
$10,000-$24,999	16.9
$25,000-$49,999	4.8
, $50,000	1.2
Marital Status	
Married	6.7
Unmarried	19.0
Employment Status	
Employed	9.4
Retired	7.9
Laid Off/Unemployed	42.5
Other Not Employed	30.3

Source: Hogarth and O'Donnell (1997).

Who Are the Unbanked?

Most of the unbanked are low-income individuals. Table 1, calculated by Hogarth and O'Donnell (1997), lists demographic characteristics of the unbanked from the 1995 Survey of Consumer Finances. Income stands out as an important indicator of whether or not someone owns a checking account. Among those with $9,999 or less in annual income, 38.4 percent do not own bank accounts. This percentage drops dramatically to 16.9 percent for those with incomes

of $10,000 to $24,999, and to less than 5 percent for those with incomes of $25,000 to $49,999. Other demographic characteristics, such as whether one is a minority, unemployed, young, or single, or possesses a low level of education, are also highly correlated with not owning a checking account. Because these characteristics are negatively correlated with income, Hogarth and O'Donnell (1997) ran a multivariate logistic regression and determined that only three characteristics—having low income, being unemployed, and being of Hispanic descent—remained statistically significant. The implication is that the other demographic characteristics—age and minority, marital, and educational status—were only correlated with being unbanked because they were also correlated with these variables.

How Do the Unbanked Use the Payment System?

People need two types of payment services. One is a means for paying bills. The other is a means for converting a received payment into a usable form, such as a deposit or cash. For people who own a checking account, these services (along with savings services) are bundled together.

Making Payments

For people without checking accounts, the two primary means of making payments are with a money order or in person with cash. A money order is issued by an institution for payment of a specified sum of money collectible from itself. If someone wants to pay using a money order, that person can purchase the order (usually with cash), make it out to the recipient, and mail it. The recipient can then deposit the money order at the recipient's bank. Money orders are sold by banks, convenience stores, grocery stores, check-cashing outlets, and the U.S. Postal Service. At present, the Post Office charges 80 cents per money order.

Some companies allow customers to pay bills in person with cash. For example, utility companies frequently have in-person bill payment offices. Often, bills can be paid in this manner at a third-party location, such as a bank or grocery store. The bank or store accepts cash and in turn transfers funds to the biller's account. The store that collects the payment usually offers these services free of charge and receives payment for the service from the billing institutions.

Dove (1999) reports that 55 percent of the unbanked paid some of their monthly bills with cash, and 50 percent paid some by money order. Caskey (1997) also found that money orders were an important means for bill payment. In Caskey (1997), 84 percent of respondents without deposit accounts reported using a money order at least once a year, while 39 percent reported using money orders more than 30 times in a year. In Dove (1999), for those who do write money orders, the mean number written per month by the unbanked

was 3.3. (The unbanked who do not write money orders were excluded when calculating this average.) At a rate of 80 cents per money order, this payment method would cost on average, $31.68 a year.[8]

Finally, Dove (1999) reports that 20 percent of unbanked respondents sometimes paid some bills via someone else. The other quantitative surveys do not consider this option, and in the qualitative research this bill payment option was not mentioned. Consequently, we do not know much about it. We did find one individual in the focus groups, however, who cashed checks through a relative. Presumably, networks of family and friends are also being used to pay bills.

Receiving Payments

As mentioned previously, banks, thrifts, and credit unions are the most important check-cashing sources for the unbanked. In Caskey's survey, 48.5 percent of the unbanked report that they regularly cashed checks at depository institutions. The percentages for government beneficiaries were 62 percent in the Treasury's telephone survey, 42 percent in the Treasury's mail survey, and 51 percent in Dove's survey. Table 2 reports Caskey's findings on sources of check-cashing services. Table 3 reports the results from the Treasury and Dove surveys.

After depository institutions, the next most important source of check-cashing services is grocery stores. Two of the surveys report that 25 percent of the unbanked regularly use the stores for this purpose, another reports 30 percent, and the final survey reports 36 percent.

The third most important source is check-cashing outlets. Caskey finds that 17.2 percent of unbanked respondents regularly use outlets; the other surveys report that approximately 10 percent use them. According to Caskey, other sources of check-cashing services that charge fees, such as convenience and liquor stores, are regularly used by 4.5 percent of respondents.[9] The remaining sources are used less frequently than check-cashing outlets. For example, friends and relatives are used by 12 percent of the respondents in the Treasury's mail survey and by 7 percent in the Dove survey.

Considering how much attention check-cashing outlets have received regarding their fees, it is interesting that these outlets are only a minor source of check-cashing services. For government and payroll checks, outlets will

[8] Focus group participants indicated substantial variation in the price of money orders. According to respondents, banks were the most expensive while convenience stores were relatively inexpensive, even as low as 39 cents. Presumably, these prices are set so low in order to draw customers with cash into the store.

[9] In the Richmond focus groups none of the respondents reported regularly using a check-cashing outlet or convenience store to cash checks, though they knew fellow community members who did.

Table 2 Check-Cashing Sources (Caskey Survey)

	Percent
Bank, savings & loan, or credit union	48.5
Grocery Store	23.2
Convenience or liquor store	4.5
Check-cashing outlet	17.2
Employer	1.5
Elsewhere	1.5
Did not cash any checks	3.5

Table 3 Check-Cashing Sources

	Treasury's telephone survey (Percent)	**Treasury's mail survey** (Percent)	**Dove's survey** (Percent)
Bank or Credit Union	62	42	51
Grocery Store	30	24	36
Friend or Relative	1	12	7
Check-Cashing Service	10	12	12
Other Retail	3	10	11
Other			10

often charge a fee of 1 to 3 percent of the face value of the check.[10] If a personal check is cashed, the fee to cash it is higher still. Whether these fees are excessive is an open question, but because outlets bear the risk of a bad, forged, or stolen check, and because they often operate in high-crime locations for long hours, there is good reason to think that the fees are not excessive (Caskey 1994). Regardless, the finding that check-cashing outlets are used infrequently is important because it bears directly on our analysis of the decision to own a checking account, as we will see in the next subsection.

We offer a note of caution about this finding on check-cashing outlets. Outlets tend to be more prevalent in larger cities, particularly Chicago and

[10] In addition, check-cashing outlets frequently provide services and products such as bill and tax payment, money orders, and money wires. Where not forbidden by state law, many check-cashing outlets also offer payday loans. To obtain one of these loans, a borrower writes a personal check to the cashier, who agrees not to cash it until the borrower's payday. Such loans tend to be made only to people with stable employment histories. All reported information on check-cashing outlets is taken from Caskey (1994).

New York City (Caskey 1994).[11] The Dove survey reports that 27 percent of the unbanked in urban areas use check-cashing outlets, while only 8 percent of the unbanked in small towns and the same percentage in rural areas use them. Apparently, there are differences between urban and non-urban markets.

Why Are the Unbanked Unbanked?

Many discussions on why the unbanked do not own checking accounts compare the cost of owning a checking account (exclusive of fees for bounced checks) with the cost of using a check-cashing outlet. For example, Doyle, Lopez, and Saidenberg (1998) assume that the cost of owning a checking account with no bounced checks is $44 per year. They compare this cost with that of using check-cashing outlets to cash paychecks at a rate of 1.1 percent of face value. Under their assumptions, the cost of not owning an account is $110 plus the cost of money orders for a family with an income of $10,000, while it is $172 plus the cost of money orders for a family with an income of $15,600 (the 1997 poverty level for a family of four). Since this sum is substantially higher than the $44 estimate, why would anyone choose to live without a checking account?

We argue that for many people, forgoing a checking account is a rational choice. First, we contend that being unbanked is not as expensive as the numbers above indicate. More specifically, we demonstrate that check-cashing outlet fees incorrectly measure the costs of not owning an account. Second, we argue that owning a checking account can be more expensive than $44, because maintaining a very low balance, as many low-income people do, can often result in overdraft fees.

What Are the Costs of Being Unbanked?

As we saw earlier, expensive sources of check-cashing services, like check-cashing outlets and convenience stores, are only used regularly by approximately 20 percent of the unbanked population. The critical issue then is to determine how much the unbanked are paying to cash checks through banks and grocery stores. Unfortunately, none of the surveys explicitly asked respondents how much they paid for check-cashing services but the Caskey and Dove surveys asked respondents if they usually paid fees to cash their checks. Caskey (1997) reports that 59 percent of the unbanked in his survey did *not* usually pay a fee to cash their checks. For its sample, Dove (1999) reports a similar number of 61 percent.

[11] There are fewer check-cashing outlets in midsize cities such as Richmond, Virginia, a metropolitan area with fewer than one million residents. The 1999 Greater Richmond Yellow Pages lists only seven locations that provide check-cashing services. However, many sources of check cashing, such as convenience stores, are not included in this list.

We do not know how much people pay in fees or precisely how these fees are broken down by the check-cashing source. Dove (1999), however, reports that 81 percent of the unbanked who cash their checks at banks do not pay fees. The focus group interviews provide insight as to why some banks do not charge fees. In one case, a non-account holder was able to cash checks for free at a particular bank because her employer held an account there. Presumably, others are finding banks that will cash checks for free—particularly those that are government or payroll checks of locally known companies. Indeed, an employer with many unbanked employees might choose a bank that would cash its employee paychecks without charging fees.

One can further speculate that a bank in a small community, where fraud is difficult, would be more willing to cash a check than a bank in a large city. One of the Dove (1999) findings is consistent with this speculation. In its sample, 53 percent of urban unbanked recipients paid check-cashing fees, while only 29 percent of small-town unbanked recipients paid these fees.

The evidence also indicates that it is inexpensive to cash checks at grocery stores. Most of our information on their practices comes from the focus groups and other qualitative sources. Respondents in the Richmond focus groups reported that the grocery stores they frequented did not charge fees to cash their checks, but that the stores sometimes required a minimum purchase to cash a check. In addition, they reported that using a grocery store for check-cashing services was not always convenient for those without a car. We followed up on these findings by contacting two grocery stores in Richmond to ask them about their check-cashing policies. Both cashed payroll and government checks for free. In addition, we discovered that grocery stores sold money orders and collected bill payments for some companies. Companies would contract with them to collect bill payments. The grocery stores would not charge the consumers but would instead charge the company on a per-bill basis. They did not consider this service costly to provide, since an employee assigned to this duty could usually perform other duties as well. Furthermore, offering these services attracted customers with cash into their stores.

Finally, friends and relatives are cited as a minor source for cashing checks. In the Treasury's mail survey, 12 percent of the respondents mentioned this source and at least one focus group respondent used a relative to cash checks. We can probably assume that these sources provide their service for free.

In summary, $172 (from Doyle et al.) overestimates the cost of cashing checks for a substantial portion of the unbanked population. As the Caskey survey, the Dove survey, and the focus group interviews indicate, many of the unbanked are cashing their checks at banks, grocery stores, and even with friends and relatives at no cost. Furthermore, many payments are being made free of charge. For about two-thirds of the unbanked, particularly those not located in urban markets, the evidence suggests that the costs of not owning a

checking account are very small and probably best approximated by the cost of writing money orders, around $30 by our earlier estimate.

What Are the Costs of Owning a Checking Account?

Both the Caskey and Treasury surveys asked people why they did not own checking accounts. The results are reported in Tables 4, 5a, and 5b.[12] The most common reply was that respondents did not have enough money or enough savings for an account to be worthwhile. It is difficult to evaluate this response or similar responses such as "[I] don't have a need for any." These responses suggest judgment about the relative costs and benefits of owning versus not owning a checking account; they are not informative about the actual costs of owning an account or the relative importance of different costs.

Ultimately, the problem is that these questions ask about motives, and the answers are less reliable than those to questions that require factual responses. For this reason, we will only mention the survey responses when we feel they are useful.

Some Speculation

We speculate that overdrafts are an important reason that checking accounts are unappealing to the unbanked. The possibility of an overdraft is a key difference between payments made by cash and personal checks: No credit is extended with cash payments, but credit is extended, albeit short term, when payments are made by personal check.[13] Overdrafts, because they do not include check-writing services, are not possible with ETAs. If our speculation is correct, removing check writing from the standard checking account has value.

From a customer's perspective, overdraft fees could be a significant deterrent to owning an account. While overdraft fees are avoidable, overdrawing an account is easier to control in theory than in practice, particularly for an account that is frequently near a zero balance. One miscalculation that results

[12] The two Treasury surveys are not directly comparable. Unlike the mail survey, the telephone survey did not give respondents a list of answers from which to choose. However, in the telephone survey respondents were also asked if they agreed whether the following reasons were important for not owning a checking account. Included were "bank fees are too high," "I have no need for bank services," "I don't want anyone else to have records of how much money I have," "I don't trust banks with my money," "bank hours don't match my schedule," and "there are no banks conveniently located near me." At least 20 percent of respondents strongly agreed that each reason was important while the highest number, 40 percent, reported "bank fees are too high." The differences between the aided and unaided responses are a bit troubling. We do not have a theory for this discrepancy and feel that more focused investigation through qualitative methods is warranted.

[13] Technically, the retailer is extending the credit since a bank may return a check for insufficient funds. However, the bank could still accept the check, and in that case it would bear the risk.

Table 4 Reasons for Not Owning a Deposit Account (Caskey survey)

Survey Responses	Percent
Bank account fees are too high	23.1
Fees considered to be the biggest problem	
Monthly account maintenance fees	40.0
Check-writing fees	10.0
ATM fees	11.1
Bounced-check fees	28.9
Banks require too much money just to open an account	22.1
Don't need account because we have no savings	53.3
Not comfortable dealing with banks	17.6
No banks have convenient hours or location	8.5
Banks won't let us open an account	9.5
We want to keep our financial records private	21.6

in two overdrawn check charges can produce a memorably expensive financial experience. For example, a bank's overdraft fees can range from $20 to $35 per check, while merchants will often charge an additional fee.

There is some support in the surveys that overdraft fees deter the unbanked from choosing to own an account. In the Treasury's mail survey, 13 percent of the respondents cited problems managing their money as a reason for not owning a bank account.[14] (In the Treasury's mail survey, respondents could explicitly choose this option; in the telephone survey, respondents could only give unsolicited reports of this response.) Also, 28 percent of respondents in the Caskey survey who complained about fees said that overdraft fees were their main concern.

In the focus groups and other qualitative information sources, money management problems were frequently considered important. Our discussions with bank staff underscored the greater likelihood of significant overdrafts on low-balance accounts, as compared to those with a higher balance. Often with overdrafts, the low-balance account holder tends to close the account, while a high-balance account holder simply pays the service charges. In the Richmond focus groups, several unbanked participants reported that they previously owned bank accounts and suffered losses from overdrawn accounts. Overdraft problems may help explain Caskey's notable finding that 70.7 percent of the unbanked previously had checking accounts.[15] Admittedly, our analysis at this point is merely speculation. However, we feel the connection between

[14] Innumeracy could be one reason for money management problems.

[15] Another argument is that increases in fees explain why account ownership rates decline. Stegman (1999) argues that banks began to charge fees to low-balance customers because of changes in the regulatory environment.

Table 5a Reasons for Not Owning a Deposit Account (Treasury's telephone survey)

Survey Responses	Percent
Don't have enough money to justify/make worthwhile	47
Don't have need for any	21
Fees/costs are too high	6
Problems with managing an account	3
Don't know	20

Table 5b Reasons for Not Owning a Deposit Account (Treasury's mail survey)

Survey Responses	Percent
Don't have enough money to justify/make worthwhile	67
Don't have need for any	27
Fees/costs are too high	24
Problems with managing an account	13
Use another person's account	11
Poor credit history/turned down for one	10
Banks inconveniently located	4
Difficult to get to a bank	4
Keep records private from government	4
Don't want money frozen in event of divorce/lawsuit/judgment	4

overdrafts, bank fees, and the decision to own a checking account is worth further investigation.[16]

In some cases, we can identify specific reasons that respondents do not own a checking account. For some people, the fact that creditors could access a debtor's bank account is reason enough not to own such an account. For example, if someone defaults on a debt, creditors may attach the defaulter's bank account. This concern, primarily raised in the focus groups, is only mildly apparent in the Treasury's surveys. In Caskey's survey, however, 21.6 percent mentioned privacy as a reason for not owning a checking account. Presumably, this reason includes fear of attachment, though it could also include motives such as evading taxes, avoiding the savings limitations on welfare beneficiaries (Edin and Lein 1997), or hiding income from other household members.

[16] Interestingly, roughly 10 percent of respondents in the Caskey and Dove surveys reported that banks would not accept them as customers.

Furthermore, it is worth noting that even though federal benefits are protected from attachment by law, in practice the depositor is responsible for proving that the particular funds may not be attached. When benefits are commingled with other funds, determining which funds are protected and which are not can be complicated, effectively making it too costly for a low-income individual to stop the attachment.

Finally, some have argued that there are people who do not have accounts because (1) banks are inconveniently located and have poor service hours or (2) the unbanked are unaware of check-cashing fees. The surveys report minimal support for the first argument. Four percent cite location and service hours in the Treasury's mail survey and 8.5 percent cite them in Caskey's survey. In one of the Richmond focus groups, location was mentioned as an issue for people who did not own cars. As for the second argument, undoubtedly it is possible that some of the unbanked are naive about price differences, but we are skeptical that this is an important reason for not owning an account. The respondents in the Richmond focus groups were well aware of the costs of using check-cashing outlets or convenience stores but still used them occasionally.

In summary, we find that payment services are relatively inexpensive for many of the unbanked and that check-cashing outlet fees are not representative of the true costs of cashing a check. Furthermore, we speculate that because of bounced check fees, a checking account might be more expensive than the $40 to $50 often estimated. We think that together these factors explain why many of the unbanked do not own a checking account. Our analysis also finds that the cost of being unbanked varies across different groups of people. For example, someone who lives close to a grocery store may be able to obtain payment services at no cost. Someone in a neighborhood with neither a grocery store nor a bank willing to cash the checks for free would probably pay substantial check-cashing fees.

2. EFT '99

In this section, we use our findings on the unbanked to analyze the implementation of EFT '99—the Treasury's plan to encourage government beneficiaries to use direct deposit. Earlier strategies to implement the plan were altered to respond to concerns that the law would unfairly burden the unbanked. We believe our previous analysis explains why that opposition was so strong. Early proposals would have shifted costs to the unbanked.

The driving force behind EFT '99 was the pressure on Congress to reduce federal expenditure as part of the balanced budget compromise. The budget-scoring rules adopted by Congress required that new expenditures be matched by corresponding decreases in spending. Switching government payments from paper to electronic means was scored as savings. The Treasury's Financial Management Service estimates that a fully implemented EFT system would

save the government $100 million per year in printing, processing, and postage costs (U.S. Treasury 2000).[17]

Early Strategies for Implementation[18]

The Debt Collection Improvement Act of 1996 mandated that all government beneficiaries receive their payments electronically, but it gave the Treasury the authority to grant waivers on the basis of four categories: financial hardship, impossibility, cost-benefit, and law enforcement and national security interests.[19] Early proposals to implement this mandate did not make liberal use of the waivers. The first proposal would have required that all government beneficiaries open a bank account in order to receive their payments. Another proposal would have given unbanked beneficiaries a year to open an account. A third proposal would have required that only those beneficiaries who already owned an account had to switch to electronic receipt. Community groups reacted negatively to these proposals, arguing that mandated accounts would adversely affect some low-income people.

The Adopted Strategy

In response to the criticisms of the earlier proposals, the Treasury adopted a strategy of making participation voluntary. The earlier proposals were modified so that anyone who did not sign up for direct deposit would automatically be granted a waiver to receive a check instead. The Treasury also developed the electronic transfer account, specifically designed to appeal to the unbanked. The goal was to encourage banks to offer ETAs and the unbanked to sign up for them.

ETAs

ETAs are low-cost accounts that are designed to receive government payments by electronic direct deposit. These accounts would be available only at federally insured financial institutions that offer them *voluntarily*. When an institution chooses to offer ETAs, the Treasury will reimburse it $12.60 for the one-time cost of setting up each account. The financial institutions offering ETAs would enter into contractual agreements with the Treasury that stipulate the account's specifications. These specifications require that ETAs:

† be an individually owned account at a federally insured financial institution;

[17] Roughly 30 percent of Treasury payments are made by check. From October 1998 to March 1999, approximately 130 million payments were made in this manner.

[18] Stegman (1999) describes the implementation process up until early 1999.

[19] The latter category includes law enforcement payments to informers, who for obvious reasons would prefer not to have an electronic record of a payment from the federal government.

- be available to any individual who receives a federal benefit, wage, salary, or retirement payment and other such deposits as a financial institution agrees to permit;
- charge no more than $3 per month;
- allow at least four free cash withdrawals and balance inquiries per month through any combination of proprietary ATM and/or over-the-counter transactions;
- provide the same consumer protections that are available to other account holders at financial institutions;
- allow access to point-of-sale networks, if this service is available to non-ETA holders;
- require no minimum balance, except as required by federal or state law; and
- send each account holder a monthly statement.

How Does EFT '99 Affect Market Participants?

There are four parties that have been affected by the various possible implementations of EFT '99: the government; the unbanked, who are represented at the policy level by low-income advocacy groups; the banks; and the alternative institutions to banks, such as check-cashing outlets. As noted earlier, the driving force behind EFT '99 is the belief that switching to electronic payments would save the federal government a substantial amount of money. The question for the other affected parties is whether it is worth adopting this means of payment.

The Government

The government's interest in costs and benefits is relatively clear. The more people who switch to electronic payment, the more money it saves. Presumably, there will be cost savings from ETAs, even with the $12.60 payment per account.

The Unbanked

As previously discussed, the decision to forgo owning a checking account is entirely rational. Early EFT '99 proposals that would have mandated beneficiaries to own checking accounts in order to receive payments would not have saved resources for the economy but instead would have shifted costs from the government to the unbanked.

However, even with voluntary participation in the ETA program where there is no danger that EFT '99 merely shifts costs, it is still uncertain whether ETAs will be widely adopted by the unbanked. The critical issue is the elasticity

of demand with respect to price and ETA characteristics.[20] Our assessment is that these elasticities are small.

For the majority of the unbanked who are already cashing checks for free, the pecuniary benefits of adopting an ETA seem small or even negative. An ETA holder will pay $36 a year mainly for check-cashing services already obtained for free. It follows that ETAs may stand their best chance of success in urban areas where the prevalence of costly check-cashing outlets indicates a lack of costless alternatives.

Additional considerations enter the cost-benefit calculations of a potential ETA holder. For example, while the ETA removes the credit extension inherent in a traditional checking account, it does so at a cost in bill payment services.[21,22] An ETA holder will still have to withdraw cash from the account in order to pay bills in person or to purchase a money order. There seems to be little advantage to this method over going to a grocery store to cash checks and to pay bills all at once. For this reason, Caskey (1998) argues that the unbanked need accounts with access to something like a low-cost ATM that would not only supply cash but also money orders. Still, an earlier attempt to encourage the unbanked to own these or similar accounts is discouraging. In the 1980s several states mandated that banks offer "life-line accounts," low-cost checking accounts designed to appeal to the unbanked, but they were not widely adopted (Doyle, Lopez, and Saidenberg 1998).

The Banking Industry

Banks, like the community groups, disliked the early proposals. Many banks feared that political pressures would require them to offer low-cost checking

[20] The price elasticity of a good is the percentage change in demand for it divided by the percentage change in price. It is a measure of how much demand will increase in response to a price change.

[21] Parallels exist with an interesting banking experiment that sought to simplify making small payments by mail in Britain during the late 19th century. Although the British wrote checks during this period, the lower and middle classes did not use them as extensively. Jevons (1897) argues that check use had not extended to these segments of society because the extension of credit involved in a check invited fraud. He was a strong advocate of the Cheque Bank, an institution that avoided this problem by issuing checks with limits on the amount for which they could be written. For example, a depositor who deposits £100 would receive any desired combination of checks as long as the maximum amounts did not sum to more than £100. The depositor would be able to write a check for any amount up to the maximum listed on it. This device greatly facilitated making payments through the postal services since there was no need for change. The Cheque Bank would maintain two balances for each account, the amount in the account, and the amount of credit extended by the checks. When a check was paid, both balances would be adjusted. Ultimately, the Cheque Bank failed in 1900 for reasons that included increased handling of small accounts by the rest of the banking sector and forgeries of its checks, which were apparently easy to cash (Banker's Magazine 1901).

[22] Point-of-sale purchases can still cause an overdraft if the purchases are off-line, that is, there is a delay between purchase and communication with the bank.

accounts to the unbanked. They claimed that regardless of the technological advances of electronic banking, they would lose money by carrying transaction accounts for those who were currently unbanked. In any event, the banks believed a large portion of the $100 million in government savings would be a transfer of cost rather than a savings to society.

Of course these concerns about cost shifting are not an issue with ETAs. Furthermore, in recognition that low service fees may not cover the costs of one of these accounts, the current ETA proposal requires the Treasury to provide the banks a one-time reimbursement of $12.60 to open each ETA. We do not know whether this is enough money to induce banks to offer the accounts. We suspect that this fee plus the service charges will be the only reliable sources of revenue for banks from these accounts. The experience that states have had with a similar type of account—electronic benefit transfer (EBT) account—is that it quickly draws down to a low balance. One study of an EBT pilot estimated interest income from balances to be 19 cents (U.S. Treasury 1997).

Banks do have a noneconomic incentive for offering these accounts. They may receive Community Reinvestment Act credit during the examination process if they offer ETAs. The Federal Financial Institutions Examination Council, an umbrella organization of all bank supervisory agencies, released a notice that financial institutions offering ETAs will be given positive consideration under the service test in the examination process (Federal Register, May 3, 1999).

The Check-Cashing Industry

Blessed with an extensive distribution network but threatened with the likely loss of business from ETAs, the check-cashing industry is trying to ally with financial institutions. For example, Citigroup has signed a deal with the National Check Cashiers Association to issue a debit card to government beneficiaries. A beneficiary would open an account with Citigroup and could use the debit card at ATMs, point-of-sale terminals, and check-cashing outlets that are members of this association. Charges on the account would range from $3 to $6 per month, with $1 to $2 fees for withdrawals and point-of-sale purchases (Keenan 1999).

Any attempt to distribute government benefits through check-cashing outlets has been controversial, mainly because low-income advocacy groups view the check-cashing outlet fees to be exploitative. These concerns have led to recent rules banning financial institutions from providing ETAs in partnership with institutions like check-cashing outlets and liquor stores. Furthermore, the Treasury recently requested public comments asking whether to regulate partnerships between check-cashing outlets and financial institutions that offer non-ETAs (*Federal Register,* January 8, 1999). Some fear that a regulated financial institution may encourage its customers to have their government

checks electronically deposited into a standard deposit account and arrange for a nondepository institution to dispense those funds. Under such an arrangement the nondepository institution could then charge fees, and there would be no regulatory control over that arrangement.

3. CONCLUSION AND AN ALTERNATIVE IMPLEMENTATION

There are good reasons to think that the elasticities of demand for ETAs are small. The most important reason is our finding that many of the unbanked presently obtain their payment services at no or low cost. Still, the unbanked are a heterogenous group and ETAs may well appeal to a portion of them. Since this group has varying needs, one could imagine an alternative implementation that allows for some government cost savings while still letting individuals decide what is best for themselves: Let beneficiaries face the marginal tradeoff between different means of payment. Beneficiaries could be paid to receive electronic payments, could themselves pay to receive a check, or could realize some combination of the two. The point is to have beneficiaries bear the costs of the means of payment, which in the absence of externalities or some sort of market failure align individual tradeoffs with those of society.[23] Those who find it worthwhile to continue receiving a check will do so and those who do not will switch. Ultimately, the beneficiary is best positioned to determine the tradeoff.

APPENDIX

Here we provide background information on field research with focus groups, discuss in more detail their advantages and limitations, and describe how those discussed in the article were conducted. Focus groups are just one type of field research for gathering qualitative information. Other related methods for gathering information include interviews with key informants, community interviews, structured direct observation, and small-scale surveys. Kumar (1993) contains a wealth of information on these methods, including a description of an investigator's experience in using each method. Townsend (1995) contains a good example of qualitative field research followed by a small-scale survey.

[23] One complication with this suggestion is that most government payments (73 percent) are already made electronically. In view of the government's budget constraint, it does not make sense to pay beneficiaries who already have switched. Instead of offering the program to everyone, the government could target subsets of existing check users and offer them cash to switch. The results of these targeting efforts could be used to estimate elasticities of demand, much like the credit card companies do now with their offers.

There can be enormous advantages to using qualitative field research. This type of research can be conducted quickly and inexpensively. Also, the format allows the researcher to learn from the process itself. The give-and-take of open-ended interviewing allows topics to be explored in detail. Further, respondents may raise issues that the researcher may not have been aware of before the interview. Another advantage of field research is that results can be used to develop large-scale formal survey instruments. A survey that does not ask the right questions is of no use.

The idea behind focus groups is that the group interaction generates data and so can itself be used as a source of data. Focus groups historically have been used heavily in marketing, but also have been used in sociology, nursing, and the health sciences. This method can be effective in gathering information from multiple individuals at the same time. A classic source on focus groups is Merton, Fiske, and Kendall (1956).

Like any source of qualitative information, data from focus groups require cautious interpretation. Among other things, participant samples are often non-random. Furthermore, interviewers must be careful that they do not ask leading questions of respondents, and investigators must make sure that they are not just seeing what they want to see when interpreting the interviews. Not surprisingly, there is a large literature in the fields mentioned above that discusses these problems and presents strategies for avoiding them.

The Richmond Focus Groups

Each group of participants was drawn from a Richmond area housing development that is run by a nonprofit housing organization. The first group consisted of 11 individuals, and the second group consisted of five. Participants were recruited by individuals from the nonprofit organization and were not randomly chosen. Their incomes were low; some worked and others received government aid. Interviews lasted two hours and were conducted on site. The moderator was given an outline of questions to guide the discussion, though occasionally the investigators would interject more specific questions. Generally, participants were asked about their uses of and need for financial services, including how they and other members of their community made and received payments. For example, did they own checking accounts? How did they pay their bills? If they did not have accounts, why not?

We have chosen to report these interviews as supplementary sources of information. Though the usual caveats apply to any interpretation of this evidence, several findings were substantiated by the quantitative surveys. For this reason, we are confident that any additional details reported are important. We believe qualitative data-gathering methods of the sort described here will prove to be particularly valuable for studying the low-income household's use of and need for all financial services, not just those involving payment services.

REFERENCES

Bankers', Insurance Managers', and Agents' Magazine. "Liquidation of the Cheque Bank," vol. 71, (January-June 1901), pp. 273–76.

Booz, Allen & Hamilton, and Shugoll Research, *Mandatory EFT Demographic Study,* OMB #1510–00–68. Washington: U.S. Department of the Treasury, September 15, 1997.

Caskey, John P. "Serving the Unbanked." Comments prepared for the Symposium on Financial Services in a Post-Welfare-Reform Society, Federal Reserve Bank of Richmond, April 1998.

_____. *Lower Income Americans, Higher Cost Financial Services.* Madison, Wisconsin: Filene Research Institute, 1997.

_____. *Fringe Banking: Check-Cashing Outlets, Pawnshops, and the Poor.* New York: Russell Sage Foundation, 1994.

Dove Associates, Inc. *ETA Conjoint Research: Final Report and Market Model, Unbanked Federal Check Recipients,* OMB #1510–00–71, May 26, 1999.

Doyle, Joseph J., Jose A. Lopez, and Marc R. Saidenberg. "How Effective is Lifeline Banking in Assisting the 'Unbanked'?" Federal Reserve Bank of New York *Current Issues in Economics and Finance,* Vol. 4 (June 1998).

Edin, Kathryn, and Laura Lein. *Making Ends Meet: How Single Mothers Survive Welfare and Low-Wage Work.* New York: Russell Sage Foundation, 1997.

Hawke, John D. "Mandatory EFT '99," *Financial Access in the 21st Century: Proceedings of a Forum.* Washington: Office of the Comptroller of the Currency, 1997, pp. 37–41.

Hogarth, Jeanne M., and Kevin H. O'Donnell. "Banking Relationships of Lower-Income Families and the Government Trend toward Electronic Payment," *Federal Reserve Bulletin,* vol. 85 (July 1999), pp. 459–73.

_____, and _____. "Being Accountable: A Descriptive Study of Unbanked Households in the U.S.," *Proceedings of the Association for Financial Counseling and Planning Education.* (1997 Conference), pp. 58–67.

Jevons, William Stanley. *Money and the Mechanism of Exchange.* New York: D. Appleton and Company, 1897.

Keenan, Charles. "Citi to Issue Debit Cards through Check Cashers," *American Banker,* (January 19, 1999), p. 18.

Kennickell, Arthur B., Martha Starr-McCluer, and Annika E. Sunden. "Family Finances in the U.S.: Recent Evidence from the Survey of Consumer Finances," *Federal Reserve Bulletin*, vol. 83 (January 1997), pp. 1–24.

Kennickell, Arthur B., Martha Starr-McCluer, and Brian J. Surette. "Recent Changes in U.S. Family Finances: Results from the 1998 Survey of Consumer Finances." *Federal Reserve Bulletin*, vol. 86 (1) January 2000, pp. 1–29.

Kumar, Krishna. *Rapid Appraisal Methods.* Washington: The World Bank, 1993.

Merton, Robert K., Marjorie Fiske, and Patricia L. Kendall. *The Focused Interview: A Manual of Problems and Procedures.* Glencoe, Illinois: Free Press, 1956.

Office of the Federal Register. "Notices," *Federal Register,* vol. 64 (May 3, 1999), p. 23622. Washington: Government Printing Office.

_____. "Proposed Rules," *Federal Register,* vol. 64 (January 8, 1999), pp. 1149–52. Washington: Government Printing Office.

Stegman, Michael A. *Savings and the Poor: The Hidden Benefits of Electronic Banking.* Washington: Brookings Institution Press, 1999.

Townsend, Robert M. "Financial Systems in Northern Thai Villages," *Quarterly Journal of Economics*, vol. 110 (November 1995), pp. 1011–46.

U.S. Treasury, Financial Management Service. *Commonly Asked Questions.* Available: http://www.fms.treas.gov/eft/question.html [February 11, 2000].

_____. "Governmentwide Treasury-Disbursed Cumulative Payment Volume," *Electronic Funds Transfer.* Available: http://www.fms.treas.gov/eft/agency/v33199.html. [December 26, 1999].

_____. *Direct Payment Card Pilot, Statistical Evidence of Float Earnings*, (June 26, 1997).

Explaining the Increased Variability in Long-Term Interest Rates

Mark W. Watson

Monetary policy affects the macroeconomy only indirectly. In the standard mechanism, changes in the federal funds rate, the Federal Reserve's main policy instrument, lead to changes in longer-term interest rates, which in turn lead to changes in aggregate demand. But the links between the funds rate, long rates, and demand may be far from tight, and this potential slippage is a fundamental problem for monetary policymakers. In particular, long-term interest rates sometimes move for reasons unrelated to short-term rates, confounding the Federal Reserve's ability to control these long-term rates and effect desired changes in aggregate demand. Has the link between long rates and short rates weakened over time, therefore making it more difficult for the Federal Reserve to achieve its macroeconomic policy objectives through changes in the federal funds rate?

Such questions naturally arise when one observes the behavior of long-term interest rates. For example, Figure 1 plots year-to-year changes in ten-year Treasury bond yields from 1965 through 1998. (The volatile period of the late 1970s and early 1980s has been masked to highlight differences between the early and later periods.) The most striking feature of the plot is the increase in the variability of long-term rates in the recent period relative to the earlier period. Indeed, the standard deviation of long rates essentially doubled across the two time periods. What caused this increase in variability? Did a change in

I thank Michael Dotsey, Robert Hetzel, Thomas Humphrey, Yash Mehra, Pierre Sarte, and Ross Valkanov for comments on a previous draft of this article. This research was supported by National Science Foundation grant SBR–9730489. The views herein are the author's and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

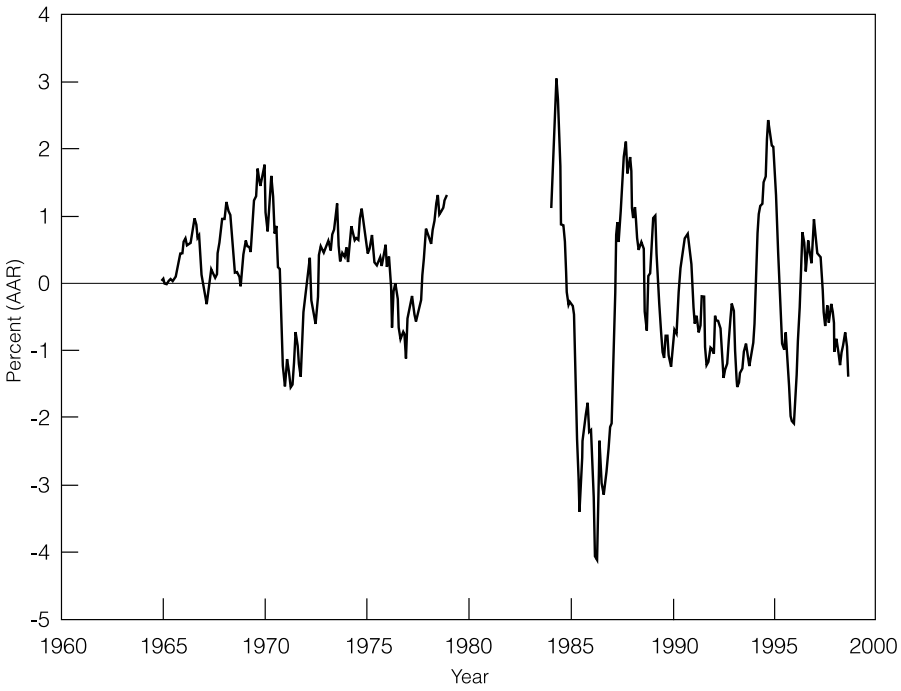
the behavior of short-term interest rates (caused, for example, by a change in Federal Reserve policy) lead to this dramatic increase in long-rate variability? Or, rather, is this change in variability caused by changes in factors unrelated to short-term rates, often described under the rubric of "term" or "risk" premia?

In what follows, we study the behavior of short-term interest rates over the two sample periods, 1965–1978 and 1985–1998, highlighted in Figure 1. It focuses on two key questions. First, has the short-term interest rate process changed? Second, can these changes in the behavior of short-term interest rates explain the increased volatility in long-term interest rates? The answer to both of these questions is yes; our findings suggest no weakening of the link between short rates and long rates and thus no weakening of the link between the Federal Reserve's policy instrument and its ultimate objectives.

The variability in long-term interest rates is tied to two distinct features of the short-rate process: (1) the variability of "shocks" or "innovations" to short-term interest rates, and (2) the persistence (or half-life) of these shocks. In the standard model of the term structure, changes in the variability of short-rate innovations lead to proportional changes in the variability of the long rate. Thus, holding everything else constant, doubling the standard deviation of the innovation in short-term interest rates would lead to doubling the standard deviation of long rates evident in Figure 1.

The relationship between short-rate persistence and long-rate variability is more complicated. To explain this relationship it is useful to consider an example in which the short-term interest rate process can be described by an autoregressive model with one lag (an AR(1)). Let ρ denote the autoregressive coefficient associated with the process. When $\rho = 0$, short rates are serially uncorrelated, and shocks have only a one-period effect on the short-term interest rate. In contrast, when $\rho = 1$, short rates follow a random walk so that shocks to the current value of short rates lead to a one-for-one change in all future short rates. When long-term interest rates are viewed as discounted sums of expected future short-term rates, these different values of ρ imply very different behavior for long-term rates. For example, when $\rho = 0$, a change in the current short rate has no implications for future values of short rates, so long rates move very little. In contrast, when $\rho = 1$, any change in the current short rate is expected to be permanent and all future short rates are expected to change. This change in expected future short rates leads to a large change in the long-term rate. Values of ρ between 0 and 1 are intermediate between these two extremes, but in a subtle way that will turn out to be important for explaining the increased variability in long-term interest rates evident in Figure 1. In particular, for long-lived bonds, a short-rate process with $\rho = 0.9$ generates long rates that behave much more like those associated with $\rho = 0$ than with $\rho = 1$. Put another way, changes in the autoregressive parameter ρ have large effects on the behavior of long-term rates only when ρ is very close to 1. Such a result is familiar from studies of consumption behavior using

Figure 1 Ten-Year Treasury Bond Yields Annual Differences



the present-value model, where the variability of changes in consumption increase dramatically as income approaches a "unit-root" process (Deaton 1987, Christiano and Eichenbaum 1990, Goodfriend 1992, and Quah 1992).

As a preview of the empirical results in later sections, we find that the variability of short-term interest rate shocks was *smaller* in the later sample period than in the earlier period. If there were no other changes in the short-rate process, this decline in short-rate variability should have led to a *fall* in the standard deviation of long-term interest rates of approximately 50 percent, as opposed to the 100 percent *increase* shown in Figure 1. However, we also find evidence of an increase in persistence: for example, the estimate of the largest autoregressive root in the short-rate process (the analogue of ρ from the AR(1) model) increased from 0.96 in the early period to nearly 1.0 in the later period. By itself, the increase in persistence should have led to a three-fold increase in the standard deviation of long rates. Taken together, the decrease in short-rate variability and increase in persistence explain remarkably well the increase in the variability of long rates evident in the data.

The estimated change in the persistence of the federal funds process has important implications for the Federal Reserve's leverage on long-term rates. For example, the estimated autoregressive process for the early sample period implies that a 25 basis point increase in the federal funds rate will lead to only a 3 basis point increase in ten-year rates. The autoregressive process for the later period implies that the same increase in the federal funds rate will lead to a 15 basis point increase in ten-year rates. Alternatively, the increase in persistence makes it possible to achieve a given change in the long rate with a much smaller change in the federal funds rate. The "cost" of increased leverage is the implicit commitment not to reverse changes in the federal funds rate, that is, to maintain the persistence in the short-rate process. The benefit of increased leverage is the reduced variability in the short-term rate. These costs and benefits are discussed in detail by Woodford (1999), who argues that it may be beneficial for the monetary authority to commit to making only persistent changes in its policy instrument.

The article is organized as follows. Section 1 documents changes in the variability of both long-term and short-term interest rates from the 1960s to the present. Here we document the decrease in variability of short-term interest rates (the federal funds rate and three-month Treasury bill rates) but an increased variability in longer-term rates (one-, five-, and ten-year Treasury bond rates). The relative increase in variability is shown to depend on the horizon of the interest rate—it is much higher for ten-year bonds than for one-year bonds, for example.

Section 2 studies changes in the persistence of short-term interest rates over the two sample periods. It begins by using a hypothetical AR(1) model for short-term interest rates to quantify the potential effects of short-rate persistence on the variability of long-term interest rates. The calculations are carried out using a standard model linking long rates to short rates—the expectations model with a constant term/risk premium. In this model, changes in long-term interest rates reflect changes in current and future values of short-term interest rates. The persistence of short-term interest rates is important because it affects the forecastability of short-term rates and thus the effect of changes in the short rate on long rates. The results indicate that, when ρ is very near 1, a relatively small change in ρ can lead to a large change in the variability of long-term interest rates.

Also in Section 2 we present empirical estimates of the short-term interest rate processes for the early and later sample periods using monthly values of the federal funds rate. These estimated processes show a fall in the variance of the short rate but an increase in persistence. Statistical inference about persistence is complicated by the near unit-root behavior of the short rate. This behavior leads to bias in the ordinary least squares (OLS) estimates and a nonstandard sampling distribution for test statistics for shifts in the process across the two sample periods. The article corrects the OLS estimates for bias using a procedure

developed in Stock (1991) and develops a new statistical test for a change in an autoregression that can be applied when data are highly persistent.

In Section 3 the variance of long-term interest rates is calculated using the expectations model together with the estimated processes for the short rate. These calculations show that the changes in the estimated short-rate process lead to increases in long-rate variability quite similar to the change found in the long-rate data.

Finally, Section 4 discusses the robustness of the empirical conclusions to specifics of the econometric specification, and Section 5 concludes. Econometric details concerning tests for changes in the persistence of the short-rate process are given in the Appendix.

1. CHANGES IN THE VARIABILITY OF U.S. INTEREST RATES

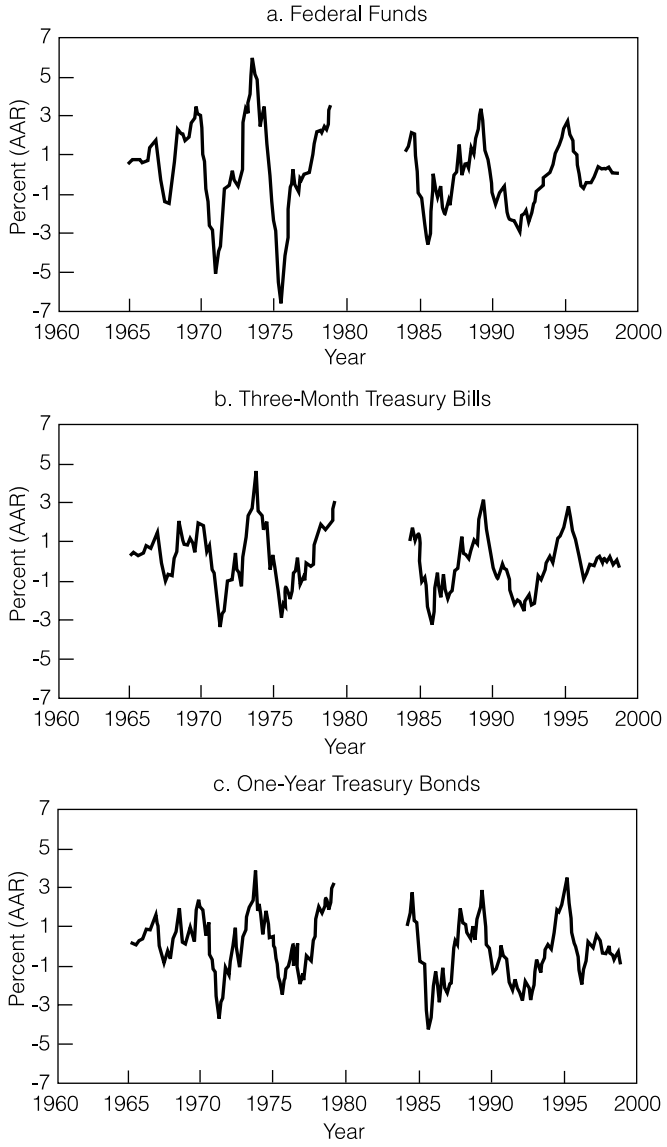
The first task is to examine shifts in the volatility of market interest rates. Figure 2 plots year-over-year changes in six different interest rates over 1965 to 1998. As in Figure 1, the data from 1979 to 1984 are masked to highlight differences between the early sample period (1965:1–1978:9) and the more recent period (1985:1–1998:9). The interest rates range from very short maturity (the federal funds rate) to long maturity (ten-year Treasury bonds and AAA corporate bonds).[1] Each series is a monthly average of daily observations of the interest rates measured in percentage points at annual rates. Table 1 presents standard deviations for changes in interest rates over different sample periods. Panel a reports results for the year-over-year changes plotted in Figure 2, panel b reports results for monthly changes ($R_t - R_{t-1}$), and panel c reports standard deviations of residuals from estimated univariate autoregressions.

As seen in the figures and table, the volatility of long-term rates is much higher in the recent period than in the 1965–1978 sample period, but that is not the case for short-term rates. For example, from panel a of Table 1, the standard deviation of year-over-year changes in ten-year Treasury bond rates increased from 0.69 (69 basis points) in the 1965–1978 period to 1.29 (129 basis points) in the 1985–1998 period. A similar large increase is evident for AAA Corporate bond rates and for five-year Treasury bonds. At the shorter end of term structure, volatility did not increase. Indeed there is a substantial fall in the variability of the federal funds rate from 2.44 (244 basis points) to 1.50 (150 basis points).

The remainder of the table investigates the robustness of this conclusion about volatility both with respect to sample period and data transformation.

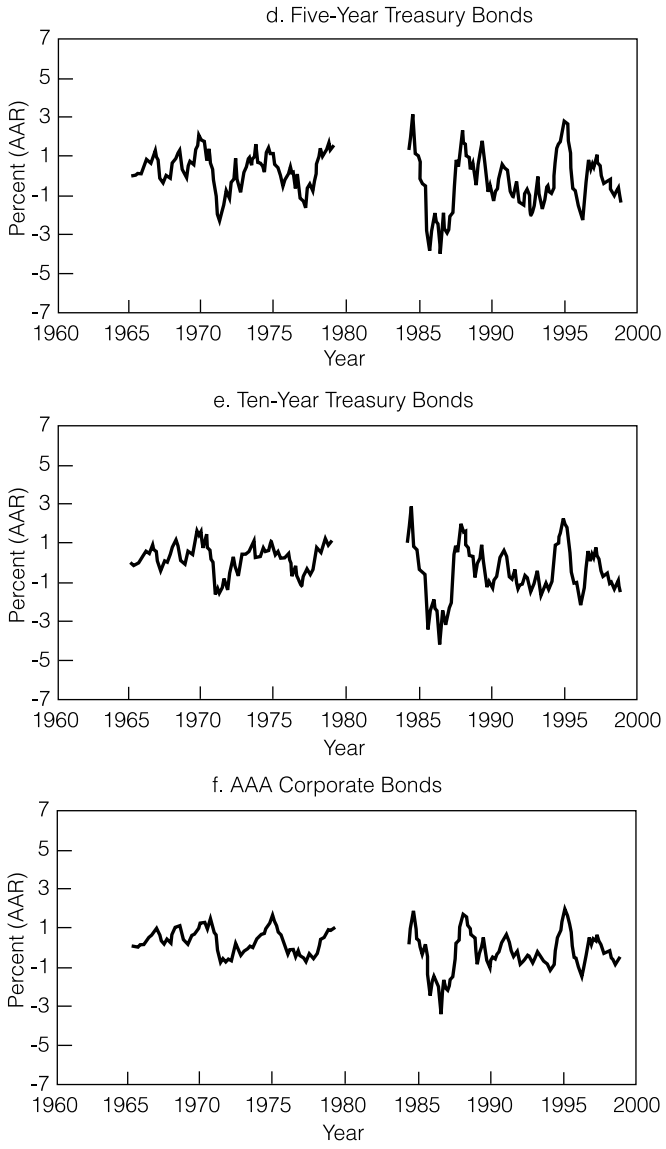
[1] All of the data are from the DRI database. The series are FYFF, FYGM3, FYGT1, FYGT5, FYGT10, and FYAAAC.

Figure 2 Annual Differences of Interest Rates



As shown on the table, this conclusion does not depend on the precise dates used to define the "early" and "recent" periods. These 1965–1978 and 1985–1998 dates were chosen somewhat arbitrarily, and the same volatility results hold for a wide range of cutoff dates used to define the sample periods. Consequently,

Figure 2 Annual Differences of Interest Rates



defining the early period as 1955–1978 and the recent period as 1992–1998 leads to the same conclusions. However, results do change if the volatile period of the late 1970s and early 1980s is included: from Table 1 interest rates were much more volatile in this period than they were either before 1979 or after

Table 1 Standard Deviations of Interest Rate Changes
(Percent at Annual Rates)

a. Year-over-Year Differences

	Interest Rate					
Sample Period	**FedFunds**	**3M-TB**	**1Y-TB**	**5Y-TB**	**10Y-TB**	**Corp**
1965:1–1978:9	2.44	1.50	1.40	0.89	0.69	0.60
1985:1–1998:9	1.50	1.37	1.54	1.40	1.29	1.00
1978:10–1984:12	4.12	3.29	3.12	2.35	2.06	1.81
1955:1–1978:9	2.02	1.33	1.30	0.83	0.62	0.51
1992:1–1998:9	1.28	1.17	1.38	1.24	1.07	0.81

b. First Differences

	Interest Rate					
Sample Period	**FedFunds**	**3M-TB**	**1Y-TB**	**5Y-TB**	**10Y-TB**	**Corp**
1965:1–1978:9	0.44	0.37	0.37	0.26	0.20	0.13
1985:1–1998:9	0.23	0.21	0.28	0.30	0.27	0.21
1978:10–1984:12	1.36	1.10	1.03	0.68	0.57	0.47
1955:1–1978:9	0.38	0.32	0.32	0.22	0.17	0.11
1992:1–1998:9	0.15	0.16	0.23	0.26	0.23	0.17

c. AR Innovations

	Interest Rate					
Sample Period	**FedFunds**	**3M-TB**	**1Y-TB**	**5Y-TB**	**10Y-TB**	**Corp**
1965:1–1978:9	0.38	0.35	0.34	0.25	0.19	0.12
1985:1–1998:9	0.21	0.18	0.25	0.27	0.25	0.19
1978:10–1984:12	1.23	0.94	0.87	0.56	0.50	0.40
1955:1–1978:9	0.35	0.30	0.29	0.21	0.16	0.10
1992:1–1998:9	0.14	0.15	0.21	0.23	0.20	0.16

Notes: Entries are the sample standard deviations of the series over the sample period given in the table's first row. Year-over-year differences are $R_t - R_{t-12}$, first differences are $R_t - R_{t-1}$, and AR innovations are residuals from AR(6) models that incorporate a constant term.

1984. Finally, the results from different panels show that the same qualitative conclusion follows when year-over-year differences are replaced with monthly differences or with residuals from univariate autoregressions. For example, the standard deviation of the residuals in a univariate autoregression for ten-year Treasury bond rates increased from 19 basis points in 1965–1978 to 25 basis points during 1985–1998 (see panel c). The corresponding standard deviation for the federal funds rate fell from 38 to 21 basis points.

Since the variability of short-term rates was smaller in the later sample period than in the early period, it is clear that changes in the variability of short rates cannot explain the increased variability of long rates. We will have to look elsewhere, and with that in mind, the next section investigates changes in persistence in the short-rate process.

2. CHANGES IN THE PERSISTENCE OF U.S. INTEREST RATES

Before examining the empirical results on the persistence of short-term interest rates, it is useful to review the mechanism that links changes in short-rate persistence with changes in long-rate variability. This mechanism can be described using a simple expectations model of the term structure. Thus, let R_t^h denote the yield to maturity on an h-period pure discount bond, and assume that these yields are related to short-term rates by

$$R_t^h = \frac{1}{h}\sum_{i=0}^{h-1} E_t R_{t+i}^1,$$

where R_t^1 is the corresponding rate on a one-period bond. This relation can be interpreted as a risk-neutral arbitrage relation. Now, suppose that short-term rates follow an AR(1) process

$$R_t^1 = \rho R_{t-1}^1 + \varepsilon_t$$

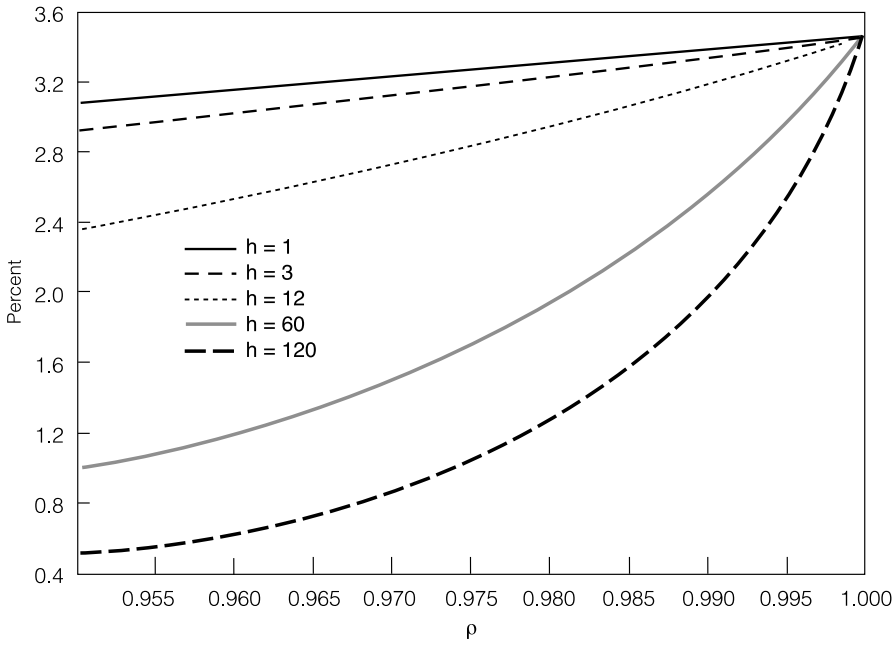
so that $E_t R_{t+i}^1 = \rho^i R_{t+i}^1$ for $i \geq 1$. Then

$$R_t^h = R_t^1 \left[\frac{1}{h}\sum_{i=0}^{h-1} \rho^i\right]$$

so that long rates are proportional to short rates, with a factor of proportionality that is an increasing function of the persistence parameter ρ. Complications to the model (incorporation of term/risk premia, allowance for coupon payments, etc.) change details of the link between long rates and short rates. They do not, however, change the key feature of the model—namely, that long rates depend on a sequence of expected future short rates and that the variance of this sequence depends critically on the persistence of shock to short-term rates.

Of crucial importance is the quantitative impact of short-rate persistence on long-rate variability. Figure 3 gives a sense of this impact. Using the expectations relation given above, it plots the standard deviation of year-over-year changes in R_t^h (that is, $R_t^h - R_{t-12}^h$) as a function of ρ. Results are shown for different maturities h, and the scale of the plot is fixed by setting the innovation

Figure 3 Annual Differences in Interest Rates Implied Standard Deviation from AR(1) Model



variance of short rates (σ_ε^2) equal to 1. The plot shows the functions for values of ρ between 0.95 and 1.00, which is the relevant range for the monthly data studied in this article. For short maturities (small values of h) ρ does not have much of an effect on the standard deviation interest rate. For example, as ρ increases from 0.95 to 1.00, the standard deviation of one-period rates increases by a factor of 1.1 (from 3.1 to 3.5). However, ρ has a large effect on the variability of long-term interest rates. When $h = 120$ (a ten-year bond when the period is a month), then as ρ increases from 0.95 to 1.00, the resulting standard deviation of long rates increases by a factor of 7 (from 0.5 to 3.5). Moreover, the rate of increase in the standard deviation increases with the value of ρ. Thus, the implied changes in the volatility of long rates across sample periods will depend both on the level of ρ and on its change.

Having considered the analytical importance of persistence, we now examine the empirical evidence on it. Table 2 contains estimates of the persistence in short-term rates for the two sample periods. Results are presented for both the federal funds and the three-month Treasury bill rate. Univariate autoregressions are fit to the series, and persistence is measured by the largest root of the implied autoregressive process. This largest autoregressive root determines the effect of

Table 2 Largest Autoregressive Roots for Short-Term Interest Rates

	Sample Period							
	1965:1–1987:9			1985:1–1998:9			Chow Test	
Interest Rate	ρ_{ols}	ρ_{mub}	90% CI	ρ_{ols}	ρ_{mub}	90% CI	F_ρ	P-Value
Federal Funds	0.97	0.96	0.91-1.01	0.98	1.00	0.94-1.02	1.19	0.30-0.64
3-Month TBill	0.96	0.98	0.93-1.02	0.98	0.99	0.94-1.02	1.17	0.31-0.64

Notes: ρ_{ols} is the OLS estimate of ρ constructed from an AR(6) model that included a constant term. ρ_{mub} is the median-unbiased estimator of ρ constructed from the Dickey-Fuller τ^μ statistic as described in Stock (1991). The 90 percent confidence interval is also computed from τ^μ using Stock's procedure. F_ρ is the Chow F-statistic for testing for change in ρ across the two sample periods. The column labeled P-value shows the upper and lower bound for the F-statistic P-value using the procedure described in the Appendix.

shocks on long horizon forecasts of short rates and therefore summarizes most of the information about the link between short rates and long-term interest rate variability. We denote the parameter by ρ as in the AR(1) model discussed above.

The first entry for each sample period is the OLS estimate of ρ (denoted ρ_{ols}) computed from an AR(6) model. (The next section will discuss the robustness of results to the lag length in the autoregression.) The values of ρ_{ols} are very large both for the two interest rates and the two sample periods. The implication is that short rates were apparently highly persistent in both sample periods. There is some evidence of a small increase in ρ in the latter sample period: the value of ρ_{ols} increases from 0.97 to 0.98 for federal funds and from 0.96 to 0.98 for three-month Treasury bills. However, interpreting these changes is difficult because of statistical sampling problems associated with highly persistent autoregressions. These problems are well known in autoregressions with unit roots, but similar problems also arise when roots are close to unity. To aid the reader, we digress with a short statistical primer before discussing the other entries in Table 2.

When values of ρ are close to 1 and the sample size is moderate (as it is here), then the sampling distributions of OLS estimators and test statistics differ markedly from the distributions that arise in the classical linear regression model. In particular, ρ_{ols} is biased, and the usual t-statistics have non-normal distributions. One cannot construct confidence intervals for ρ in the usual way. Of course, as long as ρ is strictly less than 1, the usual asymptotic statistical arguments imply that these difficulties disappear for a "suitably" large sample size. Unfortunately, the sample size used in this article (like that commonly used in empirical macroeconomic research) is not large enough for the conventional asymptotic normal distributions (based on stationarity assumptions)

to provide an accurate approximation to the sampling distribution of the usual OLS statistics. We must use alternative and more accurate approximations.

In empirical problems when ρ is close to 1 (say in the range 0.90–1.01) and the sample size is moderate (say less than 200 observations), econometricians have found that "local-to-unity" approximations provide close approximations to the sampling distribution of OLS statistics.[2] In the present context, these approximations will be used to construct unbiased estimators of ρ, confidence intervals for ρ, and Prob-values in tests for changes in ρ over the two sample periods. Specifically, "median-unbiased" estimates and confidence intervals for ρ are constructed from the Dickey-Fuller τ^{μ} statistic using the procedures developed in Stock (1991).[3] Tests for changes in ρ across the two sample periods are carried out using the usual Chow-F statistic. This statistic is computed as the Wald statistic from changes in the values of ρ_{ols} over the two sample periods. The regressions are estimated separately in each sample period, so that all of the coefficients are allowed to change, but the Wald statistic tests for a change in the largest root only. (Changes in the other autoregressive parameters will have little effect on the variance of long rates, so we focus the test on the largest root.) The statistical significance of the Chow statistic can be determined using Prob-values computed from the local-to-unity probability distributions. These alternative Prob-values are described in detail in the Appendix. As the Appendix shows, the Prob-value depends on the true, and unknown, value of ρ. Thus, rather than reporting a single Prob-value, we report an upper and lower bound.

With this background, the reader can now understand other entries in Table 2. The unbiased estimates are reported in the column labeled ρ_{mub} (the *mub* subscript stands for "Median UnBiased"), and these are followed by the 90 percent confidence interval for ρ. The point estimates suggest that persistence was higher in the second period; for example, using the federal funds rate, the value of ρ_{mub} increased from 0.96 to 1.00. However, the confidence intervals show that there is a rather wide range of values of ρ that are consistent with the data—the confidence interval, which for federal funds in the first period is 0.91–1.01, shifts up to 0.94–1.02 in the second period. The overlap in these confidence intervals suggests that the apparent shift in ρ is not highly statistically significant, and this conjecture is verified by the Chow-statistic, which has a Prob-value that falls between 0.30 and 0.64. Thus, there is some evidence of a shift in the largest root, in a direction consistent with the behavior of long-term rates, but the shift is small and the exact magnitude is difficult to determine because of sampling error. However, when ρ is near 1, small changes in its value can cause large changes in the variability of long-term interest rates.

[2] Important early references in econometrics include Cavanagh (1985), Phillips (1987), and Stock (1991).

[3] The median-unbiased estimator, which will be denoted ρ_{mub}, has the property that $\text{Prob}(\rho_{mub} \leq \rho) = \text{Prob}(\rho_{mub} \geq \rho) = 0.5$.

3. IMPLICATIONS OF THE CHANGES IN THE SHORT-RATE PROCESS ON LONG-RATE VARIABILITY

The changes in long-rate volatility associated with the changes in the short-rate process depend on the specifics of the model linking short rates to long rates. Before we compute the variability of long rates associated with the estimated short-rate processes from the last section, three issues need to be addressed in the present context.

First, the data used here, while standard, are not ideal. The data are not point sampled but rather are monthly observations of daily averages. The bonds contain coupon payments, which were missing in the simple theory presented above. The calculations presented below are based on two approximations. First, the process for one-month rates is estimated using the federal funds data. This is a rough approximation that uses a monthly average of daily rates as a monthly rate. As it turns out, similar results obtain if the federal funds process was replaced with the estimated process for the three-month Treasury bills, so the precise choice of short rate does not seem to matter much. The second approximation adjusts the present-value expectations model for coupon payments using the approximation in Shiller, Campbell, and Schoenholitz (1983). Specifically, the expectational equation for long rates becomes

$$R_t^h = \frac{1-\beta}{1-\beta^h} \sum_{i=0}^{h-1} \beta^i E_t R_{t+i}^1,$$

where $\beta = 0.997$.

The second issue involves the expectations theory described above. That model used an AR(1) driving process for short rates, and constructed expectations using this process. The univariate process for short rates is more complicated than an AR(1) process; moreover, one can form short-rate expectations using a richer information set than one containing only lags of short rates. Extending the calculations to account for a higher-order univariate AR process is straightforward, as the exercise merely involves computing the terms $E_t R_{t+i}^1$ from a higher-order AR model. However, to account for a wider information set is more problematic. A standard and powerful approach to this problem is to construct bounds on the implied variance of long rates from the short process, using, for example, the approach in Shiller (1981). Unfortunately, this approach requires stationarity of the underlying data, so the bounds are likely to be inaccurate for the highly persistent data studied here. West (1988) proposes bounds for the expectational present-value model based on the innovations in the univariate processes and shows that these bounds hold for integrated as well as stationary processes. But as it turns out, West's results hold only for the

infinite horizon model, and the model here is finite horizon.[4] Another approach is simply to specify a more general information set and carry out the analysis using, say, a vector autoregression (VAR) instead of a univariate autoregression. However, the statistical analysis becomes increasingly complicated in a VAR with highly persistent variables. For all of these reasons, the analysis here will be carried out using a univariate AR.

Finally, the calculations reported here ignore all term/risk premia and other deviations from the simple expectations theory. As mentioned above, even in more complicated versions of the models, the first-order impact of short-rate persistence on long-rate variability occurs through the expected present-value expression from the version of the model used here.

With these limitations in mind, consider now the implied variability in long-term rates. The results are summarized in Figure 4 and in Table 3, which shows the implied variability of interest rates computed from the expectations model, using the estimated short-rate process over the different sample periods and for different values of ρ. Results are shown for four maturities. Each panel of Figure 4 shows the variability of year-over-year changes in the interest rate implied by the estimated AR(6) model for the federal funds rate, where the estimates are derived by imposing the value of ρ shown on the x axis. Results are shown for both sample periods. Highlighted on the graphs are the results that impose the OLS and the median-unbiased estimates of ρ from Table 2. (A circle denotes the value of ρ_{ols}; a square denotes ρ_{mub}.) In each panel, the variance function for the second period lies below the function for the first period. This shift is caused by the decrease in variance of the AR errors estimated for the second period. The vertical distance between the curves shows the change in variance for a given value of ρ. To compute the variance across periods, the value of ρ in each sample period must be specified. In terms of the figures, the vertical displacement of the plotted circles gives the change in variability across the two periods using the OLS estimates of ρ (ρ_{ols}). The displacement of the squares gives the change using the median-unbiased estimator (ρ_{mub}). The implied standard deviation for the four maturities in both sample periods and for ρ_{ols} and ρ_{mub} are given in Table 3. For comparison, the table also gives the period-specific sample standard deviations for the federal funds rate and the rates on one-, five-, and ten-year Treasury bonds.

There are substantial differences in the results across the four panels in Figure 4. For one-month rates (panel a), variability is essentially independent

[4] In West's present-value model $y_t = E_t \sum_{i=0}^{h} \beta^i x_{t+i}$, the key restriction is that $E_t \beta^h x_{t+h}$ converges to zero in mean square as $h \to \infty$. This suggests that West's bounds will provide a good approximation in the finite horizon model so long as $E_t \beta^h x_{t+h}$ is small. Thus, the quality of the approximation will depend on the size of $(\beta\rho)^h$, where ρ is the largest autoregressive root. In the term structure model $\beta = 0.997$, and the x_t process is highly persistent, with a largest autoregessive root of, say, $\rho = 0.99$. Thus, for $h = 120$, $(\beta\rho)^h = 0.16$, which implies that $E_t \beta^h x_{t+h}$ will often be substantially different from 0.

Table 3 Standard Deviation of Annual Changes in Interest Rates

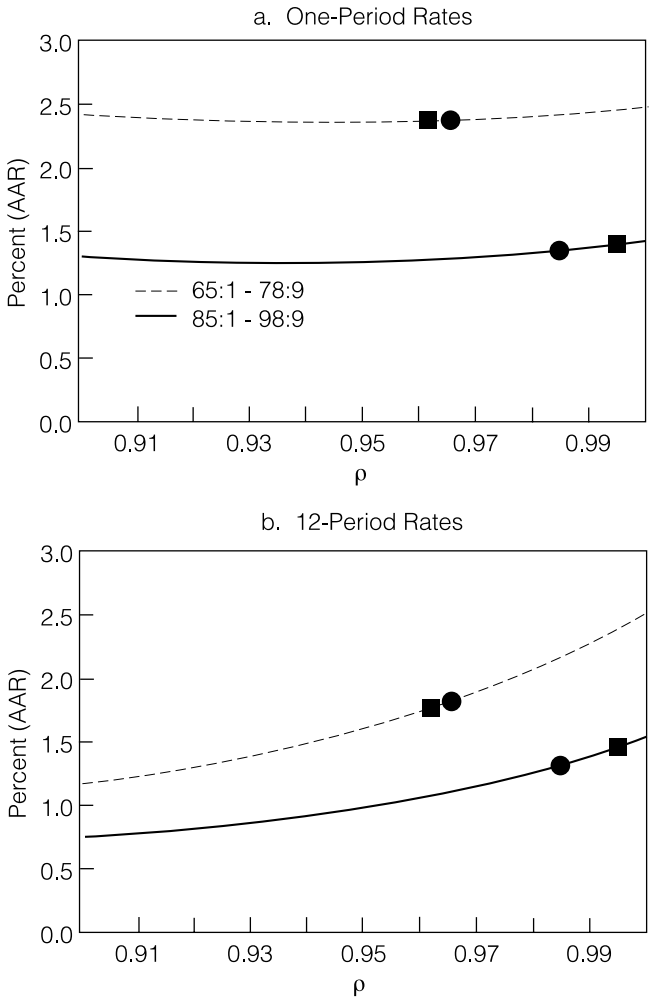
Maturity	Sample Period					
	1965:1–1978:9			1985:1–1998:9		
	Actual	Implied by		Actual	Implied by	
		ρ_{ols}	ρ_{mub}		ρ_{ols}	ρ_{mub}
1 Month	2.44	2.39	2.39	1.50	1.36	1.41
12 Month	1.40	1.80	1.76	1.54	1.31	1.46
60 Month	0.89	0.57	0.52	1.40	0.76	1.22
120 Month	0.69	0.31	0.29	1.29	0.46	0.98

Notes: Entries show the actual (sample value) and implied standard deviations of year-to-year changes in interest rates ($R_t^h - R_{t-12}^h$) for different horizons. Entries labeled Actual are taken from Table 1 and are the sample values for the federal funds rate and the rates for one-, five-, and ten-year Treasury bonds. The columns labeled ρ_{ols} and ρ_{mub} were computed using the expectations model and the estimated AR(6) processes using the federal funds rate over the sample periods shown and imposing the values of ρ_{ols} and ρ_{mub} listed in Table 2. These values correspond to the circles and squares shown in Figure 4.

of ρ and thus the model predicts a substantial decrease in variability during the second period. Since the federal funds rate data were used to estimate the short-rate process, this decrease in variability is essentially equal to the sample values—see the first row of Table 3. For one-year rates (panel b of Figure 4 and the second row of Table 3), variability is also predicted to decrease in the second period, but the decrease is far less than for one-month rates and depends on which estimator is used for ρ. (The implied decrease in the standard deviation is 49 basis points using ρ_{ols} and 30 basis points using ρ_{mub}.) In the sample, there was a small increase (14 basis points) in the standard deviation of one-year interest rates. At longer maturities (panels c and d of Figure 4 and the last two rows of Table 3), variability is predicted to increase in the second period, and again, the amount of the increase depends on the estimator of ρ that is used. The increase is not particularly large using ρ_{ols} (less than 20 basis points); however, it is much larger using ρ_{mub} (70 basis points). The small bias correction incorporated in ρ_{mub} results in this large difference because it pushes the second-period estimate of ρ very close to 1 and because the variance function is rapidly increasing in this region.

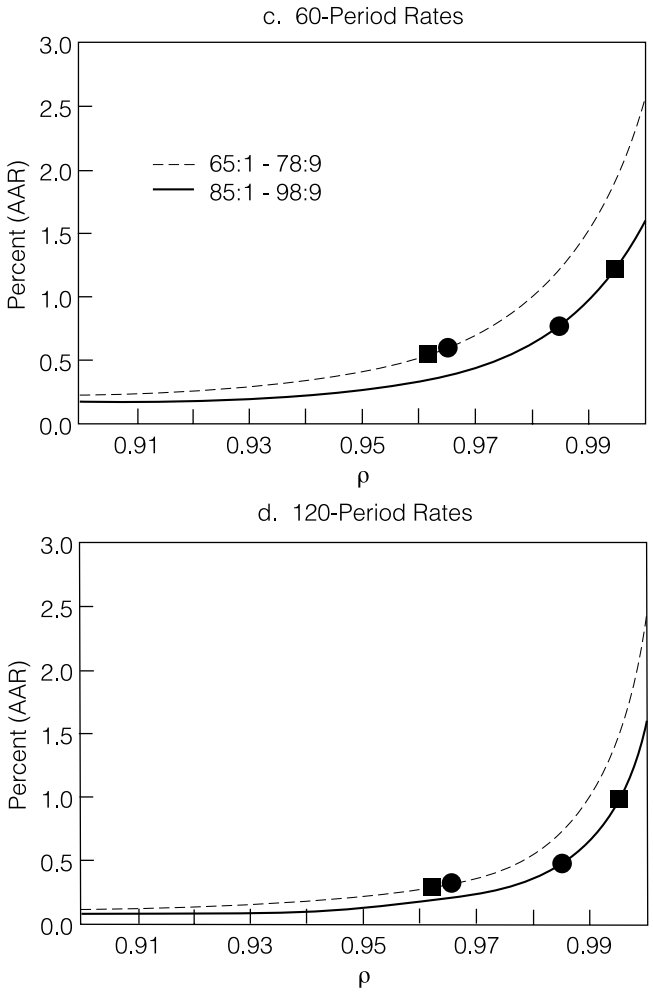
While the estimated difference in persistence, as measured by ρ_{mub}, explains much of the increase in variability in long-term interest rates, much of that variability is still unexplained. For example, in the first sample period the model's implied standard deviation for five-year rates is 52 basis points, while the sample standard deviation of actual five-year rates is 89 basis points. This leaves a "residual" component, orthogonal to short-term rates, with a standard

Figure 4 Annual Differences in Interest Rates Implied Standard Deviation from Fitted AR Models
(Circle = ρ_{ols}, Square = ρ_{mub})



deviation of 72 basis points ($72 = \sqrt{89^2 - 52^2}$) representing the difference between the actual five-year rates and the value implied by the expectations model. Interestingly, a residual component of similar size (69 basis points) is necessary in the second sample period. (A somewhat larger residual is required for ten-year rates.) Thus, although the simple expectations model with constant term/risk premia and simple information structure leaves much of the variability

Figure 4 Annual Differences in Interest Rates Implied Standard Deviation from Fitted AR Models
(Circle = ρ_{ols}, Square = ρ_{mub})



in long rate unexplained in both sample periods, it does explain the lion's share of the increase in variability across the two sample periods.

The results derived here, based on a simple version of the expectations theory of the term structure, are consistent with results derived by other researchers using reduced-form time-series methods. For example, the expectations theory, together with a process for the short-term interest, can be used to calculate

the change in the long rate associated with a given change in the short rate. Using the first-period estimates (and the values of ρ_{mub} shown in Table 2) the model predicts that a 25 basis point change in the federal funds rate would lead to a 3 basis point change in the ten-year bond rate. The second-period estimates imply that the same 25 basis point change in the federal funds rate would lead to a 15 basis point change in the long rate. Mehra (1996) estimates a reduced-form time-series model (a vector error correction model) of long rates and inflation over the 1957–1978 and 1979–1995 sample periods. His estimated models predict that a 25 basis point change in the federal funds rate led to a 3 to 7 basis point change in long rates in the early period and a 7 to 12 basis point change in the later period.

4. ROBUSTNESS OF RESULTS

This section discusses the robustness of the article's main findings to specification of lag length in the autoregression and to choice of sample period. The empirical results are summarized in Table 3. The first row in each panel of the table shows the results from the specification used in the last section, so these results are the same as reported in Table 2. Each of the following rows summarizes results from a different specification of either lag-length or sample period. Panel a of Table 4 shows results for the federal funds rate and panel b shows results for the three-month Treasury bill rate.

The AR lag length of 6 used in the baseline specification was suggested by the Akaike Information Criteria (AIC) and by t-tests on the autoregressive coefficients. Much shorter lag lengths were suggested by the Schwartz criteria (BIC). Table 2 shows results from specifications using 2, 4, and 8 lags. Each of these alternative specifications yield first-period estimates of ρ that are lower than the estimates from the AR(6) model; second-period estimates are essentially unchanged. The first-period differences in ρ_{ols} are small, but the differences are more substantial for the ρ_{mub}. Ignore for the moment the large amount of sampling error associated with these estimates. Even so, the new point estimates have little effect on the variance of long-term rates. From Figure 3, the long-rate variance function is relatively flat over the range of first-period ρ estimates given in Table 3. Thus, from Figure 3, the implied first-period standard deviation of long-term interest rate changes is 0.11 when $\rho = 0.93$ and increases to only 0.18 as ρ increases to 0.96. (The first ρ figure is the value of ρ_{mub} from the AR(4) first-period model; the second figure is the corresponding value of ρ_{mub} in the AR(6) model.) Both of these specifications imply a much larger second-period standard deviation (1.48 and 0.975 for the AR(4) and AR(6) models, respectively) since the second-period values of ρ_{mub} are very close to 1.0 in both specifications. Thus lag-length choice appears to have little effect on the qualitative conclusions.

Table 4 Largest Autoregressive Roots for Different Specifications

a. Federal Funds Rate

Specification Change from Baseline	First Sample Period			Second Sample Period			Chow Test	
	ρ_{ols}	ρ_{mub}	90% CI	ρ_{ols}	ρ_{mub}	90% CI	F_ρ	P-Value
None	0.97	0.96	0.91-1.01	0.98	1.00	0.94-1.02	1.19	0.30-0.64
AR(2)	0.96	0.92	0.86-1.00	0.99	1.00	0.95-1.02	2.46	0.13-0.43
AR(4)	0.96	0.93	0.87-1.01	0.99	1.00	0.95-1.02	2.34	0.14-0.45
AR(8)	0.96	0.95	0.90-1.01	0.99	1.00	0.95-1.02	1.65	0.22-0.55
SD 1955	0.98	0.98	0.96-1.01	0.98	1.00	0.94-1.02	0.06	0.81-0.93
SD 1992, AR(2)	0.96	0.92	0.86-1.00	0.98	1.01	0.95-1.04	1.55	0.24-0.57

b. Three-Month Treasury Bill Rate

Specification Change from Baseline	First Sample Period			Second Sample Period			Chow Test	
	ρ_{ols}	ρ_{mub}	90% CI	ρ_{ols}	ρ_{mub}	90% CI	F_ρ	P-Value
None	0.96	0.98	0.93-1.02	0.98	0.99	0.94-1.02	1.17	0.31-0.64
AR(2)	0.95	0.97	0.92-1.01	0.99	1.00	0.95-1.02	1.73	0.21-0.54
AR(4)	0.95	0.95	0.90-1.01	0.99	1.00	0.95-1.02	2.66	0.12-0.40
AR(8)	0.95	0.97	0.91-1.01	0.98	0.99	0.94-1.02	1.85	0.20-0.52
SD 1955	0.98	1.00	0.97-1.01	0.98	0.99	0.94-1.02	0.01	0.93-0.98
SD 1992, AR(2)	0.98	1.00	0.97-1.01	0.98	0.99	0.94-1.02	0.00	0.96-0.99

Notes: The first column shows the change in the specification from the baseline AR(6) model incorporating a constant (from Table 2). The baseline specification was estimated over the sample periods 1965:1–1978:9 and 1985:1–1998:9. AR(p) denotes an AR(p) model when a constant was used. "SD 1955" denotes a specification with the first sample period from 1955:1–1978:9. "SD 1992, AR(2)" denotes an AR(2) with second sample period from 1992:1–1998:9.

The choice of sample period has a more important effect. The baseline sample periods 1965:1–1978:9 and 1985:1–1998:9 were chosen to eliminate the large variability in interest rates during the late 1970s and early 1980s. With this volatile period eliminated, two samples of equal size were chosen (with 1998:9 being the last sample period available when this research was started). There is no compelling reason, other than equating statistical power in each sample, why the early and recent samples should be of equal size. The last two rows of the table show results from increasing the early sample period (by changing the beginning date to 1955:1) and decreasing the recent sample period (by changing the beginning date to 1992:1). Since the 1992–1998 sample period is very short, an AR(2) model was used for this specification. Evidently, the choice of the second period has little effect on the estimates of ρ, but the choice of first sample period does. Estimates of ρ are larger for both interest rates in the extended sample period 1955–1978 than in the 1965–1978 period. This increase should not be surprising given the behavior of interest rates over

the 1955–1978 period, where the dominant feature of the data is an increase in the "trend" level of interest rates. However, since this article's analysis focuses on the behavior of long rates as they are affected by expected future short rates, the question is whether investors in the late 1950s anticipated this trend rise in interest rates, as would be suggested by ex post fitted values from the univariate autoregression. Such prescience seems unlikely.

5. SUMMARY AND DISCUSSION

We have documented the increase in the variability of long-term interest rate changes during the 1985–1998 period relative to the 1965–1978 period. In contrast, the variability of short-term interest rates decreased in the later period. A possible explanation for this differential behavior is a change in the persistence of changes in short-term rates: expectations theories of the term structure imply that such shifts in persistence will have a large effect on the variability of changes in long-term rates but have little effect on the variability of changes in short rates. Point estimates of the largest autoregressive root for short rates show an increase in persistence that is large enough to explain the increased variability in long rates. However, the short-rate persistence parameter is imprecisely estimated, so that it is impossible to reach definitive conclusions based on this analysis. The lack of precision raises two issues: one related to statistical technique and one related to learning about changes in central bank policy.

The first issue concerns using the behavior of long rates to infer the persistence of the short-rate process. This is appropriate if long rates and short rates are connected by the present-value model. This procedure is used in Valkanov (1998), where the model's implied cointegration between long and short rates yields improved estimators for ρ. Valkanov then uses the improved estimator to overcome inference problems identified by Elliott (1998) in his critique of cointegration methods. Indeed, in a comment on a preliminary draft of this article, Valkanov (1999) uses his method to construct estimates of ρ together with 90 percent confidence intervals for the time periods 1962:1–1978:8 and 1983:1–1991:2 using data on the federal funds rate and ten-year Treasury bonds. He finds an estimate of ρ of 0.96 (with a 90 percent confidence interval of 0.93–0.98) in the early period and an estimate of 0.99 (with a 90 percent confidence interval of 0.99–1.00) in the later period (Valkanov 1999, Table 2c). His point estimates are essentially identical to the values of ρ_{mub} reported in our Table 2, but as expected from the use of a more efficient procedure, his confidence intervals are considerably narrower than the results presented in Table 2.

The large sampling uncertainty associated with estimates of the short-rate persistence suggests that the market will learn about changes in persistence very slowly from observing short-term interest rates. A central bank interested

in increasing the persistence of short-term interest rates (for the reason suggested in Woodford [1999], for example) would have to follow this policy for a considerable time to convince a market participant who relied only on econometric evidence that such a change had indeed taken place. For example, suppose that the federal funds process changed from one with a largest root of 0.96 to one with a largest root of 0.99, and after ten years in the new regime an econometrician tested the null hypothesis that $\rho = 0.96$ versus the alternative that $\rho > 0.96$ using a standard t-test at the 5 percent significance level. The econometrician would (correctly) reject this null only about 50 percent of the time. (That is, the power of the test using ten years of data is roughly 0.50.) Thus, it is likely the econometrician would have to observe the new federal funds process for quite some time before he concluded that the process had changed. This failure immediately to recognize policy shifts highlights the importance of other devices (institutional constraints, public statements, etc.) to more quickly convince a wary public that such shifts have occurred.

This article has presented econometric evidence suggesting that changes in the federal funds rate are more persistent now than they were in the 1960s and 1970s. Why did this change occur? We can offer but a few remarks on this important question. Here is one possible explanation. Suppose we decompose the funds rate into a real rate and an inflation component. If movements in the real rate are transitory, then the persistence in the funds rate will be driven by the inflation component. Therefore, an increase in the persistence of inflation possibly explains the increased persistence in the funds rate. This explanation, however, does not seem promising. For example, the values of ρ_{mub} computed using CPI inflation *fell* from 0.98 in the earlier sample period to 0.92 in the later period. As a result, inflation seems to have become *less* persistent, and this implies that some of the explanation must lie in the persistence of the real component of the funds rate. There is growing econometric evidence that the Federal Reserve's "reaction function" linking the federal funds rate to expected future inflation and real activity has been quite different under Chairmen Volker and Greenspan than under the previous three chairmen. For example, Clarida, Galí, and Gertler (1999) present evidence suggesting that the Federal Reserve responded more aggressively to expected future inflation after 1979 than in the previous two decades. Their evidence also suggests that the Federal Reserve more aggressively smoothed the funds rate in this latter period, consistent with the increased persistence found here. Changes in this reaction function undoubtedly contain the key to explaining the increased persistence in the federal funds rate process.

APPENDIX

A. Computing Prob-Values for the Chow Test Statistic for the Largest Autoregressive Root

This Appendix describes the method used to compute the Prob-values for tests of changes in the largest autoregressive root of a univariate autoregression. The specification is the AR(p) autoregression

$$x_t = \mu + u_t$$

with

$$u_t = \sum_{i=1}^{p} \phi_i u_{t-1} + \varepsilon_t,$$

where x_t denotes the level of the interest rate, μ is a constant denoting the average level of the process in the stationary model, and u_t is a stochastic term. The u_t process can be rewritten as

$$u_t = \rho u_{t-1} + \sum_{i=1}^{p-1} \pi_i (u_{t-i} - u_{t-i-1}) + \varepsilon_t,$$

where $\rho = \sum_{i=1}^{p} \phi_i$ and $\pi_i = -\sum_{j=i+1}^{p} \phi_j$. The parameter ρ is thus the sum of the AR coefficients. When one root of the AR polynomial $1 - \sum_{i=1}^{p} \phi_i z^i$ is close to 1 and all of the other roots are larger than 1, then ρ is also approximately equal to the inverse of the root closest to unity. In this case ρ is usually called the "largest" root because its inverse is the largest eigenvalue of the companion matrix of the model VAR(1) representation.

We study the behavior of statistics in a setting where ρ is modeled as close to 1.0, written as

$$\rho_T = 1 + \frac{c}{T}.$$

The artificial dependence of ρ on the sample size T facilitates the analysis of continuous asymptotic limits as $T \to \infty$.[5] To simplify notation, we will present the AR(1) model, so that $\pi_i = 0$, for $i = 1, ..., p-1$. For the test statistics

[5] To see this, contrast the discontinuous results

$$\lim_{T \to \infty} \rho^T = \left\{ \begin{array}{c} 0 \text{ when } |\rho| < 1 \\ 1 \text{ when } \rho = 1 \\ \infty \text{ when } \rho > 1 \end{array} \right\}$$

with the continuous result

$$\lim_{T \to \infty} (\rho_T)^T = e^c \text{ when } \rho_T = 1 + c/T.$$

used in this article, the inclusion of extra lags has no effect on the limiting distribution, and in this sense the presentation here is without loss of generality. Following the discussion of the limiting distribution of the Chow test statistic, Appendix A2 discusses the numerical procedure used to compute the Prob-values shown in Tables 2 and 3.

A1 Asymptotic Distribution in the AR(1) Model

Assume

$$u_t = \rho_T u_{t-1} + \varepsilon_t,$$

where u_0 is a finite fixed constant, $t = 1, ..., T$, and ε_t is a martingale difference sequence with $E(\varepsilon_t^2 \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) = 1$, and with $\sup_t E\varepsilon_t^4 < \infty$, where $\rho_T = 1 + \frac{c}{T}$.

Let $\widehat{\rho}_1$ denote the OLS estimator of ρ constructed from the regression of x_t onto $(1, x_{t-1})$ using the early sample period $t = 1, ..., T_1$, and let $\widehat{\rho}_2$ denote the corresponding estimator constructed using the later sample period $t = T_2, ..., T$. Assume

$$\lim_{T\to\infty} \frac{T_1}{T} = \tau_1$$

and

$$\lim_{T\to\infty} \frac{T_2}{T} = \tau_2$$

with $0 < \tau_1 < \tau_2 < 1$. Denote the sample means by

$$\overline{x}_{1,T} = \frac{1}{T_1} \sum_{t=1}^{T_1} x_t$$

$$\overline{x}_{2,T} = \frac{1}{T - T_2 + 1} \sum_{t=T_2}^{T} x_t$$

and the demeaned series by

$$x_{1,t}^{\mu} = x_t - \overline{x}_{1,T}$$

$$x_{2,t}^{\mu} = x_t - \overline{x}_{2,T}.$$

The limiting behavior of these series is related to the behavior of the diffusion process $J_c(s)$, generated by

$$dJ_c(s) = cJ_c(s)ds + dW(s)$$

for $0 \le s \le 1$, where $W(s)$ is a standard Wiener process. In particular,

$$\frac{1}{\sqrt{T}} x_{1,[sT]}^{\mu} \Rightarrow J_c(s) - \tau_1^{-1} \int_0^{\tau_1} J_c(r)dr \equiv J_{1,c}^{\mu}(s) \text{ for } 0 < s \le \tau_1$$

$$\frac{1}{\sqrt{T}} x^{\mu}_{2,[sT]} \Rightarrow J_c(s) - (1-\tau_2)^{-1} \int_{\tau_2}^{1} J_c(r)dr \equiv J^{\mu}_{2,c}(s) \text{ for } \tau_2 \leq s < 1.$$

The Chow F-statistic for testing $H_0 : \rho_1 = \rho_2$ is

$$F = \frac{(\widehat{\rho}_1 - \widehat{\rho}_2)^2}{[\sum_{t=1}^{T_1}(x^{\mu}_{1,t-1})^2]^{-1} + [\sum_{t=T_2}^{T}(x^{\mu}_{2,t-1})^2]^{-1}}.$$

The limiting behavior follows from considering the terms

$$U_{1,T} \equiv \frac{1}{T}\sum_{t\equiv 1}^{T_1} \varepsilon_t x^{\mu}_{1,t-1} \Rightarrow \int_0^{\tau_1} J^{\mu}_{1,c}(s)dW(s) \equiv U_1$$

$$U_{2,T} \equiv \frac{1}{T}\sum_{t\equiv T_2}^{T} \varepsilon_t x^{\mu}_{2,t-1} \Rightarrow \int_{\tau_2}^{1} J^{\mu}_{2,c}(s)dW(s) \equiv U_2$$

$$V_{1,T} \equiv \frac{1}{T^2}\sum_{t\equiv 1}^{T_1} (x^{\mu}_{1,t-1})^2 \Rightarrow \int_0^{\tau_1} (J^{\mu}_{1,c}(s))^2 ds \equiv V_1$$

$$V_{2,T} \equiv \frac{1}{T^2}\sum_{t\equiv T_2}^{T} (x^{\mu}_{2,t-1})^2 \Rightarrow \int_{\tau_2}^{1} (J^{\mu}_{2,c}(s))^2 ds \equiv V_2.$$

Defining

$$\gamma_{1,T} = T(\widehat{\rho}_1 - \rho) \text{ and } \gamma_{2,T} = T(\widehat{\rho}_2 - \rho),$$

the F can be written as

$$F = \frac{(\gamma_{1,T} - \gamma_{2,T})^2}{V_{1,T}^{-1} + V_{2,T}^{-1}}.$$

Since

$$\gamma_{1,T} = \frac{\frac{1}{T}\sum_{t=1}^{T_1} \varepsilon_t x^{\mu}_{1,t-1}}{V_{1,T}} = \frac{U_{1,T}}{V_{1,T}} \Rightarrow \frac{U_1}{V_1} = \gamma_1$$

and

$$\gamma_{2,T} = \frac{\frac{1}{T}\sum_{t=T_2}^{T} \varepsilon_t x^{\mu}_{2,t-1}}{V_{2,T}} = \frac{U_{2,T}}{V_{2,T}} \Rightarrow \frac{U_2}{V_2} = \gamma_2$$

by the continuous mapping theorem, then

$$F \Rightarrow \frac{(\gamma_1 - \gamma_2)^2}{V_1^{-1} + V_2^{-1}},$$

which provides a representation for the limiting distribution of F in terms of functionals of the diffusions $J_c(s)$.

A2 Approximating Prob-values

The limiting distribution of F is seen to depend on three parameters τ_1, τ_2 (through the limits in the integrals), and the value of c (through the mean reversion in the diffusion process J_c). Quantiles of the limiting distribution (and hence Prob-values for the test statistic) can be approximated by repeated simulations of F using a large sample size and for fixed values of τ_1, τ_2, and c, and ε_t chosen as $Niid(0, 1)$ random variables. The Prob-values reported in the article resulted from 10,000 replications from a sample size of 500. The parameters τ_1 and τ_2 were chosen as T_1/T and T_2/T, where T_1 denotes the first break point and T_2 denotes the second break point. The distribution also depends on c, which governs how close ρ is to unity. Unfortunately, this parameter cannot be consistently estimated. (Equivalently, in finite samples the distribution of F depends on ρ, and small changes in ρ—like those associated with sampling error—lead to large changes in the quantiles of this distribution.) Thus, selecting the correct distribution of F requires knowledge of c (equivalently, ρ). Since c is unknown, the distribution is computed for a range of values in $-25 \leq c \leq 10$ and the resulting minimum and maximum Prob-value over all of the values of c is reported in the table. Viewing c as unknown, classical approaches (which must hold for all values of the "nuisance parameter" c) would use the upper Prob-value. The lower bound gives the smallest Prob-value that would be obtained if c were known.

REFERENCES

Cavanagh, C. L. "Roots Local to Unity." Manuscript. Department of Economics, Harvard University, 1985.

Christiano, Lawrence J., and Martin Eichenbaum. "Unit Roots in Real GNP—Do We Know, and Do We Care?" *Carnegie-Rochester Conference Series on Public Policy*, vol. 32 (1990), pp. 7–32.

Clarida, Richard, Jordi Galí, and Mark Gertler. "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *Quarterly Journal of Economics*, vol. 115 (February 2000), pp. 147–80.

Deaton, Angus S. "Life-Cycle Models of Consumption: Is the Evidence Consistent with the Theory?" in Truman F. Bewley, ed., *Advances in Econometrics*, Vol II. Amsterdam: North Holland, 1987, pp. 121–48.

Elliott, Graham. "On the Robustness of Cointegration Methods when Regressors Almost Have Unit Roots," *Econometrica*, vol. 66 (January 1998), pp. 149–58.

Goodfriend, Marvin. "Information-Aggregation Bias," *American Economic Review*, vol. 82 (June 1992), pp. 508–19.

Mehra, Yash P. "Monetary Policy and Long-Term Interest Rates," Federal Reserve Bank of Richmond *Economic Quarterly*, vol. 82 (Summer 1996), pp. 27–29.

Phillips, P. C. B. "Toward a Unified Asymptotic Theory for Autoregression," *Biometrika*, vol. 74 (1987), pp. 535–47.

Quah, Danny. "The Relative Importance of Permanent and Transitory Components: Identification and Some Theoretical Bounds," *Econometrica*, vol. 60 (January 1992), pp. 107–18.

Shiller, Robert J. "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?" *American Economic Review*, vol. 71 (June 1981), pp. 421–36.

_____, John Y. Campbell, and Kermit L. Schoenholitz. "Forward Rates and Future Policy: Interpreting the Term Structure of Interest Rates," *Brookings Papers on Economic Activity*, 1:1983, pp. 173–223.

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.

Valkanov, Rossen. "Notes on Watson's 'Explaining the Increased Variability in Long-Term Interest Rates.' " Research Memorandum. Princeton University, 1999.

_____. "The Term Structure with Highly Persistent Interest Rates." Manuscript. Princeton University, 1998.

West, Kenneth D. "Dividend Innovations and Stock Price Variability," *Econometrica*, vol. 56 (January 1988), pp. 37–61.

Woodford, Michael. "Policy Inertia." Manuscript. Princeton University, 1999.