Spring

1987

Federal Reserve Bank
of San Francisco

Economic Review

Number

2

Opinions expressed in the Economic Review do not necessarily reflect the views of the management of the Federal Reserve Bank of San Francisco, or of the Board of Governors of the Federal Reserve System.

The Federal Reserve Bank of San Francisco's Economic Review is published quarterly by the Bank's Research and Public Information Department under the supervision of Jack H. Beebe, Senior Vice President and Director of Research. The publication is edited by Gregory J. Tong, with the assistance of Karen Rusk (editorial) and William Rosenthal (graphics).

For free copies of this and other Federal Reserve publications, write or phone the Public Information Department, Federal Reserve Bank of San Francisco, P.O. Box 7702, San Francisco, California 94120. Phone (415) 974-3234.

Major Trends in the U.S. Financial System: Implications and Issues

Robert T. Parry*

The major trends shaping the current evolution of the U.S. financial system conflict with an outmoded legal and regulatory framework. Reform of that framework is needed to promote an efficient and stable financial system. Specifically, reforms are needed in the deposit insurance system, bank powers, and the large-dollar, electronic payments system .

This paper was originally presented at a Federal Reserve System Management Conference held April 1987.

Shaped by the interaction of economic, technological, legal, and regulatory forces, the U.S. financial system is undergoing signficant change. During the next five to ten years, it increasingly will be characterized by:

- reliance on primary securities markets, with a diminishing role for traditional bank-provided intermediation;
- institutional realignment of functions in the provision of financial services, including clearing and settlement;
- expanded access to the payments system; and
- geographic integration, including internationalization of financial activity, with around-the-clock trading and settlement.

The present legal and regulatory structure often conflicts with fundamental economic and technological forces. Moreover, the piecemeal efforts to resolve these conflicts and to accommodate market forces have resulted in several undesirable consequences. First, financial change is occurring through the exploitation of legal and regulatory loopholes rather than in a manner that ensures the evolution of an efficient financial system. The proliferation of new instruments, the transfer of traditional banking activity to nonbanks, and the staggering volume of daily payments activity, for example, may be as much a result of efforts to avoid regulation as a response to fundamental economic needs.

Second, although partial integration of financial activities and of financial and commercial activities is occurring, the important issues of how to reform the federal safety net and how far to extend its coverage are not being resolved. Third, as activity shifts to international financial centers and less-regulated nonbank firms, domestic banking firms are left with a diminishing proportion of overall financial activity.

The legal and regulatory framework should be reformed to accommodate market-driven forces for change. However, such reform also must be consistent with the goal of preserving financial stability. These criteria imply that federal supervision, regulation, and protection of the financial system should be structured and conducted in a way that promotes stability while limiting the perverse incentives for risk-taking and the possibility of large government expenditures that government intervention can create. In this paper, I presume that limited government guarantees, directed at payments balances and savings balances held by depositories, are

* President, Federal Reserve Bank of San Francisco. Staff contributors are Barbara Bennett, Economist, Michael Keeley, Senior Economist, Randall Pozdena, Assistant Vice President, and Jack Beebe, Senior Vice President and Director of Research.

desirable. The optimal extent and structure of such guarantees, however, are issues that are addressed, but not resolved, in the paper.

The purpose of this paper is to provide a conceptual framework for both understanding the changes occurring in the financial system and analyzing the policy implications of those changes. The paper is organized as follows: in Section I, the economic and regulatory forces that are driving the evolution of the financial system are described. Then the major emerging trends are outlined in Section II. Finally, the policy implications of these changes are considered in Section III. The paper concludes in Section IV that policymakers should focus on reforming three key areas: the federal safety net, bank powers, and the payments system.

I. Forces for Change

The changes that are likely to take place in the U.S. financial system are the result of the interaction of basic economic forces with regulatory and legal constraints. In some cases, the economic forces will overwhelm these constraints; in other cases, regulatory and legal forces will dominate.

Economic Forces

The following primary economic forces appear to be at work today. First, electronic information processing is reducing the cost of gathering, managing, and transmitting the data required to produce financial services. This trend not only makes it possible to provide certain traditional financial services at reduced cost but makes new financial services feasible, thereby altering ways of raising and investing funds. Moreover, the effect of technological change is amplified by the rising demand for time-saving innovations, such as certain integrated financial services, and devices that improve access to retail funds balances, such as ATMs and point-of-sale terminals.

Second, the growth in wealth and the level of commercial activity worldwide is increasing the number, size, and complexity of transactions in financial markets. The changing nature of transactions, in turn, is stimulating the demand for sophisticated financial services.

Third, greater volatility in interest rates, exchange rates, and asset prices is expanding the demand for ways to manage risk. Not only is it increasing reliance on specific risk-management products, as exemplified by the growth of options and futures markets, but also the complexity of traditional financial instruments. Moreover, widespread loan losses in recent years have heightened the need for risk management and diversification within institutions.

Legal and Regulatory Forces

The present financial system has in place extensive legal and regulatory structures to promote the stability of the banking industry and the payments mechanism and to facilitate the conduct of monetary policy. To the extent that these structures conflict with private economic incentives, the private market will attempt to take advantage of or avoid them. As a result, our system of laws and regulations itself is a force for change in financial markets.

The first, and perhaps most important, example is the effect of the federal safety net on bank behavior. Underpriced deposit protection gives banks and thrifts incentives to reduce equity capital and to expand the scope of insurance coverage. Similarly, nondepository institutions may seek to own depository firms in order to raise insured funds.

A second force for change is reserve requirements. It has been argued that reserve requirements are needed for monetary policy purposes. Yet the requirement to hold noninterest-earning reserves against transactions deposits, together with the prohibition on the payment of explicit interest on demand deposits, provides incentives to create functionally similar instruments, such as overnight RPs, that arbitrarily increase activity in the payments system.

A third regulatory force for change is the system of legal restrictions on the ownership and powers of banking firms, which is particularly cumbersome at a time when nonbanking firms are expanding rapidly into the provision of many banking services. Geographic restrictions facing banking firms also are affecting industry structure by shifting activity away from traditional banks.

Fourth, payments system policies affect not only the level of risk in the payments system but also the structure of the financial services industry. The absence of real-time settlement and mechanisms to price intraday credit extended by the Federal Reserve arguably contribute to the level of payments system risk. The exclusion of nondepository firms from access to the payments system may be inducing commercial enterprises to buy depository institutions in order to gain the access they desire.

Finally, regulatory, tax, and other policies vary across countries (and in some instances, states) even for the same class of institution. These international (and interstate) differences affect not only the locus of financial activity, but also the ability of any one government or agency to pursue policies that ensure prudent practices in the financial system. However, the recently proposed agreement between U.S. and British banking authorities is an important step in the attempt to internationalize supervision and regulation.

II. Emerging Trends

Several key trends are emerging from the economic, technological, regulatory, and legal forces affecting the U.S. financial system. These include trends toward direct placement and securitization, functional realignment in the provision of financial services, expanded access to the payments system, and geographic integration of financial services and markets. They are described below.

Direct Placement and Securitization

Increasingly, borrowers are placing debt securities directly with investors and relying correspondingly less on traditional financial intermediaries — most notably, commercial banks — as a source of funds. This rise in direct placement is the outcome of a number of underlying economic forces, as well as constraints, imposed by the current financial regulatory framework. Growth in the sheer size of transactions, together with the declining costs of transmitting credit information and effecting transactions, have made direct placement cost-effective for an increasing number of transactions.

Historically, banks have enjoyed cost-advantages in evaluating a given borrower's creditworthiness and in taking on interest rate, liquidity, and credit risks. As a result, financial intermediation through the use of two instruments — a deposit liability and a loan asset — with the bank as a party to each has been more economical than direct dealings between borrowers and investors. Today, however, banks' cost-advantages in executing intermediation functions are diminishing. The declining cost of technology permits the sale of credit information by various rating services and, in one sense, even by banks themselves through the issuance of standby letters of credit. The effects of declining information costs are especially evident in the shrinking share of large and middle-market corporate borrowers that continue to rely on banks as their primary source of funds. More of these borrowers now obtain Moody's or Standard and Poors' ratings in conjunction with standby letter of credit backing to enable them to raise funds directly in the commercial paper and bond markets.

In addition, the rapidly growing depth and liquidity of futures and options markets enable investors to manage interest-rate risk inexpensively and directly, thereby reducing the relative cost advantage that banks traditionally have enjoyed. Similarly, secondary markets for a broadening array of primary securities provide a growing source of liquidity which, in the past, only banks were able to provide economically. Finally, pension funds, tax-sheltered savings plans, and money market and bond mutual funds now offer numerous channels to satisfy the diverse denomination requirements of borrowers and investors. In fact, money market mutual funds continue to hold over $250 billion in assets even though banks no longer are subject to interest-rate ceilings on most deposit products. As a result of these developments, financial market participants can now purchase — in relatively small denominations — their desired mix of liquidity, credit, and interest-rate risks directly, without having to turn to the banking system in the traditional sense.

In addition to the economic forces favoring direct placement, regulatory constraints continue to make intermediation by depositories less attractive than otherwise. Reserve requirements, in particular, raise the cost of attracting reservable funds. Moreover, the decision of bank regulators in recent years to require banks to maintain higher minimum (book) capital-to-asset ratios probably is raising the effective cost of holding assets in portfolio. With underpriced deposit insurance, it generally has been more profitable for banks to raise insured deposits than to raise capital to fund loans.

As a result of these economic forces and regulatory constraints, banks increasingly have been shifting to so-called off-balance-sheet activities. In recent years, for example, banks have offered standby letters of credit (for a fee) to back debt securities of prime issuers rather than funding loans to those issuers. By assuming these contingent liabilities, banks can increase their effective leverage without actually violating formal capital requirements. In effect, the implicit government protection of large institutions encourages their entry into the growing market for financial guarantees. Approximately 15 percent of all commercial paper and 29 percent of newly issued municipal bonds now have some form of financial guarantee as backing.[1]

Because banks retain some comparative advantages in evaluating the creditworthiness of smaller businesses and households as well as in servicing debt obligations, they continue to originate loans but are selling and then servicing an increasing number of them. Similarly, banks are "securitizing" loans (that is, pooling and using them as security for marketable debt instruments that are sold outright to investors). In addition to residential mortgage loans, which were first securitized in 1968, commercial mortgages, automobile finance loans, and credit card receivables are also being securitized now. For many of these products, the rate of growth of the derivative security far exceeds the growth rate of the underlying asset. For example, residential mortgage backed securities grew by 155 percent over the past five years, while real estate loans in bank and thrift portfolios grew by only 24 percent.[2]

Clearly, the trend is for banks to take on the role of broker, or even underwriter, to facilitate transactions in the primary market. However, banks still will function as traditional intermediaries, as most will continue to hold loans in portfolio, particularly loans to borrowers whose creditworthiness is relatively costly for the market to evaluate or whose funding needs are not standard.

Functional Realignment

Economic forces such as the demand for greater convenience in financial services, the declining cost of effecting transactions, and the growth of securitization and direct placement are causing a breakdown in institutional specialization. Commercial banks, thrift institutions, securities firms, insurance companies, and other types of financial and nonfinancial companies increasingly are offering products that overlap their traditional markets. Although these developments do not necessarily portend full-scale integration of financial service firms, they do suggest that the old institutional boundaries governing firms' activities are breaking down and that a realignment of the types of services each firm chooses to provide is taking place.

For wholesale commercial banks and investment banks serving the needs of the corporate sector, this process of realignment is especially apparent. For commercial banks, the push towards investment banking is a logical extension of their expertise in lending as their corporate borrowers rely more and more on direct placement. By the same token, investment banks believe that their ability to offer certain commercial banking services would be advantageous. For example, investment banks want to be able to offer payments services and to settle transactions directly because doing so themselves is more efficient and profitable than obtaining the same services from commercial banks.[3]

So far, nonbank firms have been more successful at circumventing barriers than have commercial banks simply because they are regulated less extensively. Technically, regulatory and legal restrictions on the ownership of commercial banks prevent nonbank firms from offering banking services. However, through such innovations as checkable money market mutual funds and cash management-type sweep accounts, nonbank firms now offer services that are functionally similar to commercial banks' payments and deposit services. Moreover, with the expansion of thrift institutions' lending, payments, and deposit-taking powers, the ownership of nonbank firms (which is not restricted in the same way as commercial bank ownership) confers many banking powers. Likewise, the innovation of nonbank banks, which skirt the legal definition of a commercial bank by offering only commercial loans or demand deposit services but not both, will enable nonbank firms such as Merrill Lynch to offer banking services.

While nondepository firms are now able to offer virtually the full range of banking services — albeit, less efficiently than through outright ownership of commercial banks — banks are trying to broaden their nonbanking activities. To a certain extent, regulators are accommodating these pressures. Bank regulators have expanded permissible activities to include credit-related insurance, discount brokerage, (limited) securities underwriting, data processing, financial planning, and investment advisory services. Moreover, regulators now sanction bank holding companies' purchases of failing thrift institutions, thereby expanding the opportunities for those banking organizations.

For the most part, regulatory accommodation still does not allow banks the degree of freedom that nonbank firms have. In particular, commercial banks are prevented from underwriting the vast majority of municipal debt and all domestic corporate debt and equities. Where they are not as constrained by Glass-Steagall restrictions, large commercial banks underwrite a significant and growing proportion of securities in international capital markets, including interest rate and currency swaps. (There are, however, limitations on the amount of corporate debt and equity underwriting they can do even in foreign markets.[4])

Expanding Access to Payments System

Many of the forces that are encouraging realignment in the provision of financial services also are motivating nonbanks' desire for access to the payments system. In particular, the increasing integration of payments and securities activities and the trend towards direct placement are making direct access to the payments system more valuable than in the past. Also, the high and volatile interest rates of a few years ago induced corporations and households to invest in more sophisticated cash management technology, which they continue to use even in the current lower interest rate environment. The use of such technology makes direct access to the payments system for the purposes of consolidating and investing idle balances especially attractive. Such forces are behind brokerages' cash management accounts and the establishment of nonbank banks by brokerage firms.

Combined with these economic forces are several regulatory constraints that encourage the use of alternatives to bank-provided payments balances. Noninterest-earning reserves and the prohibition on the payment of explicit interest on demand deposits raise the effective cost of using demand deposits to settle transactions. As a result, corporations in particular employ cash management techniques that minimize such balances. Their actions, especially those related to the overnight RP market, undoubtedly are part of the explanation for the extraordinary volume of transactions over Fedwire, the Federal Reserve's electronic funds transfer network.

Nonbank firms traditionally have been denied direct access to the payments system in general, and to Fedwire in particular, because of concerns about increased payments system risk. However, as noted above, ownership of thrifts and nonbank banks enables nonbank firms to circumvent restrictions on access and may, in time, render the current legal framework governing access to the payments system obsolete.

Geographic Integration

The growth of international trade and commerce, the integration of financial markets and payments media, and the declining cost of information technology appear to be increasing the optimal geographic scope of firms in banking and finance. As a result, there is a trend towards internationalization of capital markets and interstate provision of domestic financial services. Domestic firms of stature can now raise funds economically in the rapidly growing euromarkets. A large California utility, for example, has at times raised a significant proportion of its longer-term funding in the euronote and bond markets even though its operations are confined largely to domestic markets.

Commercial and investment banking firms are expanding their international activities not only to "follow their customers" but also to take advantage of international regulatory discrepancies. They have, for example, been attracted to the London and Tokyo markets, which recently have loosened restrictions on the activities of market participants. At the same time, domestic restrictions also are pushing U.S. financial institutions overseas. Restrictions on commercial banks' securities underwriting activities as well as reserve requirements and deposit-rate regulations, for example, have induced U.S. banks to shift business to international markets where they can avoid domestic regulations.[5]

The trend towards interstate provision of domestic financial services is even more pronounced.[6] Banks are seeking to establish regional and even national deposit-taking networks to broaden and diversify their core deposit, financial services, and lending bases and to provide customers engaged in interstate transactions with improved access to the payments system. Regulations already have accommodated these forces to a large extent, although perhaps not in the most economical way.

Through holding company subsidiaries, banks and thrifts now can perform virtually all banking functions across state lines except deposit-gathering. (With the advent of brokered deposits and nonbank banks, they are not fully constrained even in this last area.) Moreover, individual states are now accommodating interstate entry. Thirty-seven states have passed legislation permitting entry by banking firms located out of state. Eighteen of those states permit, or will soon permit, entry by banks headquartered anywhere in the country.

III. Implications and Issues

These developments raise a number of public policy concerns. First, the present approach to regulation of the financial system encourages an inefficient use of resources. For example, resources are devoted to discovering and exploiting loopholes in the current legal and regulatory system. More importantly, the result of this process is a structurally inefficient financial industry that is characterized by a proliferation of new instruments, transfer of traditional banking activity to nonbanks, and payments volumes that are excessive in relation to economic activity.

Second, without deposit insurance reform, the expansion of financial activity of banks or the integration of financial and commercial activities may lead to an undesirable propagation of the deposit insurance subsidy. One concern, for example, is that a stressed nonbank affiliate might draw financial support from the bank, endanger the bank, and indirectly be supported by the deposit insurance fund.

Third, the growth of international financial centers and of unregulated firms' involvement in the provision of financial services implies diminished federal supervisory leverage over financial activity that may be essential to financial stability. Diminished supervisory control is particularly troublesome in light of concern about the potential for undesired or unintended *de facto* extension of the federal safety net.

The current legal, regulatory, deposit insurance, and payments frameworks are inadequate for addressing these policy concerns. Reform is needed to preserve financial stability and to accommodate a changing financial environment. However, such reform must balance the benefits from enhancing stability against the costs. For example, stability could be enhanced (in the short run) if deposit insurance were extended to every financial firm. Absent deposit insurance reform, however, such an approach would distort risk-taking decisions (or require a vast expenditure of supervisory resources to prevent the distortion).

Since it is not feasible or desirable to insure every firm or activity, we must decide what truly needs to be protected. Although the extent of insurance coverage is a subject of intense debate, nearly all agree that protection of transactions balances (whether held at commercial banks or at nonbanks) is essential.[7] Also, because many observers are concerned that a serious contraction in the availability of intermediated credit from depository institutions could have destabilizing consequences, there also is a view that a significant part of all nontransactions balances needs to be protected as well.[8] At the same time, however, there is the concern that this protection not extend too far. Clearly, we must not protect the owners of credit-granting intermediaries from the consequences of their decisions lest we run the risk of excessive risk-taking on their part.

Although the question of what ought to be protected has no simple answer, most observers conclude that both the payment and credit intermediation functions of depositories need partial, if not fairly extensive, protection. Many of the issues discussed below regarding the structure of the deposit insurance system, the boundaries of bank powers, and the operation of the payments system are predicated on this conclusion. The conclusion itself, however, is open to debate.

Deposit Insurance and the Federal Safety Net

Our present deposit insurance system actually has performed remarkably well over the last fifty years. Although there have been runs on individual banks, spillover effects have been limited and there have not been any banking panics at federally insured institutions. In addition, payouts from insurance funds were very modest prior to the 1970s.

More recently, however, many observers have begun to question the viability of the system in the wake of a record number of bank failures, the large foreign debt exposures of the money center banks, and the well-publicized problems of the FSLIC. One can argue that some of the recent problems stem from an implicit (and at times, explicit) extension of the federal safety net well beyond the stated coverage of deposit insurance. One might even say that the safety net has been spread so thinly it may soon tear. Moreover, because the current system relies so heavily on supervision and regulation, it has become increasingly unable to accommodate the market forces and trends enumerated above.

The Status Quo

It is now widely recognized that the current deposit insurance system introduces a moral hazard; that is, it gives insured institutions an incentive to take on excessive risk. The combination of flat-rate premia unrelated to risk, possible coverage of all deposit and nondeposit liabilities (at least at large banks), and a willingness to let insolvent banks and thrifts continue to operate has seriously undermined the discipline on risk-taking that would otherwise be imposed by the market.[9]

As a result, regulation and supervision bear the main burden of limiting risk-taking. However, should the implied protection of deposit insurance continue to expand, the prospects of containing bank risks with supervision and regulation would dim and leave the government to underwrite risks for larger and larger segments of the economy. Thus, reform of the deposit insurance system is central to and a prerequisite for financial reform. Indeed, it may be needed just to deal with the current economic environment, as exemplified by the problems of the savings and loan industry and its insurance fund.

Approaches to Reform

There have been many proposals for reforming the deposit insurance system. Some involve restricting the explicit or implicit scope of deposit insurance coverage while others seek to "reprice" insurance to reduce the moral hazard problem. Below, the pros and cons of various alternatives are discussed.

Reducing the Scope of Deposit Insurance

Perhaps one of the oldest reform proposals dates back to Henry Simons' 1948 proposal for 100 percent reserve banking as modified by Milton Friedman in 1959 to include the payment of interest on reserves.[10] This idea, which in essence has been revived by Robert Litan and John Kareken among others,[11] would turn banks into institutions similar to money market mutual funds — that is, banks' liabilities would be used to fund only safe assets, such as short-term government securities, cash, and reserve balances at the Federal Reserve.

If banks were required to back their liabilities with only "perfectly safe" assets, they could not fail. Moreover, no restrictions on the ownership of such "eunuch" banks would be necessary since there would be no opportunity for the bank to support failing nonbank affiliates. (The bank's deposit liabilities would be used to fund only safe assets and not to fund any form of credit to affiliates, including intraday payments credit.) Under these conditions, which imply complete legal and economic separation of the bank from nonbank affiliates, the failure of a nonbank affiliate could not impair the bank. Of course, if banks held assets that were "fairly," but not perfectly, safe and were allowed to extend credit to subsidiaries, or to lend their "good name" to the subsidiaries in a way that implied legal liability, then the problems of controlling the risks undertaken by a diversified conglomerate would arise.

Implicit in this "safe assets" approach is the notion that deposit insurance should protect only the payments system or payments-related balances. In fact, under this proposal, meaningful credit intermediation would take place only in uninsured financial institutions or subsidiaries similar to current-day banks in most respects except for their inability to offer insured pure transactions accounts. Although uninsured intermediaries presumably would take on fairly conservative risk postures, they probably would still use short-term liabilities to fund risky, longer-term loans to some degree and thus could be subject to problems with depositor runs. These runs have the potential for destabilizing the credit system. Thus, although the safe assets proposal might provide adequate protection for the payments function of depositories, it would offer no protection for credit intermediation.

Another approach to limiting the scope of deposit insurance focuses on explicitly restricting the payouts made to depositors to encourage depositor surveillance of depositories' risk-taking. Traditionally, the insurance system has restricted payouts by fully insuring each deposit only up to some maximum amount. But other approaches could be taken, such as also explicitly insuring a given percentage of deposits above the maximum amount. Of course, to be meaningful, maximum insurance coverage would have to be enforced strictly and uniformly for all failed banks *and* "assisted" mergers that expose the insuring agency to losses.

The FDIC's "modified payout" proposal would work well in this context if it were applied uniformly to "purchases and assumptions" as well as to "payouts." If this were done, deposits not fully insured would be subject to an immediate markdown and might never be fully repaid. (Under the experimental modified payout plan, uninsured depositors of some *closed* banks received only a prorated portion of the estimated value of the failed banks' assets immediately. Uninsured depositors of such failed banks could receive additional payments if, upon disposal of the assets, the realized value exceeded the FDIC's estimate. The actual plan, however, did not shift losses to depositors if a "failed" bank were handled through a purchase and assumption.)[12]

Increasing depositor surveillance through these avenues would reduce *ex ante* risk-taking because uninsured depositors would require premium rates for the riskier institutions and the threat of a run by depositors would induce institutions to operate prudently. Over time, the level of supervision and regulation could be scaled back, although there would be a need for *more* public information about the conditions of banks. Both of these developments would accommodate the natural evolution of the financial system. However, the proposal provides little protection against runs by uninsured depositors. To the extent that one believes that runs by uninsured depositors are potentially destabilizing to the financial system, as apparently was the view of regulators in the episode involving the failure of Continental Illinois in 1984[13], proposals of this nature do not offer sufficient protection.

In sum, limiting the scope of deposit insurance, either by limiting the functions of insured institutions or by limiting insured deposit coverage, could reduce or eliminate the moral hazard implicit in the current deposit insurance system. However, these proposals would increase the potential for credit runs that could destabilize the financial system. An alternative approach to reform is to maintain fairly broad insurance coverage of the payments *and* credit functions of financial intermediaries while "repricing" that coverage to reduce the moral hazard.

Repricing Deposit Insurance

The most obvious way to reprice deposit insurance is to charge an insurance premium that rises with the *ex ante* risk of the insured institution's portfolio. This is a sound concept because it would penalize bank equityholders for excessive risk-taking and thus would internalize the costs of risk-taking along with the benefits.

In practice, however, this proposal could prove extremely difficult to implement because it would require charging an insurance premium based on examiners' assessments of the *ex ante* market values and risks of a bank's portfolio of assets, many of which are not traded or readily marketable, as well as judging the risks and potential profitabilities of its non-portfolio activities. Moreover, to have a significant impact on *ex ante* risk-taking, examiners' risk assessments would have to look well to the future, and premia might have to be adjusted fairly dramatically on the basis of subjective risk assessments. (The FDIC's legislative proposal to double the annual premium to one-sixth of a percent of deposits for banks in the high-risk category would not be sufficient to deter risk-taking.)

A second method of internalizing risk requires that insured institutions be closed before the market value of their equity could fall below zero. If this could be accomplished without error, a closed institution's assets necessarily would be sufficient to discharge its liabilities at the time of liquidation. As a result, failed (that is, closed) institutions would not impose losses on the insurance fund. Instead, bank equityholders would bear the full costs and benefits of their decisions and would have no incentive to take excessive risks.

Moreover, as long as depositors were confident that regulators would be successful in closing banks before the market value of equity became negative, and thus assure them of protection from losses, they would not run on a "troubled" bank. In this manner, it would be possible, in concept, to protect deposits and prevent runs while simultaneously confining risk to bank equityholders.

To be effective, however, this approach would require increases in both the scope and frequency of federal supervision of insured institutions to monitor the market values of their equity closely. One major practical difficulty in increasing supervision lies in assigning accurate market values to non-traded assets and liabilities. Such valuation might be even more difficult if banks took on added powers or acquired commercial firms. Another difficulty is the lack of legal authority for the insuring agency to require chartering agencies promptly to close institutions deemed insolvent on the basis of a market value assessment of equity.

As a result, any practical implementation of this approach would have to allow for errors in closure. If depositors believed a bank might be closed too late, for example, they would run unless they could be assured that losses would be covered by a third party, such as subordinated debt holders and/or the deposit insurance fund. (To be effective, subordinated debt should be perpetual and subordinated to both bank deposits and the insurance fund.)

There are other ways of accommodating errors in assessing market values. One would be to give regulators the authority to err on the safe side either by closing a bank that might still have a positive market value of equity or by requiring the bank to increase its equity to reduce the risk of *ex post* uninsured depositor or insurance fund losses. Although politically impractical today, yet another method would be to hold bank equityholders liable for losses exceeding their original capital, as was the case prior to the 1930s, when stockholders of nationally chartered banks were liable for losses up to twice the par value of the stock owned.

The implementation of prompt market-value closure would raise many political problems, especially during a transitional period. For example, the closure of institutions that are currently insolvent would raise major problems for the FSLIC, and possibly even the FDIC. However, these are the very institutions that now pose the gravest threat to the insurance funds. Nevertheless, it would be possible and desirable to move closer to market-value accounting and closure rules. Moreover, once insured institutions adjust to a truly unforgiving closure policy, they would voluntarily hold more capital in relation to the riskiness of their portfolios to reduce or eliminate the risk of being declared insolvent.

The current risk-based capital proposal, which requires banks with more risky assets and off-balance sheet activities to hold more capital, can be considered a step in the same direction. Such proposals, however, will succeed in eliminating or reducing the moral hazard in deposit insurance only if they help to ensure that insured institutions maintain a positive *market* value of capital over a wide range of possible *ex post* outcomes. Since riskier assets have a higher probability of declining in value, requiring additional capital for these assets *ex ante* increases the probability of a positive market value *ex post*.

Like a scheme of risk-based deposit insurance premia, a true risk-based capital approach would require *ex ante* estimates of the value and riskiness of each type of asset as well as its contribution to the overall riskiness of the portfolio. However, an approach requiring banks to hold additional capital (even if based on a fairly crude assessment of risk) probably would be easier to implement and less likely to generate errors that cause major distortions than a system of risk-based deposit insurance premia.

In sum, there are practical and political problems with each of the approaches to insurance reform described. But if we wish to maintain deposit insurance coverage that is as extensive as what we have now, reform is necessary. The optimal approach probably will involve a blend of reforms.

Bank Powers

At the heart of the conflict between the natural evolution of the financial system and the legal and regulatory structure governing that system is the issue of bank powers. The current restrictions on bank ownership and powers, enumerated in the Glass-Steagall and Bank Holding Company Acts, stand in the way of the trend towards functional realignment in the provision of financial services. While market forces will foster the development of alternatives to bank-provided payments and credit services, these alternatives may not be the most efficient from society's perspective.

Specifically, preservation of the current restrictions on bank powers will cause financial activity to continue to shift away from banks to nonbank banks, thrifts, and investment banks. This shift implies both a relative decline in business transacted by banking firms and a rearrangement of activity within the corporate structure of bank holding companies. Failure to resolve the nonbank bank issue will lead to a decline in the value of the traditional commercial bank charter, and may even cause banking firms to shift activities to nonbank subsidiaries. In fact, one bank consulting firm has advocated a corporate restructuring dubbed "double de-banking" in which the bank holding company relinquishes its commercial bank charter in favor of a nonbank bank charter (to retain payments system access) while placing all of its other financing, underwriting, and loan servicing activities in separate nonbank subsidiaries.[14]

The basic conflict between economic forces and regulation extends beyond domestic markets. As activity continues to shift to less-regulated international centers, bank regulators will find themselves regulating and supervising a shrinking share of total financial activity. To the extent that the quality of supervision deteriorates because of the difficulty of supervising an international banking organization in its entirety, the stability of the financial system could be threatened. These challenges to supervision could be overcome, in part, by coordinating supervision and regulation in the world's three most important financial centers — New York, London, and Tokyo. The U.S.-U.K. risk-based capital proposal is a first step. Nonetheless, because of restrictions on domestic banking powers, there remain strong incentives to shift activity toward less regulated environments.

Resolving the bank powers issue requires careful balancing of disparate concerns. On the one hand, because federal oversight and protection of some portion of financial activity is essential to stability, regulation must not be so at odds with market forces that important financial activities shift away from federal control. On the other hand, because the provision of a federal safety net creates incentives for excessive risk-taking, some minimum level of regulation, or at least supervision, is necessary.

Separation of Powers

Before we consider the extent to which bank powers ought to be expanded in response to market pressures, it may be useful to reconsider the original rationale for separating banking from other financial services and from commerce. Of primary concern to legislators in the 1930s were the problems associated with concentration of resources and the potential for self-dealing. Such problems have been addressed, with varying degrees of success, in other countries without completely separating banking and securities markets.[15] Moreover, since the 1930s, the problems may have been mitigated to some extent in the U.S. by SEC regulations and surveillance. Likewise, antitrust restrictions should serve to prevent excessive concentration and anticompetitive behavior. Finally, if greater integration of financial services were allowed, the concentration of total financial resources might increase, whereas the concentration for particular services actually might decrease because a wider variety of firms would be providing them.

Unlike the 1930s, a key concern regarding bank powers today is the possibility that banking organizations would shelter additional activities under the federal safety net. For this reason, some have argued against expanding the powers of banking organizations, while others have argued that new powers be carried out only in separate subsidiaries. Most observers agree, however, that the type of corporate separability that we have today is not very likely to insulate the bank from losses of a nonbank affiliate in times of stress.[16] Truly effective corporate separateness might require completely separate identities for the bank and nonbank affiliates, separate boards of directors, and severe limitations on inter-affiliate transactions. Such an approach might severely restrict or even eliminate any potential synergies the consolidated organization otherwise might enjoy.[17]

There is yet another view on the bank powers problem. Reform of the deposit insurance system to reduce its risk-taking incentives would make it easier to expand bank powers in response to market forces. With fewer limitations on bank powers, there would be less incentive for financial activity to shift away from federally supervised institutions.

Expand the Financial Powers of Banks

Along with a program for meaningful insurance reform, two broad reforms of bank powers might be considered. First, we might consider expanding the financial powers of banks. In other words, banks might be allowed to underwrite and trade securities, underwrite and sell insurance, manage mutual funds, and offer other financially related services. This approach would accommodate the trend towards functional realignment in the provision of financial services. It also would enhance the efficiency of the financial system by, among other things, enabling banks both to originate underlying assets and then to underwrite and sell derivative securities.

This approach may require increased surveillance of the activities of the consolidated enterprise since it increases a bank's opportunities for risk-taking. Such surveillance need not be a major stumbling block, however. In other countries where greater integration of financial services is allowed, regulators apparently have been able to supervise the activities of financial conglomerates.[18] Of course, such supervision may be easier to carry out in countries where there is only a handful of large banks.

Expand the Commercial Powers of Banks

A second general approach to the reform of bank powers would be to expand both the financial and commercial powers of banks. This approach would enable banks to own and control commercial firms and *vice versa.* Concerns regarding increased concentration of corporate control could be resolved through ownership limits, as have been established in West Germany, to prevent banks from exercising too much influence over the economy.

Once again, however, expanding bank powers in this way could complicate the assessment of the risks borne by the deposit insurance system. For example, the pressure to lend to troubled "house" firms may increase affiliate banks' risk unless federal supervisors can evaluate the soundness of all inter-affiliate transactions. (Alternatively, bank regulators could ban all inter-affiliate transactions, but if the ban were effective, it would severely reduce the benefits of conglomeration.) Reform of the deposit insurance system would, in theory, reduce the problem of increased risk. In practice, however, fully effective reform rests on the ability of regulators to monitor the market value of the consolidated enterprise — a difficult task, at best.

Given these difficulties, it is debatable just how far we should proceed in the direction of allowing banks to affiliate with commercial firms. One advantage of such affiliations would be the reduction of risk through the conglomeration of dissimilar activities. However, the operating synergies between banking and commerce do not appear to be great. Instead, there is some evidence that commercial firms are seeking banking powers primarily because they desire access to the payments system and wish to take advantage of related marketing synergies. If this were true, one way to resolve the issue of integrating banking and commerce would be to grant nondepository firms access only to the payments system provided they collateralized their transactions.

The Payments System

The major trends enumerated here bear importantly on the functioning of the payments system. Increased financial activity, securitization, and internationalization of markets presage a growing payments volume. There is legitimate concern that these trends may increase both the possibility and consequences of losses arising from a payments system malfunction or from the failure of a major participant in the system.[19]

In a payments system that uses the creation and extinction of credit to facilitate payments activity, such failures can generate liquidity problems for participants. With highly interconnected payments flows that rely on credit, a single failure can cascade into liquidity problems throughout the payments network. One of the functions of a central bank, of course, is to provide liquidity to sound institutions in such circumstances. However, central bank payments system policy should not imply protection against insolvency or even encourage frequent use of the emergency liquidity facility.

The Status Quo

The consequences of maintaining current payments conventions in light of anticipated growth trends in the volume of payments are worth considering. The payments system now entails underpriced intraday credit, delayed settlement, and access that is limited to depository institutions.

Underpriced intraday credit arises in several ways. First, the Federal Reserve encourages use of intraday credit by not charging for daylight overdrafts. Although the Fed is charging an implicit price on very large intraday credit activity as the result of a policy of limiting daylight overdrafts, it does not price most intraday Fed credit. Second, the Federal Reserve does not charge for the default risk it assumes by offering finality of payment on Fedwire. Thus, receivers of funds on Fedwire are not a potential source of discipline in the payment-credit decision. This distribution of risk differs from that of private networks where the provisional nature of transactions makes receivers evaluate the creditworthiness of payors.

Finally, some argue that there are externalities associated with payments activity that lead to the underpricing of a credit associated with payments even on wholly private networks. In particular, they argue that individual payments are transacted in ignorance of the burden that would be imposed on others should that transaction fail. If this view were correct, private charges for payments credit would be lower than the social cost of that credit. This and other causes of underpriced payments credit encourage the use of intraday credit that may be too large from the viewpoint of economic efficiency.

The delayed settlement feature of present day private payments systems adds to the concerns raised by underpricing. Delayed settlement increases the chances that an adverse event will nullify transactions that have already taken place. As payments activity grows and the interconnectedness of the payments system increases, some argue that the likelihood of such disruptions will increase. Combined with excessive use of payments system credit because of underpricing, this additional concern raises the risk of coincident liquidity or solvency problems for participants that could, in turn, precipitate a general loss of confidence in the payments system.

Although intervention by the central bank should be able to protect the economy from such liquidity problems, such intervention is not costless and, if performed frequently, could create additional incentives for risk-taking, particularly if the intervention extends beyond providing liquidity to ensuring solvency. Thus, the problem of excessive payments system risk — like excessive risk-taking in other facets of banking — is a serious concern of current payments system policy.

The third feature of concern in the current payments system is access that is limited to depository institutions. Nonbank institutions have been overcoming the limitation through thrift and nonbank bank ownership. In addition, nondepository firms are using sweep-type arrangements to provide payments services to their customers. These arrangements, however, may affect the size and timing of payments activity in an undesirable way from the standpoint of payments system risk. For these reasons, payments system access is of increasing concern whether or not there is a change in explicit access policy.

Pricing Fed Credit

Further progress toward the pricing or rationing of intraday Federal Reserve credit would remove a major stimulus to the overuse of intraday credit, both on Fedwire and on private wholesale networks. These additional steps should be taken despite issues raised by Federal Reserve payments queues and computer malfunctions, although improvements in these areas should be made in conjunction with pricing efforts.

Ideally, intraday credit pricing would embody not only the time value of funds, but also the value of the default risk implicitly assumed by the Federal Reserve in granting finality of payment on Fedwire. This approach would simulate the discipline exerted by receivers of funds in the private intraday credit market, and reduce the direct credit risk to which the Federal Reserve System would be exposed.

With a positive price for intraday credit, overall use of such credit would decline. In the short run, this decline may retard the growth of activities that have become reliant on underpriced intraday credit, such as churning in the securities market and the corporate cash management process generally. It is not clear, however, that the current level of payments activity (involving daily flows of $1 trillion or more) is efficient, whereas it *is* clear that the current system induces inadequate credit evaluation. The latter increases the risk of payments failure on private networks, or, alternatively, the risk to the Fed on Fedwire.

Pricing Fedwire intraday credit presumably would push more payments activity into the private credit market. Although such a shift might increase risk in the private market, funds receivers in the private market do have an incentive to monitor and control their risk exposure. Private bilateral payments decisions, however, may not automatically take into account the total "social" credit risk involved. Reduction of this risk requires surveillance by the appropriate regulators and the principal participants in private payments networks. Such surveillance may require minimum participant capital (or liquid reserve) requirements, net debit limits, or other risk-limiting devices such as those currently employed by private intraday credit systems such as CHIPS and Euroclear.

Analogous to charging interest on intraday overdrafts, interest also should be paid on positive balances. Symmetry in the treatment of borrowing from and lending to the Federal Reserve System would improve the functioning of the private intraday credit market. It also would decrease Fedwire congestion associated with attempts to maintain minimum required reserve balances at the end of the day, and would enhance the attractiveness of holding corporate demand deposits at banks (which also should be allowed to yield explicit interest).

Real-Time Settlement

In addition to providing better management of risk in a delayed-settlement environment, an increased price for intraday credit will encourage a transition toward "real-time settlement" whereby both monitoring of positions and matching of payments flows will occur on a continuous basis. A payments system should be a credit system only if it is more efficient to bridge temporal gaps between the payment and receipt of funds through borrowing than through expending resources to make transactions synchronous. Under the current system, borrowing and *asynchronous* payments are favored.

With costly intraday credit, participants will seek the means to synchronize transactions and settle obligations in "real time." For example, repayment of funds borrowed overnight will be more closely matched in time with funds inflows that reflect borrowing for the next night. Such operations, if exactly matched in time, will reduce overdraft exposure by substituting a relatively small net transfer (the difference between the two borrowings) for two gross transfers mismatched in time.

Real-time settlement is becoming increasingly feasible as communications and electronic accounting technologies advance. Since real-time settlement eliminates, by definition, temporal risk in the payments system, the evolution toward real-time settlement will contribute significantly to reducing payments system risk. Many transactions may be quite costly to settle in real time, of course, and the payments system will continue to involve credit extension to some degree. However, as around-the-clock and global securities trading progresses, the importance of managing temporal risk will mount, and real-time payments technology increasingly will be needed to manage risk economically.

Finally, by resolving the problems of underpriced intraday credit and delayed settlement, there would be less need to continue to limit access to the payments system. An orderly expansion of payments system access, in conjunction with these other reforms, likely would not pose undue risk and would resolve the problems created by non-depository firms exploiting various loopholes in the current policy.

IV. Summary and Conclusions

A financial revolution is underway. Already we see glimpses of the new financial world in the forms of increased securitization, a diminished role for bank-provided intermediation, functional realignment in and geographic integration of financial services, and expanded access by nonbank firms to the payments system. These are trends driven both by fundamental economic forces and attempts to circumvent regulation and to exploit government guarantees.

Many of these changes have not resulted from explicit policy choices. While most would admit that a thorough reform of financial regulatory and legal policy is long overdue, the continuing debate over just what changes are necessary apparently has paralyzed the policymaking process.

Although there are no easy or simple solutions, the time has come to move forward because failure to make the needed changes may threaten financial stability. Three areas are especially in need of thorough reform: the federal safety net, the payments system, and bank powers.

The major problem with our current deposit insurance system is that it provides an incentive for excessive risk-taking, which could propagate throughout the economy as distinctions between banks and nonbank firms diminish. Without changes in the insurance system, the government could be left underwriting the risks of an ever-increasing share of the economy.

Similarly, the implicit government guarantee behind the payments system may prove to be unsustainable in the face of rapid financial innovation. Underpriced intraday credit in conjunction with delayed settlement appears to be a major part of the problem. Without reforms in these areas, expanded payments system access poses further risks.

Finally, banks are experiencing economic pressures to expand into nontraditional activities. A major reason for preventing them from doing so is to protect the deposit insurance and payments guarantees. However, many observers question whether the U.S. banking industry will be able to compete effectively if it continues to be regulated more stringently than domestic nonbank firms and banking firms in other countries.

Clearly, market forces for change are posing serious challenges to the current financial regulatory framework and safety net. By reforming the legal and regulatory framework to accommodate these forces and to encourage more market discipline of risk-taking, we can move toward a more efficient and stable financial system. Undoubtedly, a blend of many of the approaches touched upon here will be needed to reach these goals.

FOOTNOTES

[1] Summary of the Results of the August 1985 Senior Loan Officer Opinion Survey on Bank Lending Practices conducted by the Federal Reserve System.

[2] Source: *Federal Reserve Bulletin,* various years. Additional sources on securitization are: Christine Pavel, "Securitization," *Economic Perspectives,* Federal Reserve Bank of Chicago, July/August 1986; and Randall Pozdena, "Securitization and Banking," *Weekly Letter,* Federal Reserve Bank of San Francisco, July 4, 1986.

[3] Michael Keeley, *"Interest on Business Checking Accounts?," Weekly Letter,* Federal Reserve Bank of San Francisco, May 2, 1986.

[4] For additional information on functional realignment, see: Bank for International Settlements, "Recent Innovations in International Banking," April 1986; Alan Daskin and Jeffrey Marquardt, "The Separation of Banking and the Securities Business: A View of the United Kingdom, West Germany and Japan," *The World of Banking,* May/June 1984; Steven Felgran, "Bank Entry Into Securities Brokerage: Competitive and Legal Aspects," *Bankers Desk Reference: New Topics, 1985;* Samuel Hayes, "Investment Banking: Commercial Banks' Inroads," *Economic Review,* Federal Reserve Bank of Atlanta, May 1984; and Christine Pavel and Harvey Rosenblum, "Banks and Nonbanks: The Horse Race Continues," *Economic Perspectives,* Federal Reserve Bank of Chicago, May/June 1985.

[5] See also, Bank for International Settlements, "Recent Innovations in International Banking."

[6] See Donald Savage, "Interstate Banking Developments," *Federal Reserve Bulletin,* February 1987; Federal Reserve Bank of Chicago, *Toward Nationwide Banking: A Guide to the Issues, 1986;* Constance Dunham and Richard Syron, "Interstate Banking: The Drive to Consolidate," *New England Economic Review,* Federal Reserve Bank of Boston, May/June 1984; and Frank King, "Interstate Banking: Issues and Evidence," *Economic Review,* Federal Reserve Bank of Atlanta, April 1984.

[7] See, for example: E. Gerald Corrigan, "Are Banks Special? A Summary," *Federal Reserve Bank of Minneapolis Annual Report, 1982;* and Michael Keeley and Frederick Furlong, "Bank Regulation and the Public Interest," *Economic Review,* Federal Reserve Bank of San Francisco, Spring, 1986.

[8] Ben Bernanke and Mark Gertler, "Agency Costs, Collateral and Business Fluctuations," a paper presented at the Federal Reserve Bank of San Francisco's Fall Academic Conference, 1986.

[9] For discussion of this issue, see: Frederick Furlong and Michael Keeley, "Bank Runs," *Weekly Letter,* Federal Reserve Bank of San Francisco, July 25, 1986; and Barbara Bennett, "Bank Regulation and Deposit Insurance: Controlling the FDIC's Losses," *Economic Review,* Federal Reserve Bank of San Francisco, Spring 1984.

[10] Milton Friedman, *A Program for Monetary Stability,* New York: Fordham University Press, 1959; and Henry Simons, *Economic Policy For a Free Society,* Chicago: University of Chicago Press, 1948.

[11] Robert Litan, "Taking the Dangers Out of Bank Regulation," *The Brookings Review,* Fall 1986; and John Kareken, "Ensuring Financial Stability," in *The Search for Financial Stability: The Past 50 Years,* San Francisco: The Federal Reserve Bank of San Francisco, 1985.

[12] Frederick Furlong, "FDIC's Modified Payout Plan," *Weekly Letter,* Federal Reserve Bank of San Francisco, May 18, 1984.

[13] Frederick Furlong, "Market Responses to Continental Illinois," *Weekly Letter,* August 31, 1984.

[14] "Continued Inaction by Congress May Spur Banks to Drop Charters," *American Banker,* November 6, 1986.

[15] Daskin and Marquardt, *op cit.*

[16] See Anthony Cornyn, Gerald Hanweck, Stephen Rhoades and John Rose, "An Analysis of the Concept of Corporate Separateness in BHC Regulation from an Economic Perspective," *Appendix C. to the Statement by Paul A. Volcker,* June 1986. For an opposing view, see Samuel Chase and Donn Waage, "Corporate Separateness as a Tool of Bank Regulation," Samuel Chase and Company for the American Bankers Association, October, 1983; and Carter Golembe and John Mingo, "Can Supervision and Regulation Ensure Financial Stability?," in *The Search for Financial Stability; The Past 50 Years,* San Francisco: The Federal Reserve Bank of San Francisco, 1985.

[17] Paul Volcker, "Statement Before the Subcommittee on Commerce, Consumer and Monetary Affairs of the U.S. House of Representatives," June 11, 1986.

[18] Daskin and Marquardt, *op cit.*

[19] For further discussion of payments system risk, see: David Humphrey, "The U.S. Payments System: Costs, Pricing, Competition and Risk," Monograph Series in Finance and Economics, no. 1984-1/2, New York: Salomon Brothers Center for the Study of Financial Institutions, Graduate School of Business Administration, New York University, 1984; David Mengle, "Daylight Overdrafts and Payments System Risks," *Economic Review,* Federal Reserve Bank of Richmond, May/June 1985; and E.J. Stevens, "Risk in Large-Dollar Transfer Systems," *Economic Review,* Federal Reserve Bank of Cleveland, Fall 1984.

Bank Capital Regulation and Asset Risk

Frederick T. Furlong and Michael C. Keeley*

This paper examines theoretically the effects of more stringent capital regulation on a bank's incentive to increase asset risk and on the expected liability of the deposit insurance system. Our analysis shows that regulatory increases in capital standards will not require greater regulatory efforts to restrain asset risk because a bank's incentive to increase asset risk declines as its capital increases. Thus, as long as regulatory efforts to contain asset risk, such as bank examinations, are not reduced, more stringent capital regulation will reduce the expected liability of the deposit insurance system.

Over the past several years, bank regulators have placed greater emphasis on the regulation of bank capital. For example, the three federal bank regulatory agencies have raised capital requirements for banks and bank holding companies and established more uniform standards among themselves.[1] Most recently, the federal bank regulatory agencies have put forth proposals for risk-based capital requirements that would be coordinated with the Bank of England.[2]

These regulatory measures, in part, are reactions to deteriorating capital positions, particularly among the larger banking organizations. For example, among the twenty largest bank holding companies, the average ratio of the book value of common equity to assets was over 6 percent in 1968 but only about 4 percent in 1980.[3] The increase in the number of bank failures and the correspondingly sharp rise in the Federal Deposit Insurance Corporation's (FDIC's) expenses in recent years also have intensified interest in capital regulation. Total expenses of the FDIC, which fluctuated between about $50 million and $200 million per year in the 1970s, rose to about $2 billion per year in 1985 and 1986. Such increases in expenses for the deposit insurance system have focused attention on increasing the stringency of bank capital regulation to limit the FDIC's exposure to losses and to blunt the incentives for "excessive" risk-taking by federally insured banks.

The move to more stringent capital standards in banking, however, has met with considerable controversy as well as some skepticism. Some argue that higher capital requirements will cause banks simply to invest in more risky assets, and thereby offset, or even more than offset, the desired effects of higher capital. This view often is echoed in the financial press. In a New York Times article about a Federal Reserve proposal to require banks to hold capital in connection with agreements involving interest rate and currency swaps, William McDonough, vice chairman of First National Bank of Chicago, is quoted as saying that ". . . the proposal could lead banks to take on riskier business to compensate for the lower returns they would almost assuredly get by having to maintain more capital."[4]

The effectiveness of capital regulation also has come under question in the academic literature. A study by Koehn and Santomero (1980), which assumes that banks maximize utility in a mean-variance framework,[5] is representative of the literature on the theoretical relationship between capital

* Senior Economists, Federal Reserve Bank of San Francisco. Comments from Jack Beebe, Christopher James, David Pyle, Anthony Santomero and John Scadding on earlier drafts are appreciated. Research assistance from Alice Jacobson and John Nielsen is gratefully acknowledged.

requirements and bank asset risk. They conclude that ". . . a case could be argued that the opposite result can be expected to that which is desired when higher capital requirements are imposed."[6]

In this paper, we evaluate the popular view of capital regulation and the conclusions of earlier theoretical studies on the effectiveness of capital regulation. In contrast to the popular view and the earlier academic work, we find that more stringent capital standards alone would not give a bank more of an incentive to increase the riskiness of its assets. In fact, the incentive to increase asset risk falls as a bank's capital increases. This implies that, as long as regulatory and supervisory efforts to limit asset risk in banking, such as bank examinations, are not relaxed, increasing a bank's capital will lower that bank's chance of failure and reduce the expected liability of the deposit insurance system.

We also show in the Appendix that the conclusions reached by earlier theoretical studies using the mean-variance framework were derived from internally inconsistent assumptions. In essence, these studies implicitly (but unintentionally) assume that bank failure is not possible by assuming that borrowing costs are unrelated to bank risk. Yet, they seek to analyze the effects of capital regulation on the probability of bank failure. Moreover, these studies fail to incorporate the effect of the deposit insurance guarantee on risk-taking. Although the results of these studies regarding the effects of capital regulation on the incentive to increase asset risk are technically correct when bank failure is not possible, such findings are of little policy relevance since capital regulation and concern over risk-taking are relevant only when banks can fail.

The Appendix also contains an example to show that the results of these earlier studies do not generally hold when subsidized deposit insurance and the possibility of bankruptcy are taken into account. Specifically, we show that when the asset return distribution is binomial, the incentive to increase asset risk does not increase as the stringency of capital regulation increases.

The analytic framework used in the body of this paper is the state-preference model rather than the mean-variance model used in the older literature on the topic. One reason for this choice is that the state-preference model, unlike the mean-variance model, can easily accommodate the possibility of bankruptcy and an analysis of the effects of mispriced deposit insurance on a bank's choice of leverage and asset risk.[7] Moreover, with the state-preference model, the effects of changes in capital requirements on both banks' gains from increasing asset risk and the expected liability of the deposit insurance system can be evaluated directly.

Another advantage of the state-preference framework is that it can be applied to the analysis of both value-maximizing and utility-maximizing banks. Utility maximization might be appropriate for certain smaller, closely held banks where the owners' risk preferences affect the riskiness of the banks' portfolios, whereas value maximization is more suitable for most other banks, particularly the large publicly held banks whose stockholders can hold diversified portfolios. Value-maximizing banks would seek to maximize the current market value of their equity, which is independent of the risk preferences of the owners.[8]

In the next section, we start with a discussion of the nature of bank capital and the issues that higher capital requirements raise for bank regulators. In Section II we introduce the state-preference model and use it to analyze the effects of capital regulation on asset risk and the liability of the federal deposit insurance system, under the assumption that banks choose to maximize the value of stockholders' wealth. Section III contains a similar analysis, applying the state-preference model to utility-maximizing banks. The conclusions and policy implications are presented in the final section.

I. Issues in Capital Regulation

A bank's financial capital — that is, its equity — is the difference between the value of the institution's assets and liabilities. Banks use both capital and liabilities to finance loans and investments.[9] The two sources of funding are distinguished in that variations in earnings on assets are borne first by capital holders. The larger the proportion of assets funded by capital, the greater the range of returns on assets that will be borne solely by equity holders and the more likely the promised obligation to liability holders will be met. Thus, if banks were not insured, both equity and liability holders, including depositors, would have an interest in the level of a bank's capital. As with other firms, the stockholders and liability holders (depositors) of unregulated banks would be expected to "decide" on a satisfactory combination of capital financing and promised return on bank debt.

The regulation of bank capital, then, must be predicated on the assumption that a market determination of the level of capital would not be satisfactory from a public policy perspective. While capital regulation predates federal deposit insurance, partly because of the externalities argued to be associated with bank failures, the provision of the federal deposit guarantee commonly is cited as the main reason that the level of bank capital must be a regulatory concern.

The federal deposit insurance system, by guaranteeing deposits, in essence takes on the role of a bank liability holder and has an interest in bank capital similar to that of private liability holders in an uninsured firm. Indeed, some have argued that the deposit insurance system has taken on the role of covering virtually all bank liability holders in the event of an insolvency. If so, the insurance system would be the only liability holder with an interest in bank capital.

From a regulatory perspective, a bank with more capital relative to assets will be less likely to fail, and, if it does fail, will impose smaller losses on the insurance fund, all other things equal. However, the probability of failure and the contingent liability of the insurance system also depend on the variability of the return on assets.[10] The higher the variability of the return on assets for a given amount of capital, the greater the chance of bank failure.[11]

Consequently, a central issue in capital regulation is whether banks would respond to higher regulatory capital requirements by choosing riskier assets to offset or even more than offset the effects of higher capital on the exposure of the deposit insurance system to bank risk. Below, we consider this issue and examine under what conditions, if any, regulation-induced increases in bank capital would lower the expected losses of the deposit insurance system.

II. Value Maximization

A value-maximizing bank chooses its portfolio solely to maximize the current market value of equity. Such a bank's portfolio decisions are independent of the risk preferences of its individual owners because the owners are fully able to adjust the composition of their personal portfolios to attain any level of risk they desire. Therefore, even though actual returns on the bank's portfolio are uncertain (risky), a value-maximizing bank does not consider the risk preferences of the owners.

Some of the implications of bank capital regulation for value-maximizing banks within the state-preference framework are discussed in Dothan and Williams (1980), Sharpe (1978), and Kareken and Wallace (1978).[12] All of these studies provide theoretical support for restricting leverage in banking when there is subsidized deposit insurance. They do not, however, deal with the issue of how the asset investment strategies of such insured banks might be altered by capital regulation. Nor do they consider how the behavior of a utility-maximizing bank in the state-preference framework might differ from that of a value-maximizing bank.

This portion of the paper addresses the first of these two issues by extending the examination of bank capital regulation within a state-preference framework. For the reader who is not familiar with this framework, a brief description of the state-preference model is presented in Box 1. Below, we first show why leverage constraints are necessary for insured, value-maximizing banks. Then, we assess the likely effects of changes in capital requirements on the asset risk of such banks and on the liability of the deposit insurance system. The discussion in Section III turns to the implications of deposit insurance and capital regulation for utility-maximizing banks within a state-preference model.

Value-Maximizing Banks

Although the state-preference model can be applied to an individual investor's decisions, it also can be used to analyze the portfolio and leverage decisions of an insured bank that maximizes its current value (the market value of its equity). Since the current value of such a bank is independent of the risk preferences of the owners, we can put aside any consideration of utility functions and focus instead on how an insured bank's investment opportunity frontier itself is affected by leverage and capital regulation.

The effects of leverage and the role of capital regulation can be seen most easily through a numerical example with two states and two securities. In this example, security A represents a promise to pay \$4 if state 1 occurs and \$6 if state 2 occurs, and is summarized as A(4, 6). The second security, security B, is summarized as B(1, 1). Security A is a risky investment (a different payout in each state) and security B is a riskless security (the same payout in each state). For expositional purposes, we assume that the current market price of a dollar payment in state 1 is \$.35 and the price of a dollar payment in state 2 is \$.60. The current price of a share of security A then is \$5. ($\$.35 \times 4 + \$.6 \times 6$) and the current price of security B is \$.95 ($\$.35 \times 1 + \$.6 \times 1$).[13]

The bank is assumed to invest only in the risky security. The bank's purchases of that security are funded with a combination of capital and the proceeds from issuing shares of security B. Shares of security B can be thought of as deposits that are "insured" at a fixed-premium rate by the federal government. In the example, the premium is set at zero, but the analysis and conclusions would hold even with a positive, fixed-rate premium.[14] Initial capital is set at \$500 by assumption. With no deposits, the bank would have 100 shares of security A, and leverage would be one.

The calculations in Table 1 demonstrate what happens to the total net worth (current value) of the bank's equity as leverage increases. In line 2, the bank increases leverage to 2 by issuing deposits with a current value of \$500 and purchasing an additional 100 shares of security A. In both states 1 and 2 the bank promises to repay depositors \$526.32 (\$500/.95). The net claims (future wealth) of the bank in each state after paying off deposits are shown in column 6. The current value of the bank to the owners, column 8, is derived by multiplying the net claims by the price of a dollar claim in the appropriate state. The addition to the value of the bank from the free deposit insurance, presented in the last column, is derived by subtracting the initial capital, \$500, from the total value of net worth.

As Table 1 shows, initially the bank's value (net worth) is not affected by issuing deposits. At the lower levels of leverage, the bank would be indifferent to the amount of borrowing because its value (column 8) would be unaffected.[15] Although the risk of the bank increases with leverage, as reflected in the growing disparity between the claims in the two states (column 6), bankruptcy could not occur and the deposit insurance fund would not be at risk with leverage of 4 or less. Moreover, depositors would be indifferent to the risk of the bank whether or not their funds were insured as long as leverage was less than or equal to 4.

It is easy to see why the insurance fund as well as depositors are not at risk at low levels of leverage. Up to a point, the bank is able to meet its promised payments to depositors in both states 1 and 2. The bank would not fail in either state since its liabilities would not exceed its assets. Therefore, while risk increases with leverage, as long as the bank's capital is sufficient to ensure payment, the added risk is borne only by the bank.[16]

As leverage continues to increase, the bank eventually will be unable to meet its promised obligations to depositors in the first state. Without a third party guarantee such as deposit insurance, depositors would not be willing to lend to a bank in return for a promise to pay only \$1 (per share) in each state. With leverage equal to 5, for example, the bank would issue a deposit with a current value of \$.95 per share but would have to promise to pay about \$1.03 in each state instead of \$1 (the actual pay-

Box 1
The State-Preference Model

The state-preference model can be used to analyze investors' decisions that affect their future consumption. In this model, an investor chooses among combinations of claims to wealth in all possible future states of the world.*

In its simplest form, the state-preference approach includes only two time periods: now and the future. The future is uncertain because only one of a number of possible states of the world actually will occur. The possible states of the world, for example, might be represented by inflation, deflation, and unchanged prices. An investor's claim to future wealth (measured in terms of real purchasing power) if state i were to occur can be expressed as W_i.

To secure future claims to wealth, the investor in the state-preference model can purchase (or issue) securities. The securities are characterized by the payments made in the various possible states. Consider, for example, a world in which there are two possible future states, 1 and 2, and two securities, A and B. One share of security A pays a_1 real dollars in state 1 and a_2 in state 2. We can summarize these characteristics by $A(a_1, a_2)$. One share of security B can be summarized as $B(b_1, b_2)$.

The current value or price of a share of either security can be expressed in terms of two current market prices: p_1, the current price of a future payment of one dollar if state 1 occurs, and, p_2, for a one dollar claim in state 2. (These prices are taken as given by the investor and are unaffected by his decisions.)**

The current price of security A then is: $p_A = p_1a_1 + p_2a_2$. And the price of security B is: $p_B = p_1b_1 + p_2b_2$. The key insight of the state-preference model is that all future uncertain claims have a certainty-equivalent current value as determined by the prices of a dollar claim in each state.

The number of shares of securities that an investor can purchase is determined as follows. The amount of "capital," K, that an investor decides to allocate to purchase or secure claims on future wealth is taken as given. If only security A were purchased, the investor's future claims would be the number of shares purchased times the per share payout in each state: (S_Aa_1, S_Aa_2), where S_A is the number of shares of security A and $S_A = K/p_A$. If only security B were purchased, the claims would be (S_Bb_1, S_Bb_2), where S_B is the number of shares of security B and $S_B = K/p_B$.

In Figure A, we have plotted the various combinations of wealth in state 1 and state 2, denoted as W_1 and W_2, that can be attained through various holdings of security A and security B. These combinations are represented by the negatively sloped straight line in the figure called the investment opportunity frontier. Point A represents the wealth outcomes (S_Aa_1, S_Aa_2) if only security A were purchased, and point B, the outcomes (S_Bb_1, S_Bb_2) if only security B were purchased. The points between A and B represent combinations where K is divided between holdings of both securities.

For points on the opportunity frontier below point A, the investor issues security B (borrows) and uses the funds to purchase additional shares of security A. Similarly, for the points on the frontier above point B, the investor issues security A and purchases additional shares of B. Thus, through various combinations of holding and issuing securities A and B, an investor can attain any combination of wealth in the two states along the frontier. Borrowing, however, does not affect the current value of the investor's net future claims (claims after the payments on the issued security have been made), which is equal to K.***

Figure A
The Investment Opportunity Frontier

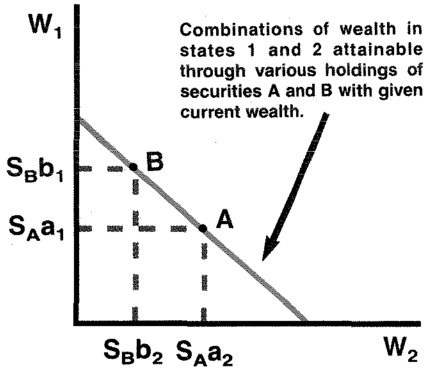
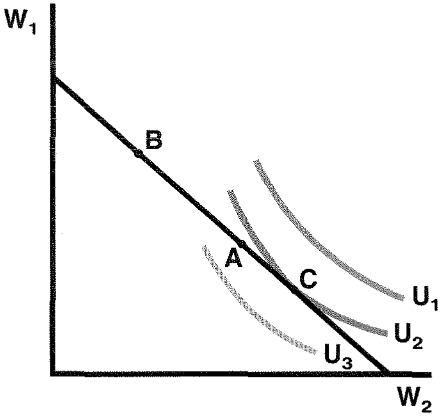

Combinations of wealth in states 1 and 2 attainable through various holdings of securities A and B with given current wealth.

Figure B
An Investors Choice of Wealth in each State



Given the constraint on current wealth, the actual combination of assets chosen is determined by the investor's preferences concerning the trade-off between wealth in the different future states. Those preferences can be represented generally as the utility function $U(W_1, W_2)$. A number of factors could affect the nature of utility functions, including the investors' assessments of the probabilities of the states occurring. As is generally the case in economic models, in the state-preference model an investor's utility function is assumed to be convex. In the state-preference model, the convexity of utility functions indicates that investors are risk-averse.

Examples of convex utility functions are shown in Figure B, where utility rises such that $U_1 > U_2 > U_3$. An investor will choose a combination of securities — a point on opportunity frontier (the line on which wealth equals K) — that results in the highest utility. The point of highest utility would be the one and only point at which a utility curve is tangent to the opportunity frontier. In the figure, that point is represented by C. Point C is attainable by issuing shares of security B and investing in shares of security A.

* For a complete discussion of the state-preference model see Sharpe (1970).

**Arrow (1964) and Debreu (1959) have shown that a competitive market equilibrium determines these prices.

***This is always true in the absence of any subsidized third party guarantee of such borrowing. However, as is shown in the body of the paper, subsidized deposit insurance does affect current wealth.

ments would be about $1.03 in state 2 and about .95 in state 1). The current value of the depositor's claims would be unaffected because the higher payment in state 2 would compensate for the lower payment in state 1. The deposit guarantee, however, would allow a bank with leverage greater than 4 to continue promising $1 to depositors in both states because the deposit insurance fund would cover the shortfall in state 1.

As seen in column 8 of Table 1, once leverage is extended to a point at which bank failure becomes possible, the current net worth of the bank begins to increase with leverage. The addition to net worth represents the current value of the deposit insurance guarantee (column 9). A bank gains from increasing leverage and simultaneously investing additional deposits in the risky security because the net claims of the owners in state 1 can never be less than zero, no matter how large the "promised" payments, while the potential claims in state 2 are unlimited. The state-preference model therefore predicts that a value-maximizing bank with an insurance premium less than the current value of the insurance payout would limit its leverage only if forced to do so by regulation.

TABLE 1

Effects of Leverage with 100% Insurance of Deposits

Security A ***A(4,6)*** ***$p_1 = \$.35$*** ***Initial Bank Capital = $500***
Security B (Deposit) ***B(1,1)*** ***$p_2 = \$.6$***

	(1)	(2)	(3)		(4)	(5)	(6)	(7)	(8)	(9)
						Future			Current	
	Initial Lever-age	Shares of Sec.A	Current Value of Depsoits *(Sec. B)*	State	Gross Claims	Deposit Payments	Net Claims *(4-5)*	Price of $1 Claim	Value of Bank *(6x7)*	Value of Deposit Guarantee *(Col. 8-$500)*
(1)	1	100	$ 0	1	$ 400.00	$ 0.00	$ 400.00	.35	$140.00	
				2	600.00	0.00	600.00	.6	360.00	
									$500.00	$ 0.00
(2)	2	200	500	1	800.00	526.32	273.68	.35	95.79	
				2	1200.00	526.32	673.68	.6	404.21	
									500.00	0.00
(3)	3	300	1000	1	1200.00	1052.63	147.37	.35	51.58	
				2	1800.00	1052.63	747.37	.6	448.42	
									500.00	0.00
(4)	4	400	1500	1	1600.00	1578.95	21.05	.35	7.37	
				2	2400.00	1578.95	821.05	.6	492.63	
									500.00	0.00
(5)	5	500	2000	1	2000.00	2105.26	0.00[1]	.35	0.00	
				2	3000.00	2105.26	894.74	.6	536.84	
									536.84	36.84
(6)	6	600	2500	1	2400.00	2631.58	0.00[2]	.35	0.00	
				2	3600.00	2631.58	968.42	.6	581.05	
									581.05	81.05
(7)	7	700	3000	1	2800.00	3157.89	0.00[3]	.35	0.00	
				2	4200.00	3157.89	1042.11	.6	625.27	
									625.27	125.27
(8)	8	800	3500	1	3200.00	3684.21	0.00[4]	.35	0.00	
				2	4800.00	3684.21	1115.79	.6	669.47	
									669.47	169.47

[1] Actual net claim in state 1 is - $105.26.
[2] Actual net claim in state 1 is - $231.58.
[3] Actual net claim in state 1 is - $357.89.
[4] Actual net claim in state 1 is - $484.21.

Capital Requirements and Risk

This brings us to the main question facing regulators: will regulatory efforts to force banks to hold more capital be partially or even totally offset by banks that then acquire riskier assets?

To answer this question, another risky asset has to be introduced. In addition to security A(4, 6), we assume that the bank also can hold the more risky security, security D(0, 8.33), where the numbers in parentheses are the dollar claims per share of the securities in the two possible states. The price of security D is $5 ($\$.35\times0+\$.6\times8.33$). A bank with a given degree of leverage can alter its net claims in future states by investing available funds in different combinations of these two risky assets.

Table 2 demonstrates how the value of a bank with an initial leverage of 3, initial capital of $500, and underpriced deposit insurance is affected by shifting from holding only security A to holding greater proportions of its assets in security D. Parallel to the case of increased leverage with asset risk held constant (Table 1), a bank with a given level of leverage benefits from increasing its asset risk with underpriced deposit insurance only when bankruptcy is possible (that is when deposit claims exceed the bank's gross claims in state 1). Once bankruptcy is possible, the value of the bank increases with asset risk (that is, with higher proportions of security D). Therefore, even if leverage were limited through regulation, a value-maximizing

TABLE 2

Effects of Asset Risk with 100% Insurance of Deposits

Security A	***A(4,6)***	***$p_1 = \$.35$***	***Initial Bank Capital = $500***
Security B	***B(1,1)***	***$p_2 = \$.6$***	***Initial Assets = $1,500***
Security D	***D(0,8.33)***		***Initial Leverage = 3***

	(1)	(2)	(3)		(4)	(5)	(6)	(7)	(8)	(9)
						Future			Current	
	Shares of Sec. A	Shares of Sec. D	Percent Shares in Sec. D	State	Gross Claims	Deposit Payments	Net Claims *(4-5)*	Price of $1 Claim	Value of Bank *(6x7)*	Value of Deposit Guarantee *(Col. 8-$500)*
(1)	300	0	0	1	$1200.00	$1052.63	$ 147.37	.35	51.58	
				2	1800.00	1052.63	747.37	.6	448.42	
									$500.00	$ 0.00
(2)	240	60	20	1	960.00	1052.63	0.00[1]	.35	0.00	
				2	1940.00	1052.63	887.37	.6	532.42	
									532.42	32.42
(3)	180	120	40	1	720.00	1052.63	0.00[2]	.35	0.00	
				2	2080.00	1052.63	1027.37	.6	616.42	
									616.42	116.42
(4)	120	180	60	1	480.00	1052.63	0.00[3]	.35	0.00	
				2	2220.00	1052.63	1167.37	.6	700.42	
									700.42	200.42
(5)	60	240	80	1	240.00	1052.63	0.00[4]	.35	0.00	
				2	2360.00	1052.63	1307.37	.6	784.42	
									784.42	284.42
(6)	0	300	100	1	0.00	1052.63	0.00[5]	.35	0.00	
				2	2500.00	1052.63	1447.37	.6	868.42	
									868.42	368.42

[1] Actual net claim in state 1 is - $92.63
[2] Actual net claim in state 1 is - $332.63
[3] Actual net claim in state 1 is - $578.63
[4] Actual net claim in state 1 is - $812.63
[5] Actual net claim in state 1 is - $1057.63

bank with deposit insurance would want to hold the risky security that maximized the value of the deposit guarantee. (In the example, this would be a portfolio that includes only security D.)

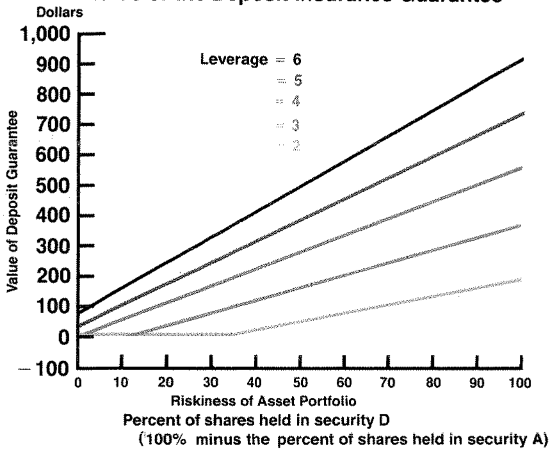
Figure 1 shows how the gains from increasing asset risk are affected by leverage. Each of the lines in the figure tracks the current value of the deposit guarantee to the bank that invests greater proportions of funds in security D (and correspondingly smaller proportions in security A), for a given degree of leverage. The marginal value to a bank from increasing asset risk (holding greater proportions of its asset in security D) is represented by the slope of a line.

With low levels of leverage and asset risk, the marginal value to increasing asset risk is zero (the lines are horizontal). However, for higher levels of leverage, the slopes of the lines increase as leverage increases, indicating that the marginal value of increasing asset risk increases with leverage. Put another way, as the capital of an insured bank increases, the marginal value to that bank of shifting to a riskier composition of assets falls. This means that more stringent capital requirements would *not* give banks a greater incentive to invest in riskier assets, and would reduce the liability of the deposit insurance system.[17]

With regulatory constraints on leverage, a bank still would want to hold the risky asset, security D.

Figure 1

Effects of Leverage and Risk on the Value of the Deposit Insurance Guarantee



Conclusions similar to those derived from the state-preference model regarding the implications of capital regulation for risk-taking (and the resulting contingent liability of the deposit insurance fund) can be derived by modeling the deposit insurance guarantee as a put option. In such an options approach, the bank is viewed as "purchasing" an option from the insurance fund to sell (put) its assets to the fund at an exercise price equal to the value of the bank's insured deposits (which we assume represent all bank liabilities). The bank would exercise this option only if it were insolvent, that is,when the assets were worth less than the liabilities.

Following Merton (1977), the Black-Scholes formula for a European option (one that can be exercised only at maturity) can be adapted to apply to the deposit insurance guarantee. Assuming all earnings are retained, the value of the insurance option can be expressed as

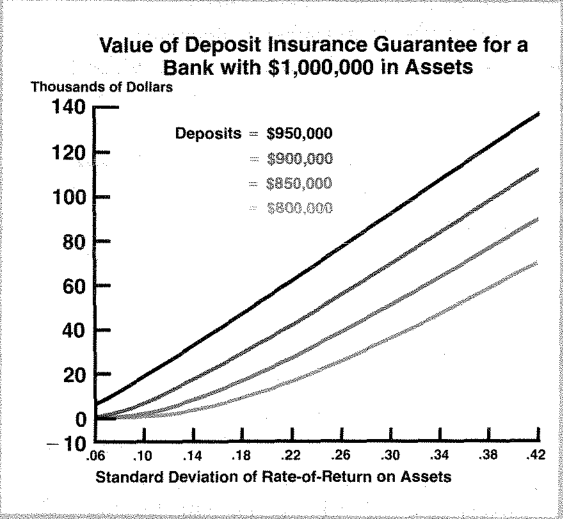
$$V_t = v(A, D, s, t),$$

where V_t is the current value of the contingent insurance liability, A is the current value of assets (excluding any insurance subsidy), D is the current value of insured deposits, s measures risk and is the standard deviation of the rate of return on assets, and t is the interval between examinations.*

Merton modified the put option model to apply to deposit insurance rather than equity securities. In his formulation, the value of assets of the bank replaces the stock price, the value of deposits represents the exercise or strike price,** the standard deviation of the rate of return on assets replaces the standard deviation of the return on the stock, and the examination interval corresponds to the time to maturity.

This approach generates the result that the value of the insurance guarantee increases with leverage.

Box 2
Capital Regulation and Asset Risk: An Options Approach



Value of Deposit Insurance Guarantee for a Bank with $1,000,000 in Assets

Greater leverage (that is, more deposits) in place of capital results in an increase in the exercise price and makes it more likely that the option would be exercised. Also, holding riskier assets, which would be reflected in a higher variation in the rate-of-return, would make it more likely that the value of assets would fall below the value of deposits, thus raising the value of the option. If a bank were not required to pay the full value of the option, then the bank could add to its owners' wealth by increasing leverage and/or asset risk.

The effects of leverage and asset risk on the value of the insurance guarantee from the options formula are illustrated graphically in the figure in this Box. For purposes of the figure, examinations are assumed to take place at the same time each year ($t = 1$) and the initial assets of the bank (excluding the value of the deposit guarantee) are set at $1 million.

In the figure, holding leverage constant (moving along a line from left to right), the contingent liability of the insurance fund increases with asset risk. Of importance to this paper is the observation that the marginal value of (and hence, incentive for) increasing asset risk increases with leverage. When the standard deviation of the rate of return on assets is held constant, the slopes of the lines in the figure increase as leverage increases. That is, at higher leverage (less capital), the change in the value of the insurance option for a given change in asset risk is greater.

As in the example using the state-preference model, then, raising capital standards (lowering leverage) reduces the marginal value of increasing asset risk for an insured bank. Unless other regulatory constraints are relaxed, the options approach also implies that regulating capital should not lead to higher asset risk and should reduce the contingent liability of the deposit insurance system.

* The options formula for deposit insurance is

$$V_t = DF(X + s\sqrt{t}\ ) - AF(X)$$

where

$$X = (\log (D/A) - (\frac{s^2}{2})t)/s\sqrt{t}$$

and F(·) is the standard normal cumulative density function evaluated from $-\infty$ to (·).

**The exercise price itself is $X = De^{rt}$, the value of deposits at time the bank would be examined. In the put options formula only D appears because the exercise price is multiplied by e^{-rt}, and $Xe^{-rt} = D$.

As a result, regulators also might put controls on asset risk to reduce the liability of the deposit insurance system. For example, regulation might limit a bank to holding less than 30 percent of its assets in security D. However, if regulatory limits on the composition of bank assets and monitoring (examinations) of banks were sufficient to constrain the asset risk of a bank for a given level of leverage, they would be sufficient for any lower level of leverage because banks would have even less of an incentive to evade them. Consequently, as long as regulators did not react to lower leverage (higher capital) by relaxing their efforts to limit asset risk, a bank would not increase its asset risk, and the liability of the insurance fund would decline.[18]

Summary

Not surprisingly, capital regulation is necessary with subsidized deposit insurance to limit the liability of the insurance fund. However, more stringent capital standards for banks do not confound regulatory efforts to limit the riskiness of bank assets because higher capital does not increase the incentives of a value-maximizing bank to hold riskier assets. In fact, the marginal value from increasing asset risk for an insured bank declines as leverage is lowered. This conclusion does not depend solely on the state-preference framework. As discussed in Box 2, a positive relation between leverage and the gains from risk-taking also can be derived from an options approach to evaluating the gains from risk and leverage in banking.

III. Utility Maximization

In this section we incorporate utility maximization into the state-preference model. With utility maximization, the state-preference model implies that capital regulation is either irrelevant because risk-averse owner-managers will hold sufficiently conservative portfolios to make bankruptcy impossible or capital regulation will limit the liability of the deposit insurance system in the same way that it can for value-maximizing banks.

Utility-Maximizing Banks

Utility maximization has been rationalized as being more applicable than value maximization to smaller, owner-managed banks because the investment opportunity set for such banks and their owners may be one and the same. The assumption behind this rationalization is that the owner-manager cannot attract capital in addition to his own and that most of his portfolio is invested in the bank. Consequently, unlike a bank that is maximizing its current market value, the owner's preference toward risk would influence the bank's portfolio decisions.[19]

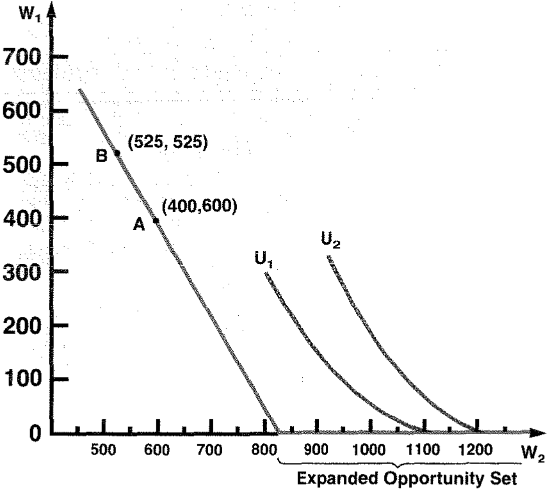
It is assumed that the owner-managers are risk-averse. As pointed out in Box 1, in a simple two-state world, risk aversion means that utility functions are convex with respect to the origin. That is, future wealth has diminishing marginal utility in each state of the world. As is shown in the figure in Box 1, in a world without deposit insurance, an investor will allocate his capital between the available securities to maximize utility along the investment opportunity frontier.

The introduction of underpriced deposit insurance expands the opportunity frontier. The opportunity frontier with free deposit insurance can be derived from the types of calculations presented in Table 1. (Recall there is only one risky security A(4,6), riskless borrowing by issuing security B(1,1), free deposit insurance, and initial bank capital of $500.) Figure 2 shows the various combinations of wealth (W_1 and W_2) that can be attained by increasing leverage and investing in the risky security (security A). Point A in Figure 2 indicates the combination attainable with no leverage, and the shaded segment includes points attainable by increasing leverage.

The key difference between the choice set with deposit insurance and the set without deposit insurance is that there is no limit to the amount of wealth that would be attained in state 2 once bankruptcy would occur in state 1. That is, various wealth outcomes (W_1, W_2), such as (0,895), (0,968), (0,1042), can be attained only with free deposit insurance. With free (or underpriced) deposit insurance and no capital regulation, there is no limit to the amount of wealth that could be attained in state 2 by increasing leverage.

Whether the effects of deposit insurance on the opportunity set will influence the owner-manager's

Figure 2
Subsidized Deposit Insurance Expands the Opportunity Set



leverage decisions depends on the owner-manager's preferences. To benefit from deposit insurance, the owner-manager must be willing to risk bankruptcy — that is, there must be some amount of wealth in state 2 that will compensate for zero wealth in state 1. If there were no such level of wealth in state 2, utility functions would approach the axes asymptotically and interior solutions (some wealth in both states) would be obtained (that is, a point on the frontier to the left of point C). Without the possibility of bankruptcy, the deposit insurance fund would not be at risk, and no capital regulations would be required. Moreover, even if capital regulations were imposed, portfolios that make bankruptcy possible would not be held.

Alternatively, if there were some level of wealth in state 2 that could compensate the bank owner for zero wealth in state 1, the utility functions would intersect the axis. Such a set of preferences is depicted in Figure 2 as indifference curves U_1 and U_2. In Figure 2, raising leverage would increase utility ($U_2 > U_1$) because wealth in state 2 would increase while wealth in state 1 would remain zero. As shown in the previous section, this is the same reason that the current value of the bank increases with leverage. Consequently, maximizing the utility of future wealth in state 2 for the type of preferences depicted in Figure 2 (that is, indifference curves that intersect the axis) is equivalent to maximizing the current value of the bank. Like a value-maximizing bank, a utility-maximizing bank that is willing to accept bankruptcy will maximize leverage.

Capital Requirements and Risk

The predictions of the state-preference model regarding the effect of capital regulations on asset risk for utility-maximizing banks (that will accept bankruptcy) also are similar to those for value-maximizing banks. As just stated, maximizing utility is equivalent to maximizing wealth in state 2 (the nonbankruptcy state) for a bank that has underpriced deposit insurance and will accept a nonzero risk of bankruptcy. Table 2 shows that if leverage were restricted by regulation to 3, wealth in state 2 could be increased by increasing asset risk (holding a larger proportion of assets in security D). Therefore, it could be necessary to regulate even utility-maximizing banks' asset portfolios to prevent such banks from increasing asset risk.[20]

Table 2 also indicates that wealth in state 2 is directly proportional to the current value of the bank when bankruptcy is possible in state 1. (With bankruptcy possible in state 1, wealth in state 2 is equal to the current value of the bank divided by .6.) This means that the marginal effect on wealth in state 2 for a given increase in asset risk declines as leverage declines, just as does the marginal effect on the current value of the bank (see Figure 1).[21] Therefore, as long as the regulatory efforts to prevent a bank from increasing asset risk are not lessened, imposing a lower leverage position would not increase the incentives for a utility-maximizing bank to increase asset risk.

Summary

In sum, incorporating utility maximization into the state-preference model does not affect our conclusion that more stringent leverage requirements will reduce payouts from the deposit insurance fund as long as the stringency of portfolio regulation remains unchanged. Some owner-managers might be so risk-averse that they would be unwilling to risk bankruptcy even with deposit insurance. However, the owner-manager that will risk bankruptcy in one state for a sufficiently high claim in the other state would seek to maximize wealth in the non-bankruptcy state. For such persons, utility maximization and value maximization are comparable and all of the results of the earlier section apply.

IV. Summary and Conclusions

This paper analyzes the theoretical relationships among capital regulation, bank asset risk, and the liability of the federal deposit insurance system. We demonstrated that a bank can benefit from underpriced deposit insurance by increasing leverage and/or asset risk. As a result, some degree of capital regulation is needed to limit the liability of the deposit insurance fund.

More importantly, the analysis suggests that regulatory increases in capital standards will *not* require greater efforts to restrain asset risk. On the contrary, the marginal value of increasing asset risk declines as leverage falls — that is, less leverage (more capital) reduces the gain from risk-taking. In other words, banks with the least capital have the most incentive to increase asset risk.

We have shown under the assumption of value maximization that more stringent capital regulation lowers the contingent liability of the deposit insurance system as long as the stringency of asset portfolio restraints is not reduced. This result follows for value-maximizing banks in both the state-preference and options models. Moreover, incorporating utility maximization into the state-preference model does not change this conclusion.

The key policy implication that stems from our analysis is that regulatory efforts to raise capital standards in banking would not by themselves lead to more risky asset portfolios.

FOOTNOTES

[1] Actions taken in 1981, 1983 and 1985 raised capital requirements for banks and bank holding companies, and made the federal bank regulatory agencies' definitions of capital more uniform.

[2] The proposal for risk-based capital standards was made public in January 1987.

[3] The measure of common equity used in these ratios is not the current regulatory definition of equity capital that includes loan loss reserves and preferred stock.

[4] See, "Fed Urgues Swap Plan for Banks," New York Times, March 5, 1987.

[5] Other studies that consider the effects of capital regulation on bank asset risk within the mean-variance framework are Kahane (1977), Blair and Heggestad (1978), and Hanweck (1984).

[6] Koehn and Santomero (1980), p. 1244.

[7] For purposes of this paper, failure and bankruptcy occur when the market value of a bank's liabilities exceeds that of its assets.

[8] We recognize that the utility maximization model also might be rationalized by appealing to the notion of the separation of ownership and control so that the firm's operating decisions depend on the manager's risk preferences.

[9] Some articles appear to confuse financial capital with physical capital. For example, Santomero and Watson (1977) view financial capital as a physical investment that could have been made in other sectors of the economy. Financial capital, however, is not directly related to physical investment, and higher bank capital does not limit the amount of physical investment in other sectors of the economy. The amount of capital relative to liabilities is simply a reflection of the way a bank finances its assets. Bank capital as well as liabilities are available to be invested in nonbank physical investment through bank loans, for example, as well as in bank facilities.

[10] From an economic standpoint, a bank (or any other firm) fails when the value of its capital falls below zero. Mathematically, the probability of failure is the probability that the value of end-of-period assets is less than that of end-of-period liabilities:

$$\text{Prob [Failure]} = \text{Prob}[(1 + \tilde{p})A < (1 + r)L], \qquad (1)$$

where: $\tilde{p}$ = rate of return on assets (which is assumed to be random),

A = initial assets,

r = promised rate on liabilities,

L = initial value of liabilities, and

Prob [ ] = denotes the probability of [ ].

Without information on the type of probability distribution governing $\tilde{p}A$, the probability of failure can be bounded by using the Tchebichef inequality (see Koehn and Santomero, 1980). However, by assuming that the normal distribution approximates the distribution of $\tilde{p}A$ (i.e., $\tilde{p}A \sim N[E(\tilde{p}A), \sigma^2(\tilde{p}A)]$), we can solve for the probability of failure:

$$\text{Prob [Failure]} = F\left[\frac{L - A + rL - E(\tilde{p}A)}{\sigma(\tilde{p}A)}\right] \qquad (2)$$

where: F = the standard ($\mu = 0$, $\sigma^2 = 1$) unit normal cumulative distribution function.

[11] Equation 2 in footnote 10 indicates that the probability of failure increases as the riskiness of the asset portfolio, $\sigma(\tilde{p}A)$, increases, and as leverage (as reflected in the quantity of liabilities relative to assets) increases. To prove this, the equation can be differentiated with respect to the applicable parameter as follows:

$$\frac{\delta \text{Prob [Failure]}}{\delta\sigma} = \frac{-f(\ )[L - A + rL - E(\tilde{p}A)]}{\sigma^2} > 0, \qquad (1)$$

$$\frac{\delta \text{Prob [Failure]}}{\delta L} = \frac{f(\ )(1 + r)}{\sigma^2} > 0 \qquad (2)$$

where f( ) is the standard normal density function evaluated at the initial position. The first inequality holds because the term in brackets is negative. Also note, in the second inequality, assets are held constant and, thus, an increase in liabilities reflects a corresponding decrease in capital.

12. Other studies such as Merton (1977) and Pyle (1984) provide useful insights into the regulation of bank leverage and asset risk by using options models to analyze the value of the federal deposit guarantee. In Box 2, we analyze the effects of capital standards on asset risk in banking using an options model.

13. This implies a risk-free real interest rate of 5.26 percent, [($1.00/$.95) – 1] x 100%.

14. The analysis would be essentially the same if the premium rates were variable as long as the premiums paid were less than the value of the deposit guarantee.

15. The indifference of a low leverage bank in our example to the degree of leverage parallels Proposition I (the market value of a firm is independent of its capital structure) in Modigliani and Miller (1958). While Modigliani and Miller do not use a state-preference model, Hirshleifer (1966) uses the state-preference approach to show that Proposition I still holds in that framework. In fact, the state-preference model can be used to show that the Modigliani-Miller theorem holds with or without bankruptcy, when there is no subsidized deposit insurance.

16. One policy implication here is that the distortions of deposit insurance could be eliminated if risk in banking were borne only by the banks. Along this line, it has been suggested that risk would not be shifted to the insurance fund if there were timely closures of banks. With continuous (and costless) monitoring of banks, this would correspond to closing a bank before its market net worth reached zero (Furlong and Keeley, 1985). With periodic examinations of banks, the state-preference approach indicates that losses to the insurance fund could be avoided only if banks had enough capital to ensure their solvency in all possible states.

17. In this two-state model, the probability of failure actually would decline only if leverage and risk were restricted in such a way that the bank could meet obligations in each state. In a model with more than two states, the probability of failure would decline with decreased leverage as the number of states in which the bank was able to meet promised payments increased.

18. In this paper, we have not dealt directly with bankruptcy costs. As shown in Dothan and Williams (1980), such costs can lead to a determinate degree of leverage for a value-maximizing *uninsured* bank. However, bankruptcy costs are not sufficient to limit leverage if banks have access to a subsidized deposit insurance. Although bankruptcy costs are not incurred in all future states, they nonetheless can be evaluated in terms of their effects on the current value in the state-preference framework. It can be shown that, with free deposit insurance, bankruptcy costs that are fixed or proportional to assets generally will be insufficient to limit leverage. This result holds whether bankruptcy costs fall on the bank or on the depositors.

19. Despite legitimate questions as to whether this assumption would apply to any real-world banks since owners of small, privately held banks can diversify their portfolios, we hold to it.

20. It is possible that, with leverage held sufficiently low, the wealth attainable in state 2 from investing in the riskier security would not be adequate to compensate a utility maximizing bank owner for risking zero wealth in state 1, even if the utility curves crossed the axis. In such a case, the bank would choose a portfolio along the AC portion of the opportunity frontier in Chart 2, and no other portfolio restraints would be required. However, at some higher level of leverage the same bank would begin to take advantage of the opportunity to increase wealth in state 2 through investing in the riskier asset, security D.

Similar results hold in a model with more than two states. With very low leverage, a bank may not be able to realize sufficient compensation in the nonbankruptcy states to justify risking bankruptcy in even one possible future state. It would not be necessary to regulate the composition of such a bank's assets. At higher levels of leverage, the bank ultimately would increase asset risk and allow for bankruptcy in at least some states.

21. In a multi-state world it also is the case that the marginal effect on wealth in each of the nonbankruptcy states with positive payouts would increase with leverage.

Appendix

Capital Regulation and Asset Risk in a Utility-Maximization, Mean-Variance Framework

Introduction

A number of studies have attempted to analyze the effects of bank capital regulation on asset risk and the probability of bankruptcy while assuming that banks maximize utility in a mean-variance framework. This literature is best typified by articles by Kahane (1977), and Koehn and Santomero (1980).

We show below that the conclusions reached by these studies were derived using internally inconsistent assumptions. Both studies assume that a bank's borrowing cost would be unrelated to bank risk. That is, a bank's borrowing cost (per dollar of liabilities) is assumed to be constant regardless of its asset risk or leverage. Thus, these studies implicitly, but unintentionally, assume that bank failure cannot occur. Yet, they seek to analyze the effects of capital regulation on the probability of bank failure. Moreover, these studies fail to take into account the effect of underpriced deposit insurance on the incentive to take on excessive risk.

A possible explanation for why these studies overlook the effects of bankruptcy on the bank's borrowing cost is that the basic mean-variance framework used is adapted from the finance literature on investment, which does not allow for bankruptcy since borrowing and lending are assumed to take place at the risk-free interest rate. While this simplifying assumption may be appropriate for certain investment decisions, it is not appropriate for the analysis of banking with underpriced deposit insurance. The reason is that concern over the exposure of the deposit insurance system to risk in banking arises only when bankruptcy is possible.

In this Appendix, we first construct a prototype of the utility-maximization, mean-variance model used in past studies to analyze the effects of bank capital regulation on asset risk. We show that when bankruptcy is not possible, and, thus, when there is no deposit insurance subsidy, the results from our prototypical model are identical to those of the previous studies. Specifically, the effect on asset risk of moving from one binding capital constraint to a more stringent one depends on the nature of the preferences of the bank's owner-manager. Restricting such an analysis to situations where bankruptcy is not possible, however, makes these conclusions irrelevant for policy purposes since capital regulation is needed only when bank failures and deposit insurance payouts are possible.

In the next section of the Appendix, we add the possibility of bankruptcy and subsidized deposit insurance to the model. Doing so changes markedly the bank's investment opportunity set. In addition, we present a specific numerical example to illustrate that, when the asset return distribution is binomial, the incentive to increase asset risk does not increase as the stringency of capital regulation increases regardless of the nature of the bank owner-manager's preferences.

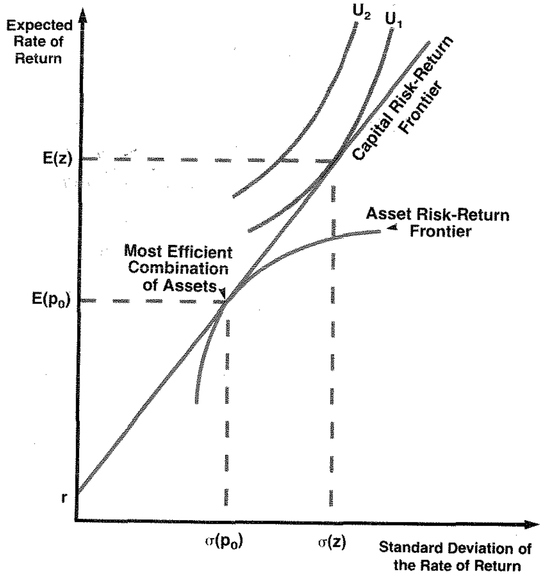
Background

The utility-maximization framework assumes that the bank owner-manager's preferences toward risk can be characterized by the expected rate of return on capital, $E(\tilde{z})$, and the standard deviation of the rate of return $\sigma(\tilde{z})$. Moreover, assuming risk-aversion, these preferences can be depicted by a set of concave, upward sloping iso-utility functions depicting the tradeoff between the expected rate-of-return and risk.

Such iso-utility functions, U_1 and U_2 ($U_i = U[E(\tilde{z}), \sigma(\tilde{z})]$), are depicted in Figure A1. The vertical axis represents the expected rate of return and the horizontal axis represents risk as measured by the standard deviation of the rate of return. Along indifference curve U_1, the investor is indifferent among the various combinations of expected return and risk. However, the investor prefers the points on U_2 to those on U_1 because, for any given level of risk (standard deviation) the expected rate of return on U_2 is larger.

The ideas behind this characterization of preferences are that initial wealth is given and that the

Figure A1
Utility-maximization Framework



investor is concerned about the expected value of future wealth along with its standard deviation. Future wealth is equal to one plus the rate of return times current wealth. Consequently, with current wealth (capital) assumed fixed, the mean and variance of future wealth are mathematically equivalent to the mean and variance of the rate of return on capital, respectively. Thus, similar to the state-preference model, utility maximization in a mean-variance framework also focuses on the probability distribution of future wealth.

Utility-Maximization Without Bankruptcy

A bank must decide on how risky an asset portfolio to hold and by how much to leverage that portfolio. Given the owner's preferences toward risk, utility will be maximized subject to a constraint that relates the expected return on capital, $E(\tilde{z})$, to the standard deviation of the rate of return $\sigma(\tilde{z})$. To derive this constraint, note that if the bank chooses sufficiently low leverage and asset risks to make *bankruptcy impossible* (that is, promised obligations to depositors are always met regardless of the asset return that is realized, the rate of return on capital, $\tilde{z}$, is given by the gross returns on assets, $A\tilde{p}$, minus promised (which equals the actual) payments to liability holders, Lr, divided by initial capital, K, or

$$\tilde{z} = [A\tilde{p} - Lr]/K \qquad \text{(A1)}$$

where

A = initial assets
L = initial liabilities
K = initial capital
$\tilde{p}$ = rate of return on assets, assumed to be random
$\tilde{z}$ = rate of return on capital, which is random
r = promised (which equals actual) rate of return paid on (and cost of) liabilities.

(As discussed more fully later, if bankruptcy were possible, the cost of liabilities would no longer be fixed since actual payments to depositors would sometimes be less than promised payments. This implies that the cost of deposits to the bank would be a random variable, which depends on the rate of return on assets realized and leverage chosen. Consequently, equation A1 would not apply to realizations of $\tilde{p}$ when bankruptcy occurred.)

Equation A1 may be rewritten by noting that $L = A - K$ to give

$$\tilde{z} = [A/K]\,[\tilde{p} - r] + r. \qquad \text{(A2)}$$

The expected rate of return on capital, $E(\tilde{z})$ may be found by taking expected values of both sides of equation A2. This gives:

$$E(\tilde{z}) = [A/K]\,[E(\tilde{p}) - r] + r, \qquad \text{(A3)}$$

as long as r is fixed and not random, which it would be as long as bankruptcy were not possible.[A1] Thus, increasing leverage, as measured by the asset-to-capital ratio increases the owner's expected rate of return on capital linearly as long as default is not possible.[A2] (We later show this result changes when bankruptcy is possible).

Similarly, the standard deviation of the return on capital, $\sigma(\tilde{z})$, may be derived from equation A2 by noting that when bankruptcy is not possible, the covariance of r and $\tilde{p}$ is zero. In this case,

$$\sigma(\tilde{z}) = [A/K]\sigma(\tilde{p}). \qquad \text{(A4)}$$

(This equation is not valid when bankruptcy is possible since the covariance of $\tilde{p}$ and the cost of deposits is not zero.)

Equations A3 and A4 may be jointly solved to eliminate the [A/K] term to give

$$E(\tilde{z}) = [\sigma(\tilde{z}) / \sigma(\tilde{p})][E(\tilde{p}) - r] + r. \qquad (A5)$$

In other words, expected return on capital varies linearly with the standard deviation of return on capital for a given expected asset return and standard deviation of the return.

This linear relationship, equation A5, is plotted as the straight line intersecting the vertical axis at r in Figure A1. It is assumed that the particular asset portfolio with expected return $E(\tilde{p}_0)$ and standard deviation $\sigma(\tilde{p}_0)$ is being leveraged. With no leverage (A = K), the expected rate of return and standard deviation of return on capital are equal to the expected rate of return and standard deviation of return on that particular asset portfolio — $E(\tilde{p}_0)$ and $\sigma(\tilde{p}_0)$, respectively. As leverage increases, the expected rate of return and standard deviation of return both increase linearly.

In general, it is assumed that a bank faces a variety of different asset risk-return combinations as determined by the availability of investment alternatives in its market (known as the asset risk-return frontier). As shown in Figure A1, asset portfolios with more risk are assumed to yield larger expected returns. Also, it is assumed that the banking sector is small enough that the asset risk-return frontier is unaffected by banks' behavior. Thus, that frontier is taken as given by banks in their optimizing decisions.

In this framework, the most efficient asset portfolio is the one where a line from the constant borrowing rate, r, is tangent to the asset risk-return frontier. This is depicted as point $E(\tilde{p}_0)$, $\sigma(\tilde{p}_0)$ in Figure A1. By leveraging this asset portfolio, the bank can obtain the highest expected return on its capital for any degree of risk. Since this tangency point does not depend on the bank owner's preferences, the asset portfolio (that is, the particular combination of assets) chosen depends only on the risk-free interest rate and the asset risk-return frontier. The degree of leverage chosen, however, is determined by the tangency of the owner's iso-utility function with the risk-return frontier for capital (the straight line in Figure A1). However, the assumption here is that the unconstrained bank would choose a degree of leverage for which bankruptcy is not possible.

Capital Requirements and Risk

In Figure A1, we showed how a particular asset portfolio may be leveraged (assuming no bankruptcy) to obtain the capital risk-return frontier. Of course, any asset portfolio may be leveraged although there would be no reason for a bank owner to leverage any asset portfolio other than the most efficient one in a world without capital or asset portfolio regulation. When capital constraints are imposed, however, the bank owner generally will be able to increase utility by leveraging asset portfolios other than the one characterized by the parameters $E(\tilde{p}_0)$, $\sigma(\tilde{p}_0)$. For example, suppose that the maximum asset-to-capital ratio allowed were 5. Then the standard deviation of return on capital would be 5 times the standard deviation of return on the asset portfolio chosen, and the expected return on capital also would be five times greater.

Figure A2
Imposing a Binding Capital Constraint Causes Asset Risk to Rise

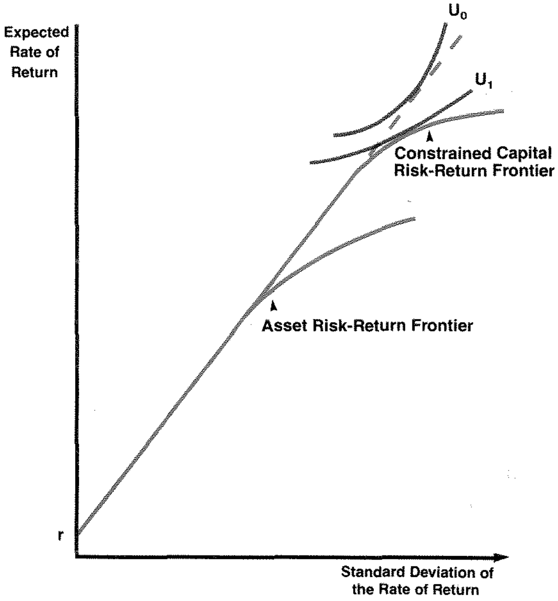
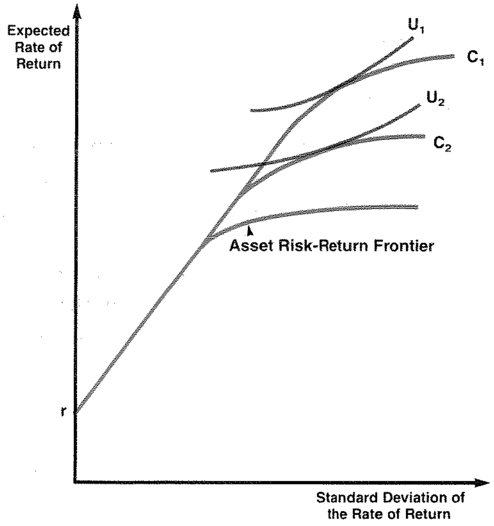


Figure A3
Increasing the Stringency of Capital Regulation



In geometric terms, the risk and return on capital for a given asset portfolio and leverage can be found by extending a ray from the constant borrowing rate through the asset portfolio chosen up to the maximum leverage allowed. As Figure A2 shows, when leverage is limited by regulation, the capital-risk return frontier becomes convex once the leverage constraint becomes binding. As a result, if a binding capital constraint were imposed on a previously unconstrained bank, the bank would choose a more risky asset portfolio. This is shown as a move from U_0 to U_1 in Figure A2.

In Figure A3, two binding capital constraints, C1 and C2, and a particular set of preferences are depicted. However, as one moves from one binding capital constraint, C1, to a more stringent one, C2, the effects on asset risk depend on the owner's risk preferences — asset risk can either increase, decrease, or remain the same, which is the basic conclusion reached by the traditional literature.

Introducing Bankruptcy and Deposit Insurance

The analysis above was derived under the assumption that an unconstrained bank would always make asset and leverage choices such that bankruptcy could not occur. Such a bank could attract deposits at the risk-free rate because it would always make the payments promised regardless of the return on assets realized. Consequently, the cost of deposits to such a bank would be fixed at the risk-free rate r, and not be a random variable.

With free deposit insurance, a bank could issue deposits at a fixed risk-free promised rate even if bankruptcy were possible. However, the cost of deposits to the bank would no longer necessarily equal the risk-free rate. When bankruptcy occurs, the bank effectively would pay less than the promised rate on deposits, r. Only when bankruptcy does not occur does the cost of deposits to the bank equal the risk-free rate. Put another way, the excess of contractual debt obligations over assets when bankruptcy occurs corresponds to the option value of deposit insurance (see Box 2). Thus, the effective deposit cost to the bank is a random variable related to the rate of return on assets and leverage.

As a result, the expected cost of deposits to the bank would decline with increasing leverage and would be less than the risk-free rate. This means that neither the expected rate-of-return equation, A3, nor the standard-deviation of the rate-of-return equation, A4, would hold. Instead, the rate of return on capital, $\tilde{z}$, to a bank with free deposit insurance is given by:

$$\tilde{z} = \begin{cases} [A\tilde{p} - Lr]/K & \text{if bankruptcy does } \textit{not} \text{ occur, that is, if } A(1+\tilde{p}) \geqslant L(1+r). \\ -1 & \text{if bankruptcy does occur, that is, if } A(1+\tilde{p}) < L(1+r). \end{cases} \tag{A6}$$

The expected rate of return on capital, $E(\tilde{z})$, found by taking the expected value of equation A6 is:

$$E(\tilde{z}) = -1 \text{ Prob[Failure]} + E[(A\tilde{p} - rL)/K \mid \tilde{p} > p^*](1 - \text{Prob[Failure]}) \tag{A7}$$

where

$$p^* = -\frac{K(1+r)}{A} + r \tag{A8}$$

is the level of $\tilde{p}$ above which bankruptcy does not occur and

$$\text{Prob[Failure]} \equiv \text{Prob}[\tilde{p} < p^*] \equiv \qquad \text{(A9)}$$

$$\text{Prob}[\tilde{p} < -\frac{K(1+r)}{A} + r].$$

Equations A6 and A7 indicate that the simple linear relationship between $E(\tilde{z})$ and $\sigma(\tilde{z})$ presented in equation A5, generally would not be valid.[A3]

Moreover, Equation A7 indicates that the bank owner would never lose more than his or her initial capital (that is, the minimum $\tilde{z}$ would be -1, even though $\tilde{z}$ would be less than minus 1 if the promised obligation to depositors were met in the event of a bankruptcy). Also, equation A9 indicates that, depending on the asset rate-of-return distribution, the probability of failure can increase up to a point as leverage increases. However, in the limit as leverage increases (and K/A goes to zero), the probability of failure approaches the probability that the rate of return on assets, $\tilde{p}$, is less than the promised rate on deposits, r.

Consequently, by increasing leverage, the owner can increase without limit the expected rate of return on capital as long as at least some part of the asset rate-of-return distribution exceeds the promised deposit rate. Thus, even if the expected rate of return on assets were less than the promised rate on deposits, a bank with underpriced deposit insurance would gain from leverage as long as this condition held. This conclusion contrasts with the results obtained when bankruptcy is not possible. In that

Table A1

Expected Return and Standard Deviation of Return on an Initial $100 Investment of Capital as Leverage Increases

(1) Leverage	(2) Assets	(3) Deposits	State	(4) End-of-Period Assets	(5) Actual Deposit Payments	(6) End-of-Period Capital (4 - 5)	(7) Deposit Insurance Payments	(8) Expected End-of-Period Capital	(9) Standard Deviation of End-of-Period Capital
1	100	0	1	110	0	110	0	105	5
			2	100	0	100	0		
2	200	100	1	220	104	116	0	106	10
			2	200	104	96	0		
3	300	200	1	330	208	122	0	107	15
			2	300	208	92	0		
4	400	300	1	440	312	128	0	108	20
			2	400	312	88	0		
⋮									
26	2,600	2,500	1	2,860	2,600	260	0	130	130
			2	2,600	2,600	0	0		
27	2,700	2,600	1	2,970	2,704	266	0	133	133
			2	2,700	2,700	0	4		
28	2,800	2,700	1	3,080	2,808	272	0	136	136
			2	2,800	2,800	0	8		
⋮									
100	10,000	9,900	1	11,000	10,296	704	0	352	352
			2	10,000	10,000	0	296		

Asset return = .5 probability of .10 (state 1)
.5 probability of .00 (state 2)

Deposit rate = .04
Initial Capital = 100

case, leverage can increase expected return only if the expected return on assets exceeds the promised rate.

A Numerical Example

Below, a simple numerical example is presented that allows for both subsidized deposit insurance and the possibility of bankruptcy. This example shows that a bank can gain from subsidized deposit insurance only by assuming a nonzero risk of bankruptcy. The example also shows that when asset returns are distributed binomially, more stringent capital regulation will not increase the incentive to increase asset risk. Thus, this example demonstrates that the results in the cited mean-variance literature do not hold generally.

The calculations presented in Table A1 demonstrate what happens to the expected return and standard deviation of return on an initial $100 investment of capital in a bank (with deposit insurance provided to it at no cost by the government) as leverage increases. To simplify the calculations, we assume that the rate of return on the asset being leveraged is drawn from a binomial probability distribution with a .5 probability of a 10 percent rate of return and a .5 probability of a 0 rate of return (for an expected rate of return of 5 percent). It is assumed that the bank is able to attract any quantity of deposits it wants at a promised 4 percent rate of return because deposit insurance (which is provided free to the bank) covers any shortfalls when bankruptcy occurs.

The calculations in the Table demonstrate what happens to expected end-of-period capital and its standard deviation as leverage increases. (The rate of return on capital, in percentage terms, is just end-of-period capital minus 100). In line 2, the bank increases leverage to 2 by issuing $100 of deposits and promising to return $104 to depositors at the end of the period. The end-of-period capital for each of the two possible asset returns after paying the deposit claims is shown in column 6. The expected end-of-period capital and standard deviation of end-of-period capital are shown in columns 8 and 9. The payment from the deposit insurance fund is shown in column 7.

As the Table indicates, initially, as leverage increases up to 26, the expected return and standard deviation of return both increase linearly and there is no bankruptcy. After this point (when leverage exceeds 26), there is a 50 percent chance of realizing the low asset return (denoted as outcome 2) and going bankrupt. However, there is also a 50 percent chance of drawing the high asset return and profiting from leverage. As leverage increases, end-of-period capital increases without limit as long as bankruptcy does not occur.[A4]

Rationality implies that a person will prefer a lottery that pays $100 with a 50 percent chance and $0 with a 50 percent chance to one that pays $10 with a 50 percent chance and $0 with a 50 percent chance.[A5] Thus, this example illustrates that even a risk averse bank owner that is willing to risk bankruptcy (that is, one who is willing to participate in the type of lottery just described) in return for a sufficiently high payoff when the higher asset return is realized would prefer unlimited leverage. Consequently, in this example, maximizing utility is equivalent to maximizing value as long as the bank owner is willing to risk bankruptcy.

In this simple model, a similar result would apply to asset risk under leverage constraints. As long as some non-zero probability of bankruptcy were acceptable, a bank owner would maximize asset risk since that would maximize end-of-period capital if bankruptcy did not occur and would not affect end-of-period capital if bankruptcy did occur. These implications are in sharp contrast to those from the cited mean-variance literature which claims that risk aversion would limit leverage and asset risk.

Moreover, as in the state preference model presented in this paper, it can be shown that the gain from increasing asset risk is positively related to leverage. Thus, in the case of binomially distributed asset returns, more stringent capital regulation does not increase the incentive to increase asset risk.

Summary

The results of previous studies using the mean-variance framework regarding the effect of capital regulation on asset risk can be replicated assuming that bankruptcy is not possible. However, when bankruptcy is possible and underpriced deposit insurance is provided to banks, the results of these studies no longer hold generally.

APPENDIX ENDNOTES

A1. An uninsured bank could attract deposits by paying a fixed rate of interest, independent of its leverage or asset risk as long as its asset risk were low enough relative to capital that the probability of bankruptcy were zero. As long as bankruptcy were not possible, bank liability holders would not be at risk of loss due to default and would accept bank liabilities as riskless. At some point, however, as leverage increased (for a given nonzero asset risk), bankruptcy would become possible and the bank would have to pay a higher deposit rate to compensate depositors for the risk of default.

The utility-maximization literature cited assumes a constant borrowing rate environment, but does not explicitly acknowledge that this would be consistent only with a zero probability of bankruptcy. Kahane does allow for a stochastic deposit rate but assumes the promised rate equals the rate the bank expects to pay. Moreover, he assumes the expected cost of deposits and the promised rate are independent of leverage and asset risk. These assumptions would hold only if bankruptcy were not possible.

A2. Thus, it is crucial to distinguish the asset from the capital risk-return frontier. Blair and Heggestad (1978) fail to do so.

A3. However, we do not mean to imply that equations A6 and A7 necessarily can be used to derive the appropriate risk-return constraint for utility maximization in a mean-variance framework. One reason is that variance no longer adequately characterizes risk when bankruptcy is possible.

A4. After the point where bankruptcy becomes possible, the relationship between the expected rate of return on capital (column 8 minus 100) and its standard deviation changes (the expected rate of return rises more rapidly and the standard deviation rises less rapidly with leverage.)

A5. This is true even though the standard deviation of the first alternative is larger.

REFERENCES

Arrow, Kenneth J. "The Role of Securities in the Optimal Allocation of Risk-bearing," *Review of Economic Studies,* April 1964.

Blair, Roger D. and Arnold A. Heggestad. "Bank Portfolio Regulation and the Probability of Bank Failure," *Journal of Money Credit and Banking,* Vol. 10, No. 1, February 1978.

Debreu, Gerard. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium,* New York: John Wiley & Sons, 1959.

Dothan, Uri and Joseph Williams. "Banks, Bankruptcy, and Public Regulation," *Journal of Banking and Finance,* No. 4, 1980.

Furlong, Frederick T., and Michael C. Keeley. "Bank Runs," *Weekly Letter,* Federal Reserve Bank of San Francisco, July 5, 1986.

Hanweck, Gerald A. "A Theoretical Comparison of Bank Capital Adequacy Requirements and Risk-Related Deposit Insurance Premia," Research Papers in Banking and Finance, Board of Governors of the Federal Reserve System, No. 72, May 1985.

Hirshleifer, Jack. "Investment Decisions Under Uncertainty: Applications of the State Preference Approach," *Quarterly Journal of Economics,* Vol. 80, May 1966.

Kahane, Yehuda. "Capital Adequacy and the Regulation of Financial Intermediaries," *Journal of Banking and Finance,* No. 1, 1977.

Kareken, John H. and Neil Wallace. "Deposit Insurance and Bank Regulation: A Partial Equilibrium Exposition," *Journal of Business,* Vol. 51, No. 3, 1978.

Koehn, Michael and Anthony M. Santomero. "Regulation of Bank Capital and Portfolio Risk," *The Journal of Finance,* Vol. XXXV., No. 5, December 1980.

Merton, Robert C. "An Analytic Derivation of the Cost of Deposit Insurance and Loan Guarantees," *Journal of Banking and Finance,* No. 1, 1977.

Modigliani, Franco, and Merton H. Miller. "The Cost of Capital, Corporation Finance and the Theory of Investment," *The American Economic Review,* Vol. 48, June 1958.

Pyle, David H. "Deregulation and Deposit Insurance Reform," *Economic Review,* Federal Reserve Bank of San Francisco, Spring 1984.

Santomero, Anthony M. and Ronald D. Watson. "Determining an Optimal Capital Standard for the Banking Industry," *The Journal of Finance,* Vol. XXXII, No. 4, Sept. 1977.

Sharpe, William F. "Bank Capital Adequacy, Deposit Insurance and Security Values," *Journal of Financial and Quantitative Analysis,* Nov. 1978.

Sharpe, William F. *Portfolio Theory and Capital Markets.* New York: McGraw-Hill, 1970.

Competition in the Semiconductor Industry

Randall Johnston Pozdena*

The semiconductor industry has played a key role in international trade disputes. Using data on Dynamic Random Access Memory devices and semiconductor chip fabrication facilities, this analysis examines the behavior of the industry for evidence of the influence of time-related technological change, economies of scale, learning curve behavior, and international differences in strategic pricing behavior. The analysis finds only weak evidence of anti-competitive behavior.

The pace of innovation in the field of electronics has been extremely rapid in the last thirty years, and high technology electronics has been a major source of strength for the American economy. The development of solid state devices — and integrated circuits in particular — has been the major contributor to the startling evolution of this field and the entry of high technology electronics into so many aspects of daily life. In addition, many place their hopes for continued growth of the national and regional economies on intensified innovation in and application of high technology electronics. Along with biotechnology, high technology electronics is seen as a kingpin of the future of the American economy.

The purpose of this paper is to evaluate popular claims that the semiconductor industry is susceptible to anticompetitive behavior, particularly on the part of foreign competitors. Specifically, we will examine the market for a particular integrated circuit (IC) device for evidence of imperfectly competitive performance. Learning and scale economies are found to be significant in this industry, and market structure — while not showing excessive concentration of market share — exhibited rigidity. Combined, these observations are consistent with what one would find where inefficient forms of strategic pricing behavior are practiced.

Production functions associated with integrated circuit fabrication facilities located in the United States and Japan are estimated to provide an insight into the origins of alleged international differences in pricing strategies. Only weak evidence is found to support the notion that Japanese integrated circuit (IC) fabrication costs are below those of their U.S. counterparts.

In Section I of this paper, a brief description of the semiconductor industry and its products is presented. Section II contains a description of the IC production process and an economic characterization of this process. Section III discusses potential implications of the production environment on industry structure and performance. In Section IV, several simple empirical investigations are performed to assess the importance of learning and scale economies in the IC industry and to investigate the origins of differences in U.S. and foreign firm pricing strategies. The paper concludes with a summary of findings and their implications for the future of the U.S. semiconductor industry.

* Assistant Vice President, Federal Reserve Bank of San Francisco. This paper benefitted from the valuable assistance of numerous people. I wish particularly to thank Dan Hutcheson of VLSI, Inc., Ione Ishii of the Semiconductor Industry Association and Lane Mason of Dataquest, Inc. for their generosity in making available much of the data used in this study. The paper also benefitted from the assistance provided by Bill Arnold of the Electronic Materials Report, Inc., Don Larsen of Tektronics, Inc., and Carolyn Rogers of Hambrecht and Quist, Inc. Finally, I wish to thank William M. Robertson for providing excellent and diligent research assistance.

I. The Industry and Its Products

The history of the semiconductor industry, its technology and products are discussed in a number of published sources[1]. It is useful, however, to review the basic features of the industry and its technology both to support the logic of subsequent discussion and to delimit the economic issues we will address.

The semiconductor industry is so named because it produces devices that exploit the special electrical characteristics of a class of natural elements and compounds known as "semiconductors" (such as silicon, germanium and gallium arsenide). The materials have the property that they can be made to behave alternately as conductors or barriers to the flow of electrical current. In the late 1940s, discovery of a means of managing the behavior of semiconductor crystals led to the development of the transistor — a device that uses small currents to control the conduction behavior of the material. Thus, the transistor can form the basis of an amplifier or electronic switch.

Through the 1950s and 1960s, the transistor rapidly replaced the vacuum tube because of its superior ruggedness, smaller size, lower power consumption, and ability to execute tasks more rapidly. Before 1958, functional electronic devices were built by connecting a number of transistors and other electronic components in a discrete manner. Then, two scientists developed the "integrated circuit" or IC, which is a single device combining the functions of a number of transistors. By so doing, ICs opened the possibility of constructing more efficient and compact electronic devices.

The first ICs were produced in commercial volume in the mid-1960s. They are produced by a complex process of etching, "doping" the crystalline material with other elements, and heat-treating the surface of a semiconductor crystal wafer. Today, a 5-inch diameter wafer of silicon can yield one hundred or more "chips", each of which may contain as many as 1 million transistors. Although "discrete" devices are still produced, the IC is now the dominant semiconductor product and has revolutionized industrial and consumer electronic products. In 1985, approximately $16.5 billion in IC shipments were made worldwide, against about $5 billion in shipments of discrete devices[2].

Despite wide variation in the types of functions that ICs can perform, the same basic production process is used in their manufacture. Microprocessors (the "brain" of computational devices), memory devices (for storing information), and a wide variety of standard circuits used in consumer electronics, telecommunications devices, and military hardware all involve similar production procedures[3]. By focusing our attention on ICs in general, and memory devices in particular later in the paper, we hope to make useful generalizations about the semiconductor industry.

II. The Economics of IC Production

The focus of this paper is on factors influencing the structure and future international competitiveness of the American semiconductor industry. We begin with a brief description of the IC production process. Certain aspects of this process are unusual and, when considered in light of U.S. patent law and the alleged industrial policies of foreign competitors, may be important determinants of the structure, performance, and international competitiveness of the American semiconductor industry.

Major Features of the Production Process

The production of integrated circuits involves very large pre-production investment. Such investment takes the form of circuit layout development, development of "maskworks" or templates to imbed the circuitry in the surface of the semiconductor material, and development and testing of prototypes. Because the prototypes often do not behave as modelled during the layout development process, many cycles of the prototype development process may be required before a useful design evolves. This basic circuit design process interacts with the design of the fabrication process and, in some cases, with the design of other chips or "firmware" (programming incorporated into ICs). In total, this preproduction investment may cost as much as $100 million in the case of a new microprocessor chip[4].

Actual fabrication of the integrated circuits takes place in a fabrication line ("fab line") facility. Wafers of the semiconductor crystal (predominantly silicon) enter one end of the fab line and the fabricated IC exits the production process after various stages of chemical and heat treatment, "dicing" of the wafer into constituent chips, and electrical and physical attachment of the chip to its plug-like base.

It is conventional to describe the capacity and activity levels on a fab line in terms of "wafer starts" per week. The relationship between wafer starts and actual production flow of ICs, however, will depend upon the design of the device being fabricated, the size of the wafer stock, and the efficiency of the fabrication process, which generally is higher on lines with newer vintage fabrication equipment and higher quality labor.

Labor and capital are substitutable to some degree in most of the steps of the fabrication process. Once a fabrication process has been configured, however, significant changes in the process can be costly and time-consuming. Similarly, although a single fab line can, within limits, be used to produce a variety of devices, different types of devices involve different processing steps and sequences, new computer programs to guide those steps, and can involve changes in the degree of cleanliness of the fab line environment. Crossovers to radically different devices, therefore, also are costly[5].

Short-Run and Long-Run Costs of Production

The characterization of IC products and the production process made above can be re-stated in conventional economic terms as follows. First, the product in the IC industry is probably best thought of not as the IC itself, but rather, the units of memory storage, switching, or logical processing functions it provides. Although there are qualitative differences across IC devices providing these various functions (such as access speed in memory devices or the compactness of the IC device that contains them), it is helpful to think of the market as demanding memory storage or other functions rather than ICs *per se*. Then, within gross functional categories at least, the elemental unit of output relates to the fundamental electronic building block of the IC, namely the transistor.

In the short run, fab line capital and the capital representing the design of the IC (the maskworks) are fixed. Output is varied by the firm by manipulating labor and materials inputs. It seems clear that average total short-run costs decline sharply with increased output because of large, fixed maskwork and fab line capital costs. At production levels above the design capacity of a firm's fab line facilities, however, problems of congestion likely arise. Each of the 50 to 100 processing steps takes a finite amount of time and few opportunities exist in the short run to accelerate the processing or to improve the yield of useful output from wafer starts[6]. Thus, in the short run, rising average variable costs likely cause average total costs to rise at high output levels.

In the long run, both fab line capital and maskwork capital are variable, and there are several potential sources of increasing returns to scale. One is that larger fab line facilities offer lower unit fabrication costs than smaller ones. The industry's practice, however, has been to manipulate the number rather than the size of fab lines to alter fab line capacity, suggesting that individual fab line scale is not a major source of economies of scale generally. Of the 1,500 or so fab lines in existence in 1986, two-thirds had design capacities between 1,500 and 4,300 wafer starts per week[7]. If fabrication were an important source of scale economies, its effects, therefore, must be derived from firm-level synergies from operating multiple lines. (The issue of fabrication scale economies is explored further below.)

Increases in the firm's stock of "maskwork capital" also could result in lower long-run average costs. Conceptually, we might view improvements in maskworks and manipulation of processing steps (that is, alteration in the design of the IC) as either an increase in the employment of maskwork "capital" or a change in technology. Technological change is usually assumed to be exogenous to the firm's labor and capital allocation decisions (that is, technical change depends only upon the passage of time) whereas investments in what we are calling "maskwork capital" have been an important component of IC firms' cost-minimization strategy.

Indeed, the commitment of resources to chip (and fabrication process) design is probably responsible for most of the widely touted, sharp declines in IC product costs that have been observed over time. IC engineers have succeeded in increasing the number of elemental components ("transistors") that can be accommodated by a single semiconductor chip of given physical dimension[8] and, hence, reducing the unit cost of fabricating IC products. (Empirical evidence will be presented in Section IV below on the relative contribution of scale economies and the passage of time to the decline in the cost of IC memory products.)

Technological Diffusion and Learning

Two other aspects of the IC production environment are relevant to understanding the current and likely future performance of the IC industry. The first is that property rights in "maskwork capital" historically have been poorly defined, making it difficult for one firm to prevent access to the fruits of its investment by other firms. It is relatively easy to reconstruct the design and manufacturing steps involved in an IC product through a process known as "reverse engineering"[9]. By a sequence of photographic analysis, disassembly by etching, and materials analysis, a rival firm can reconstruct the architecture of a functional chip and the maskworks and processing steps necessary to reproduce it. Such "reverse engineering" can cost as little as one-one thousandth of the original firm's investment[10] and permit "pirate" firms to enjoy lower total costs. The passage of the Semiconductor Chip Protection Act of 1984 foreclosed the possibility of precise "cloning" of maskwork capital by foreign or domestic competitors, although the more general practice of reverse engineering remains legal[11].

A second often-cited feature of the IC industry is the relevance of "learning curve" phenomena to IC production. The notion is simply that the cost of production may be related not only to the rate of output of a firm (economies of scale) and changes in technology over time but also to the independent effect of accumulated production experience. Such a phenomenon is considered to be relevant to complex manufacturing technologies: as output experience increases, the firm better understands the technology involved and technical efficiency increases[12].

Since integrated circuit manufacturing is an extremely complicated technical process, it seems likely *a priori* that learning-related cost adjustments may occur. The implications of learning phenomena and a test for their existence are presented below.

To summarize, the IC production process is a highly technical one involving large investments in maskwork capital that are difficult to recover if an enterprise fails and difficult to protect from exploitation by other firms. Potentially significant economies of scale are likely, and probably flow mainly from economies at the level of the firm rather than the plant (that is, the fab line). Costs also may decline over time because of technological progress and with accumulated output experience because of "learning curve" phenomena.

III. Implications for Industry Structure and Performance

The preceding discussion of the economic characteristics of the IC industry may help explain the likely structure and behavior of the industry, particularly whether the industry exhibits characteristics that may make it vulnerable to anticompetitive behavior. It has been widely alleged, for example, that Japanese producers have pursued predatory pricing strategies in certain IC products[13]. In this section, the implications of the postulated economic characteristics for structure and performance are discussed as a prelude to attempts at empirical verification.

Scale Economies and Contestability

The economic characteristics of the IC industry make it likely that the production of ICs is characterized by economies of scale. Significant scale economies, in turn, would mean that markets for IC products will tend to be concentrated in the long run, and thus have the potential for inefficiency.

Baumol and Bailey[14], however, have argued that high levels of concentration (or even monopoly) in production need not have serious effects on market efficiency if the market were "contestable". For an industry to be considered contestable (in the sense that Bailey and Baumol use the term), it must be possible for new firms to enter a market displaying abnormal profits and earn normal profits or — if extant firms cut prices to thwart the new entry — leave the industry without losing the investment associated with entry.

At least one attribute of the IC industry suggests that it may not be ideally contestable: a major cost of entry — preproduction research and development — is difficult to recover if the firm is unsuccessful in competing against extant producers and must exit the market. In this respect, the IC industry contrasts with most manufacturing, transportation and service industries for which acquisition of re-sellable, fixed assets is the dominant cost of entering a market. Thus, "contesting" for markets may not be an effective means of imposing competitive discipline within the IC industry. This makes it especially important to investigate that industry's scale economies.

Learning and IC Market Efficiency

In Section I, we also postulated that IC production occurs in the presence of a learning curve. The existence of learning effects on unit production costs may have a bearing on both industry structure and pricing behavior and, thereby, on the efficiency of the IC industry. The logic of these effects in a model of dynamic entry and pricing behavior has been demonstrated rigorously by Spence[15]. The implications of his model will only be summarized briefly here.

First, if learning (production experience) reduces costs, Spence has demonstrated that, under certain theoretical conditions, learning can confer some protection from competitive entry to the first firm into a market, in effect, simulating an entry barrier. If "first movers" do enjoy such advantages in the IC industry, then market structure might be expected to be rigid over time — that is, show little change in the rank and share of firms in the market.

The second aspect of learning curve theory of interest here are the effects of the learning curve on pricing behavior. In essence, because production experience confers subsequent cost advantages on the firm, a firm maximizing long run profits in the presence of a learning curve will charge less (and produce more) in the learning stage of production than dictated by short run profit maximization considerations alone.

Not all environments are conducive to such learning curve pricing behavior, however. Whereas firms in an industry composed of just a few firms are able to exploit a learning-curve pricing strategy, Spence argues that strategy is less effective in unconcentrated production environments where there is assertion of competitive price discipline with successful entry.

In addition, whether learning is important at all to either pricing behavior or market structure depends upon how "rapidly" learning takes place. If learning were very rapid (that is, the effects of accumulated production experience are small relative to the effects of current production on costs), then there would be few strategic advantages to deviating from short run profit maximization. A related point concerns how rapidly learning reaches other firms. If such diffusion is very rapid, accumulated output might help explain cost and price trends for the industry as a whole, but current prices would be determined by current costs.

The implications of Spence's view for the IC industry might be summarized as follows. *If* learning (cumulative output) were important to firms in the IC industry, then early entrants able to survive initial periods of low prices might gain an (at least temporary) advantage over later entrants. Market structure would be less fluid than otherwise, and such "first movers" could enjoy higher profits than subsequent rivals. Firms, in turn, would have a strategic incentive to pursue an early-entry strategy.

International Competition in the IC Industry

For American IC producers, concerns over preproduction costs and scale economies, contestability, and learning phenomena seem to be at the root of current debates over the marketing strategies of their Japanese competitors. Japanese producers have gained a growing share of the world semiconductor market; their share of combined U.S.-Japanese production has risen from 33 percent in 1971 to over 50 percent in 1982. Since 1982, the U.S. share of world IC sales has fallen from about 60 percent to 50 percent, while the Japanese share has risen from 30 to about 40 percent[15a].

It is frequently alleged that the Japanese have obtained their growing share of IC sales by pursuing "predatory" pricing strategies. In particular, Japanese IC manufacturers have been accused of selling IC products in world markets at prices below their cost of production. In 1986, for example, there were three major International Trade Commission complaints alleging such behavior filed by the American IC industry and the U.S. government[16].

If true, one explanation for such pricing behavior would be the existence of exploitable learning curve advantages, although any IC firm — not only the Japanese ones — could exploit those advantages. Nevertheless, it is argued frequently that Japanese IC producers are better able to survive the early periods of low profitability necessary to secure market dominance in a "learning" dominated production environment. They are alleged to benefit from their affiliation with conglomerate manufacturing organizations, which underwrite early periods of low profitability, and the availability of subsidies from the Ministry of International Trade and Industry (MITI) and the banking industry[17]. Whether such subsidization occurs (or differs dramatically from support the U.S. IC industry has received from the military) has been debated extensively[18].

A second alleged reason for the growth in Japanese IC market share is that low technological and legal barriers to copying U.S. designs and processes have unfairly reduced the total costs faced by Japanese producers. Particularly egregious cases of apparent cloning indeed can be documented[19]. In addition, the property rights traditionally extended by the Japanese to foreign creators of intellectual property have been criticized as being weak by international standards. In the debate over software copyright reform from 1983 to 1985, for example, the Japanese proposed standards of protection were weaker than both international copyright standards and the standards applied to domestic (Japanese) copyrights[20].

As Dasgupta and Stiglitz[21] have pointed out, ill-defined intellectual property rights can reduce innovation below the socially optimal level. However, for low barriers to cloning to have a permanent effect favoring Japanese over U.S. production, the ability to "reverse engineer" a competitor's product must be asymmetric internationally which it is not, and some other factor must operate to "cement" market dominance once dominance is achieved by this means.

Finally, it is possible that growth in the Japanese IC market share flows from legitimate differences in fabrication cost. These differences could arise from lower costs for labor of a given quality or superior Japanese management of fabrication facilities. (It is unlikely that differences in materials or equipment costs would be as important since most of the wafer stock and fab line equipment has been manufactured by one country: the U.S.) The theory of international factor price equalization[22] argues against the lower labor cost argument, but foreign producers could still have the comparative advantage if they have a greater endowment of relevant production factors[23].

IV. Empirical Examination of the IC Industry

Several empirical investigations may inform our understanding of the structure and performance of the IC industry. First, a study of market share rigidity may shed light on the structure of the IC market. Market shares that appear to be rigid over time might indicate that the market is not easily contestable, or that learning phenomena operate to retard entry.

Second, we could test directly for the existence of learning phenomena by examining the response of costs to cumulative firm output — costs should decline with accumulated output experience. In addition, if a high degree of market concentration were associated with *lower* IC prices (an association not normally expected except in the presence of a learning curve or other strategic pricing considerations), learning in the IC industry would be more likely to be firm-specific (that is, it would not diffuse so rapidly that it did not influence firm behavior).

Third, the relationship between the scale of production and cost also would be of interest. If scale economies were not extremely great, the industry would be less likely to be concentrated, and the potential distortions caused by lack of contestability or other constraints on the fluidity of the industry would be less important.

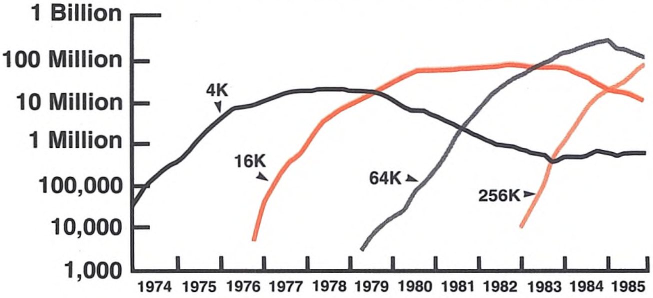
Finally, to address concerns about the behavior of foreign competitors, it would be helpful to compare fabrication costs in Japanese and U.S. facilities. If Japanese IC prices were below their American counterparts' but fabrication costs were the same, we might have evidence that Japanese producers were pricing below the full cost of production, including the cost of maskwork capital.

Unfortunately, the data available on the IC industry are not ideal for examining all of these relationships. No cost data are available and they would be suspect in any case, since the cost of producing a specific product is difficult to extract in a conglomerate enterprise. This is a particularly serious shortcoming in studying Japanese IC production costs. Fairly good price and output data are available by firm and device, however, as are data on the labor and capital employed on individual fabrication lines. In what follows, these data are exploited to provide rough information on the relationships of interest.

The Behavior of the DRAM Industry

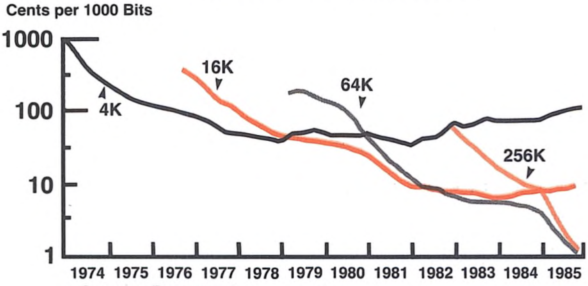
We explore here the issues of market structure rigidity, learning effects, and scale economies in the context of a particular type of IC device — the Dynamic Random Access Memory (DRAM) IC. This device stores binary bits of information in a randomly accessible manner. (The term "dynamic" simply refers to the requirement that DRAMs be powered continuously to retain implanted memory. The term distinguishes them from a related device — the Static RAM — that does not need continuous electrical power.) The memory capacity of DRAMs is measured in kilobits; each kilobit is 1,024 individual bits of memory capacity and is abbreviated by a "K". To date, DRAM devices have been

Chart 1A
Number of DRAM Devices Sold by Quarter



Source: Dataquest

Chart 1B
Time Trend in DRAM Prices



Source: Dataquest

manufactured in cómmercial volume in 4K, 16K, 64K and 256K capacities.

We focus on DRAMs for a number of reasons even though they represented only about 10 percent of total IC sales worldwide in 1985. First, the DRAM device is as close as the semiconductor industry gets to a "commodity"-type of device. Most other ICs have qualitative attributes that make them difficult to study over time or across firms. Second, unlike microprocessor ICs for example, DRAMs have been produced in significant volumes by non-U.S. firms, allowing some exploration of the influence of foreign entry on industry behavior. Indeed, DRAMs were involved in recent allegations of "dumping" by the Japanese[24]. Finally, as a practical matter, to expand sample sizes, it is necessary to combine data across devices. Such a combination is feasible with memory devices because they are unambiguously "generic" in their essential unit of service (the "bit"), and bits are substitutable across devices.

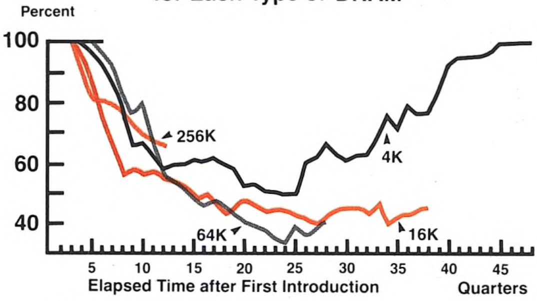
Substitutability across types of DRAMs is illustrated in Chart 1, which shows the actual quantity shipped in Panel A and actual prices per bit for four DRAM devices in Panel B. The sales of the 4K DRAM, for example, peak and decline sharply (note that all quantities are in log terms) when the price per bit of the successor device (16K DRAM) falls below the 4K price per bit. A similar pattern holds for subsequent generations of devices. The chart also illustrates vividly the observation made earlier that increasing the bit density on the chip has contributed importantly to the observed declines in price per bit of DRAM memory.

Market Rigidity

In examining the DRAM market for evidence of structural rigidity, it is instructive to trace the evolution of market structure in DRAM manufacture. As Chart 2 reveals, a new DRAM device typically is introduced by one or two firms with entry occurring gradually until concentration (as measured by the share of the market held by the largest 3 firms) declines to a relatively modest level.

In the cases of the early devices (such as the 4K DRAM), entry occurred more gradually than with subsequent generations of devices, and concentration levels did not decline below the 50 percent level. These findings probably are consistent with the existence of important scale economies and, therefore, limited "room" in the market for additional firms. Indeed, as the market for 4K devices matured and declined, the concentration ratio for 4K DRAMS gradually increased as firms exited the market.

Chart 2
3-Firm Concentration Ratios
for Each Type of DRAM

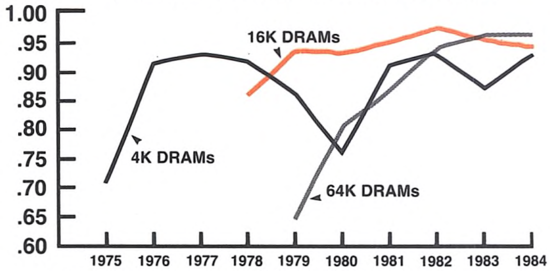


Source: Dataquest, FRBSF

Although firm entry into the market for DRAMs appears capable of reducing levels of industry concentration for at least short periods of time, new entrants have difficulty dislodging "first movers". To illustrate this observation, the orderings of firms ranked by market share from one year to the next were compared and a statistic measuring the correlation of these ranks — the Spearman Rank Correlation Coefficient — was computed for each pair of adjacent years and for each device. (Unchanged year-to-year rank ordering produces a Spearman rank correlation coefficient of 1.0)[25]. The results are presented in Chart 3.

The high correlation of firm market share rankings from one year to the next suggests that DRAM market structure is not highly fluid after the initial period of entry, and has with each successive DRAM device reached this condition of structural stability more rapidly. Moreover, data not presented in the chart suggest that the first producers of a device not only retain pre-eminence in the market for that device but often are "first movers" into production of the next generation device. Turnover of producers is greatest among those firms that are not first entrants. These observations may suggest the existence of high fixed costs (either pre-produc-

Chart 3
Time Trend in Market Structure
Rigidity as Measured by the
Spearman Correlation of Market Share



Source: Dataquest, FRBSF

tion or production) or learning phenomena that benefit incumbent firms.

Scale Economies and Learning Effects

Without specific data on IC production costs, it is not possible to test directly for the existence and importance of learning in DRAM manufacture, or to explore directly the magnitude of scale economies. However, an examination of the behavior of DRAM *prices* in addition to the inferences drawn above from the behavior of market structure may shed some light on cost behavior. In particular, except in instances of coordinated or monopoly pricing, prices and average costs are likely to move together over time. This relationship suggests that some inferences about costs can be derived from price data if the circumstances that might lead to noncompetitive pricing can be controlled.

To explore these relationships more formally (and to control for the effects of the passage of time, accumulated production experience, and market structure), we studied a simple econometric relationship using quarterly data on DRAM devices over the 1976 to 1985 period. We studied the variable the price of a *bit* of DRAM, and we pooled time series data on four devices (the 4K, 16K, 64K and 256K DRAM) to expand the sample size. The general form of the relationship studied was:

$$\begin{aligned}\text{Price}(x,t) = h[&\text{time}(x,t), \text{output}(x,t), \\ &\text{cumulated output}(x,t), \\ &\text{market structure}(x,t), \\ &\text{Japanese market share}(x,t) \text{ and} \\ &\text{size}(x,t)]\end{aligned} \qquad (1)$$

where

price(x,t) = price per bit, in U.S. dollars for device x at time t

time(x,t) = t = the date of the price observation on device x

output(x,t) = average industrywide output per firm (in number of devices), for device x at time t

cumulative output(x,t) = average device output per firm summed over time by device, for device x at time t

market structure(x,t) = 3-firm concentration ratio in market shares of device x at time t

Japanese market share(x,t) = fraction of total production by Japanese firms, for device x at time t

size(x,t) = the memory capacity in bits for device x at time t

and

$x = 1, 2, 3, 4$

$t = 1$ to 44 measured from 1974, quarter one

Table 1 presents a regression analysis of a specific configuration of Equation 1. In particular, ordinary least squares were used to estimate a relationship between the log of DRAM prices and the variables identified in Equation 1. Most of the coefficients are made a function of device size by interacting each variable with the device size measure. All of the coefficients in Table 1 are measured with considerable statistical precision, permitting a number of interesting observations.

First, the positive coefficient on the Size variable suggests that (if prices follow costs) increases in device size increase the cost of a bit of DRAM. That is, with a given technology, it is more costly (per bit) to produce large rather than small devices. The negative coefficients on Time and Time times Size, however, indicate that technological progress decreases bit price and that technological progress has been most important for large devices. The positive coefficients on Time Squared and Time Squared times Size indicate that the influence of technological change in reducing the bit price are diminishing with time and with increased device size over time. This is consistent with the idea of diminishing returns to size-related innovations and innovations generally.

Second, the negative coefficient on Output and the positive coefficient on Output Squared implies that the price per bit declines and then rises with increased output. This is consistent with the notion that there are economies of scale associated with IC production over *some* range. The rate of firm output at which prices begin to rise can be derived from the estimated coefficients and is 5.4 million devices per quarter for a 64 kilobit device. This rate is over four times the observed average output and close to the maximum of 8.53 million devices per quarter observed in the sample. It thus offers some comfort that the use of prices to study cost behavior may not be unreasonable.

Third, the negative coefficient on Cumulative Output indicates that, at least for the industry as a whole, cumulative output has an effect independent of that of current output. This finding supports the hypothesis that learning contributes to IC production behavior. Although it is not evident that

Table 1

Estimating the Effects of Market Structure, Learning, and Scale Economies on the Price of DRAMs

(Equation 1)

Dependent Variable: Natural log of DRAM price per bit (in $)

Variable	**Estimated Coefficient***	**T-ratio****
Constant	6.95	27.2
Size	.000156	9.66
Time	− .0741	6.47
Time Squared	.000931	4.45
Size x time	−7.12 D-06	9.16
Size x time squared	6.91 D-08	7.34
Output (devices per firm)	− .000154	2.42
Output Squared	2.87 D-08	3.44
Cumulative Output	−9.56 D-12	4.76
3-firm Concentration Ratio	1.28	5.29
Size x 3-firm Concentration Ratio	−2.79 D-05	4.90
Japanese Market Share	− .0282	10.57
Size x Japanese Market Share	4.94 D-07	7.64

Derived Estimates

(64K device)

Elasticity of price with respect to cumulative output:	−.08
with respect to 3-firm conc. ratio:	−.35
with respect to Japan market share:	+.13

Point at which increased output is associated with increase in DRAM prices: 5.4 million devices per quarter

R squared	.954
n	154

*Coefficients estimated using ordinary least squares. Where necessary the estimates are presented in scientific notion, with base-10 exponents.

**All of the estimated coefficients are statistically distinguishable from zero at the 95 percent confidence level or better.

Source: Dataquest, Inc., Federal Reserve Bank of San Francisco individual firms are able to exploit learning curve phenomena in devising pricing strategies, it does indicate that the theoretical potential to do so exists.

Fourth, the positive coefficient on the Concentration Ratio and the negative coefficient on Size times the Concentration Ratio suggests that the influence of market share concentration on DRAM prices is positive for small devices, but negative for larger devices. (The effect becomes negative for a "size" greater than about 40 kilobits.) As was observed earlier in this paper, a negative effect of market share concentration on price is suggestive of firm-specific strategic pricing in the presence of a learning curve. The negative coefficient for 64K and 256K devices is interesting because these are the device sizes that came to be dominated by the Japanese, suggesting, perhaps, that the Japanese introduced a different pricing strategy into the DRAM IC market.

The independent effect of Japanese market presence on DRAM prices is given by the negative coefficient on Japanese Market Share and the positive coefficient on Size times Japan Market Share. It appears that, controlling for other market and production influences, the effect of the Japanese presence was to reduce prices for smaller devices (empirically, smaller than 57 kilobits), but to elevate prices for larger devices. Once again, the markets for larger devices are ones that the Japanese are said to dominate. Since we have controlled for industrywide learning by Cumulative Output, and early strategic "underpricing" may be accommodated by the concentration ratio variables, this finding could be seen as evidence that prices can be elevated by successfully dominating the market.

There are a number of important qualifications to these findings. An obvious difficulty lies with the study of average prices and average firm output to make inferences about what inherently is relevant only to individual firm costs and output. Many relationships that hold at an individual firm level are not appropriately aggregated or averaged. A similar criticism attaches to the use of average cumulative output instead of individual firm data to detect the presence of learning curve phenomena. More complex functional forms, separate specification of the estimated relationship for each device, and recognition of qualitative differences across devices also would be useful. The available data, however, does not allow us to resolve all of these potential sources of bias.

A Study of Fabrication Facilities

Our second empirical investigation focused on the fundamental unit of fabrication: the fab line. This investigation is of interest both to verify the casual observation made in Section II about the likely lack of scale economies in fabrication and to study differences in fabrication activity on fab lines operated by Japanese and American firms. If fabrication costs were the same in both countries, then such costs would have to be eliminated as a source of differences in pricing strategies.

Once again, quite severe data limitations restrict the type and quality of analysis that can be performed. Fab line cost data are not available; only data on installed capital equipment (in 1986 U.S. dollars), the number of employees engaged on the line, and wafer start activity are available. Data on the specific type of IC device produced on the line also are not available[26]. Nevertheless, the data do permit two simple empirical tests — a comparison of capital-labor ratios and a comparison of fab line production functions.

First, using data on 386 fab lines from June 1986, we computed the capital-labor ratios separately for American and Japanese fab lines[27], and found the ratio in Japanese fab lines to be approximately 2 percent less than the ratio on American lines[28]. This finding is likely an understatement of the actual difference because the Japanese work week is one day longer than the 5-day U.S. standard. Thus, for Japanese lines, the number of employees on the fab line is a downward-biased measure of labor input flows (in man-days).

The finding of a 2 percent difference implies that, assuming the same fabrication technology and quality of labor, the *unit* cost of labor relative to capital is lower for Japanese than American firms. Since the capital equipment costs likely are very similar (since much of the equipment is American in origin), this, in turn, could be consistent with the existence of absolutely lower labor factor costs in Japan [29].

The second use of the data was to estimate a fab line production function directly for a combined sample of Japanese and American fab lines. The estimated functional form was the Cobb-Douglas representation of the production function:

$$Q = aL^bK^c \tag{2}$$

where

Q = fab line output (measured in square inches of wafer starts per week)
L = the number of fab line employees
K = the dollar value of fab line capital (in millions of U.S. dollars).

The Cobb-Douglas representation has a number of well-known limitations, the most important of which is that the rate of substitution between factors is constrained to be equal to one. It has the advantage, however, that the exponents of labor and capital provide estimates of the marginal products of these respective factors. It also can be shown that the relative sizes of these two coefficients are related to the relative contribution of each factor to total product under certain assumptions, and that the sum of these coefficients is a measure of economies of scale[30]. (Specifically, if b plus c in equation 2 is greater than one, the production function exhibits economies of scale; if they sum to less than one, there are diseconomies to large scale production.) In addition, with simple assumptions about factor prices and the profit-maximizing behavior of the firm, cost functions can be derived[31].

The estimates of the coefficients of equation 2 were obtained by taking the logarithm of both sides and using ordinary least squares regression techniques. The results are presented in Table 2, with a dummy variable introduced to identify possible coefficient differences between American and Japanese fab lines. The coefficients on fab line employment and fab line capital are, respectively, .54 and .35 for the American fab lines. It thus appears that there are no scale economies from fabrication *per se*. Indeed, the point estimate of the coefficients imply very slight *diseconomies* of scale (.89 versus 1.00 for constant returns to scale) although the estimate cannot be distinguished from 1.0 statis-

TABLE 2

Estimating the Fab Line Production Function

(Equation 2)

Dependent Variable: Log of Achieved Wafer Output (in thousands of square inches per week)

Variable	Coefficient	t-ratio
Constant	3.59	(7.08)*
Country dummy	.84	(0.67)
Fab line employment	.54	(5.95)*
Fab line employment times Country dummy	.25	(1.16)
Fab line capital	.35	(3.62)*
Fab line capital times Country dummy	−.04	(0.17)
R-squared	.46	
n	381	

Country dummy equals 1 if a Japanese fab line, 0 otherwise.
Fab line employment is in thousands of employees, in log terms.
Fab line capital is in millions of 1986 US dollars, in log terms.
Asterisk (*) indicates that the estimated coefficient differs from zero at the 90 percent confidence level or better.

Data souce: VLSI, Inc., San Jose, CA

tically. The importance of the labor component of fabrication inputs also is illustrated by the estimates, which show that labor's contribution to total product is approximately fifty percent greater than that of capital[32].

American vs Japanese Production Functions

There also appear to be no differences in fab line production functions between American and Japanese facilities. All three of the coefficients on the variables designed to capture these differences (that is, the Japanese dummy variable and its interaction with fabrication employment and fabrication capital) are not statistically different from zero. From this evidence alone, there is little to suggest that fabrication economies can explain differences in final product prices between the two countries. In other words, if manufacturers in each country were profit-maximizing and faced the same labor and capital costs, there is no statistical evidence that total fabrication cost relationships would differ[33].

Important qualifications on this finding must be offered, however. First, the output measure used in estimating equation 2 is not the number of usable ICs completed but simply square inches of wafer processed. To the extent that Japanese and American firms differ in their ability to recover usable ICs from this process, effective cost per IC would differ[34].

Second, Japanese and American producers may emphasize different products not accommodated by the simple production function estimated here. Some data are available on coarse product categories associated with each fab line. However, the use of separate dummy variables for these categories did not significantly influence the estimated parameters, perhaps because the sample sizes for some of those product variations were small.

Finally, if the differences in capital-labor ratios on fab lines observed earlier in the two countries is indicative of absolute differences in labor and capital costs in the two countries, production costs would differ accordingly.

V. Conclusions

Although the available data do not permit definitive assessment of the factors that may affect firm behavior within the IC industry, some light has been shed on two major aspects of the performance of this industry. The first is whether the industry is prone to high levels of market share concentration or other features that may result in inefficient performance.

Pricing behavior was consistent with sizeable, but not extreme, overall scale economies, which include pre-production costs. Market structure in commodity-type DRAM devices appears to be concentrated only at low industry output levels.

Pricing behavior in the DRAM market is, however, consistent with the existence of a firm-specific learning effect. In addition, large, "sunk" pre-production investments are required to enter new device markets. Both phenomena would tend to give strategic advantages to incumbent firms and hence to firms (of any nationality) that might be supported through periods of negative earnings while they acquire the advantages of production experience and incumbency.

The second major issue confronting the IC industry is the conduct of foreign competitors compared to U.S. manufacturers. The fact that estimated fab line production functions did not uncover significant intercountry differences casts some doubt on differences in fabrication costs as a source of competitive advantage for Japanese producers.

Weighing against this view is our finding (in the context of fabrication lines) that firms in Japan behave as if their labor is less costly (or of lower quality) relative to capital than firms elsewhere. Since the conventional wisdom is that Japanese labor is not of lower quality, the behavior of Japanese firms argues in favor of a cost-advantage to Japanese production of ICs. Perpetuation of this disparity runs counter to the notion of international factor price equilibration predicted by trade theory, but without additional information, the argument that cost differences are the basis for the growing presence of the Japanese in the IC market must stand.

The Future of the U.S. Semiconductor Industry

Several of our findings suggest that semiconductor markets lost to foreign competition may be difficult to recover. For one, assuming that production experience confers cost-advantages on a firm and that a growing scale of "sunk" costs is associated with *de novo* entry, it follows that new nonsubsidized firms will find it increasingly difficult to dislodge incumbent firms, whether that incumbency was achieved through cost-advantages or subsidized operation. Second, although the DRAM market, at least, is not especially concentrated at this time, markets in DRAMs have tended to become rigidly structured over time.

On a more positive note, two recent policy changes have important implications for American firms. First, the passage of the Semiconductor Chip Protection Act, by giving property rights to designers of semiconductor chip maskworks, should reduce significantly the more egregious piracy practices. The Act will reduce the likelihood that different firms will face different effective costs of maskwork capital. If, as is popularly alleged, foreign firms previously acquired maskwork capital through reverse engineering at the expense of American firms, the improvement in property rights in IC maskworks should benefit American IC producers.

The second important policy initiative is the 1986 agreement reached between the U.S. and Japanese governments regarding, among other things, international semiconductor pricing policy. The agreement was reached by negotiation between the U.S. and Japanese governments to resolve complaints about Japanese IC pricing policy brought before the U.S. International Trade Commission and in petitions filed under Section 301 of the Trade Act of 1974[35].

The ITC complaints and petitions were dropped in return for agreements from the Japanese to cease the alleged practices of (1) retarding U.S. entry into Japanese markets and (2) "dumping" of Japanese products below cost in U.S. markets. In particular, the agreement provides firms in both countries protection against "subsidized" sale of semiconductors (priced below "company specific cost of production plus 8 percent"). Also, as part of the agreement, the Japanese government is charged with monitoring the relationship between firm costs and selling prices abroad.

The agreement potentially could provide a forum for resolving the debate about whether the Japanese producers are, in fact, subsidizing IC production. To the extent that the concept "company specific cost of production" is meant to refer to short-run average costs, the agreement also could retard learning curve pricing strategies and thereby improve prospects for American firms. At this writing, however, the agreement had broken down because of the alleged failure of the Japanese government to enforce its terms.

FOOTNOTES

[1] See, for example, "The Solid State Era," Electronics, April 1980, M. S. Kiver, Transistor and Integrated Electronics, McGraw-Hill, 1972, and W. C. Hittinger, "Metal Oxide Semiconductor Technology," *Scientific American*, August 1973, pp. 48-57.

[2] World Semiconductor Trade Statistics Committee, Semiconductor Industry Association, "Semiconductor Forecast Summary," September 1986.

[3] Excellent summaries of the technical relationships among IC and other semiconductor devices are available in the annual reports of VLSI, Inc., a San Jose, California, research firm.

[4] See, for example, the F. Thomas Dunlap, Prepared Statement, "The Semiconductor Chip Protection Act," USGPO, J-98-39, pp. 152-168, and "Intel's Development of 386 Chip Took 4 Years and $100 Million," *Wall Street Journal*, August 29, 1986, p. 4.

[5] See footnote 3.

[6] Wafer processing into ICs involves such steps as application and baking of special coating materials, deposition or doping selected portions of the surface, etching and numerous measurement and testing steps. The speed of each of these steps is not easily accelerated. In addition, the wafer is carried from one step to the next mechanically and production takes place, for many of the critical steps, in hoods or rooms with highly processed atmospheres. It is difficult to add equipment, work stations, or speed up processing at will in such a confined and sequence-driven environment.

[7] These data refer to MOS technology fab lines. The source of the data is VLSI, Inc.

[8] Specifically, the original contact photolithographic techniques for transferring circuit designs to the surface of the IC have been improved first through optical projection techniques and today, electron beam lithography techniques.

[9] The prepared statement of the Semiconductor Industry Association, ibid, pp. 122-128, discusses the reverse engineering process.

[10] From the testimony of F. Thomas Dunlap. See footnote 4.

[11] Karen Ammer, "The Semiconductor Chip Protection Act of 1984," *Law and Policy in International Business*, Vol 17, 1985, pp. 395-420. See also, J. Chesser, "Semiconductor Chip Protection: Changing Roles for Copyright and Competition," *Virginia Law Review*, Vol 71, 1985, pp. 249-285.

[12] This notion was studied in an early Federal Trade Commission report and studies performed by the Boston Consulting Group. See "Staff Report on the Semiconductor Industry," Bureau of Economics, Federal Trade Commission, January 1977 and "Perspectives on Experience," Boston Consulting Group, 1972.

[13] See, for example, the complaint before the International Trade Commission regarding 256K DRAMs (ITC Docket number 731-TA-300).

[14] See E. Bailey and W. Baumol, "Deregulation and the Theory of Contestable Markets," *Yale Journal on Regulation*, Vol 1, 1984, pp. 111-137.

[15] A. Michael Spence, "The Learning Curve and Competition," *The Bell Journal of Economics*, Spring 1981, pp. 49-70. For an application of the learning curve evaluation process to another industry, see, for example, M. B. Lieberman, "The Learning Curve and Pricing in the Chemical Industries," *The Rand Journal of Economics*, Summer 1984, 213-239.

[15a] Statistics on world semiconductor sales shares are not known with precision. The statistics on US and Japanese *semiconductor* market shares are from "International Competitiveness in Electronics," US Office of Technology Assessment, Washington, DC, 1983. The data on world IC shares are from "For Chipmakers, the Game has a New Set of Rules," *Business Week*, January 13, 1986, p. 90. The trade balance in ICs also reflects Japanese competitiveness. In 1979, the US/Japan trade balance in ICs was a surplus of $90 million. By 1984, this had deteriorated to a deficit of $884 million. (Source: *Industry Week*, November 25, 1985.)

[16] These involved 64K DRAMS (ITC Docket #731-TA-270), 256K DRAMs (ITC Docket #731-TA-300) and EPROMs (ITC Docket #731-TA-288). DRAMs are Dynamic Random Access Memory ICs and EPROMs are Erasable Programmable Read Only Memory ICs.

[17] See, for example, "The Effect of Government Targeting on World Semiconductor Competition: A Case History of Japanese Industrial Strategy and Its Costs for America," Semiconductor Industry Association, 1983.

[18] This debate is presented in Okimoto, Sugano and Weinstein, eds, *The Competitive Edge: The Semiconductor Industry in the US and Japan*, Stanford University Press, 1984.

[19] In one seemingly egregious instance, an American chip design was reproduced by a Japanese producer so precisely that even a microscopic error in the chip architecture was replicated. See the document cited in footnote 4 above.

[20] See J. Chesser, ibid.

[21] P. Dasgupta and J. Sitglitz, "Uncertainty, Industrial Structure, and the Speed of R&D," *The Bell Journal of Economics*, Spring 1984. See also J. E. Tilton, "International Diffusion of Technology: The Case of Semiconductors," Studies in the Regulation of Economic Activity, The Brookings Institution (Washington, DC), pp. 24-38. Gort and Konakayama study the diffusion of the production of an innovation using a simple model and several industries, including the transistor industry. See Gort and Konakayama, "A Model of Diffusion in the Production of an Innovation," *American Economic Review*, December 1982, pp. 1111-1119.

[22] Miltiades Chacholiades, *International Trade Theory and Policy*, McGraw-Hill, 1978, Chapter 10.

[23] M. Chacholiades, Chapter 11.

[24] Specifically in the case of 64K and 256K DRAMs. See footnote 16 above.

[25] The Spearman Rank Correlation Coefficient measures the degree of correspondence between two series of numbers by examining rank differences. This statistic is useful for studies of market rigidity because in a fluid

market, market shares held by individual firms would change over time as new entrants disturbed market shares and place-switching occurred among extant firms. The fact that the Spearman Correlations observed in the DRAM market are high is illustrated by a few simple examples. Assume that an industry is structured so that the largest firm has 30 percent market share, followed by firms with 25, 20, 15, 10, 5, and 2.5 percent respectively. If the first and fourth firms switch places, the correlation between the old and new structure would be only 67 percent. If, instead, the smallest firm exits the market (with the next smallest firm taking up its market share), the correlation would be 92 percent. Both of these correlations are less than the year-to-year correlation observed in a mature DRAM market.

26. The source of this data was VLSI, Inc., San Jose, California. The data are *not* available from this author, as per agreement with VLSI, Inc.

27. The sample was confined to metal oxide semiconductor technology lines for comparability.

28. The difference is of only marginal statistical significance. Specifically, the difference is different from zero at approximately the 80 percent confidence level.

29. The capital-labor ratio also could be affected by differences in the price of capital. Indeed, it is argued frequently that Japanese producers have access to subsidized financial capital. If true, the reduction in the capital-labor ratio stimulated by lower labor costs could be offset by lower user costs of fab line capital.

30. See, for example, H. R. Varian, *Microeconomic Analysis,* Norton Books, 1976, Chapter 4 for a discussion of the econometric problems encountered in direct estimation of production relationships. See also G. S. Maddala, *Econometrics,* McGraw-Hill, 1977, Chapter 13. Suffice it here to say that the endogeneity of the right hand side variables is ignored and that estimation of production frontiers generally raises problems associated with nonnormally distributed errors.

31. See H. Varian, Chapter 1.

32. That is, the marginal product of labor is greater than the marginal product of capital by about 50 percent. If constant returns to scale were exhibited, it can be shown that under competitive market conditions, the coefficients on labor and capital could be interpreted as their respective factor shares.

33. The derivation of cost functions from assumptions of profit maximization and Cobb-Douglas production technology are presented at length in H. Varian, *ibid,* Chapter 1.

34. Because of the extremely small physical size of the constituent elements of an IC, small impurities or imperfections in the processed surface of the wafer can cause an IC to perform inadequately, thereby reducing the effective yield of ICs per wafer start. There may be international differences in the ability of producers to improve effective yields. In addition to eliminating or reducing the factors that caused the imperfections in the first place, it is also possible to build in circuit redundancy and other means of "salvaging" processed, but imperfect, chips. Even under the best of circumstances, however, the ratio of the actual to potential number of ICs per wafer may be as low as 50 percent for very large scale integrated devices such as the 256K DRAM (*Business Week,* August 18, 1986, p. 66).

35. See, "Chip Fight is Settled," Los Angeles Times, August 1, 1986, p. 11, and Semiconductor Industry Association, "US, Japanese Governments Reach Agreement on Market Access, Prevention of Dumping," *SIA Circuit,* Autumn, 1986, p. 1.

Tax Revolt or Tax Reform? The Effects of Local Government Limitation Measures in California

Carolyn Sherwood-Call*

During the past ten years, voters in many states have passed measures that limit the taxing or spending powers of local governments and thus their average level of services and ability to differentiate themselves from one another. This study of the effects of Proposition 13, a property tax limitation initiative passed in California in 1978, concludes that the initiative has reduced the overall size of local government, but that its effects on fiscal differentiation vary considerably depending on the extent of local governments' additional constraints.

During the past ten years, voters in many states have passed measures that limit the authority of governments in important ways. Most commonly, these initiatives have taken the form of limits on property tax rates or assessment practices, but spending constraints also have been passed.

These statewide initiatives potentially could interfere with the federal system of government. Under the federal system, different levels of government are responsible for providing different types of public or quasi-public goods and services. One can think of all goods as lying on a continuum between purely public goods and purely private goods. On this continuum, the national government should provide the purely public goods because the benefits of such goods are dispersed and the costs of producing purely public goods do not increase as the number of beneficiaries increases.

The degree of publicness of the goods provided should decrease as the level of government progresses from national to state to local. Thus, most services provided by local governments are nearer to the private good end of the continuum. Fire protection, for example, can be provided privately because it can be priced, its benefits can be limited to those who pay for it, and the cost of service provision increases as more households receive fire protection. At the same time, it is public in the sense that it benefits many who do not consume it directly. That is, if one house on a block has adequate fire protection, nearby houses are protected from the spread of a fire that starts in that house.

Quasi-public goods such as fire protection are particularly well-suited to local government provision because a rudimentary "market" allows individuals, by choosing a jurisdiction of residence, to select a combination of taxes and public services that suits their tastes and needs (Tiebout 1956). Thus, the abilities of local governments to provide different levels and mixes of services allows these quasi-public goods to be provided more efficiently, and the federal system to function more smoothly.

Statewide initiatives that limit local governments could interfere with this federal system of providing goods and services. Measures that tightly circumscribe local governments may limit the extent to which jurisdictions can differentiate themselves from one another. Residents of some communities may not be able to consume the level of government services that they would have chosen and for which

* Economist, Federal Reserve Bank of San Francisco. Martha Sellers provided invaluable research assistance. The paper also benefited from conversations with Larry McCarthy and Rebecca Taylor, both of the California Taxpayers' Association, and from comments by the editorial committee.

they would have paid in the absence of such restrictions.

In this paper, I examine the effects of Proposition 13, a property tax limitation initiative passed by California voters in 1978, both on the average level of services provided by local governments in California and on the extent to which different localities could continue to provide a variety of service levels. The effects differ substantially among the various types of local jurisdictions depending on their reliance on the property tax, alternative financing sources, and the extent of other constraints.

To establish the context for the passage of Proposition 13, Section I describes California's local government institutions and trends during the early and middle 1970s. Then, Section II describes and interprets Proposition 13 and other initiatives that comprised California's "tax revolt." In Section III, California data are used to examine the hypothesis that the statewide limits on local government reduced the extent to which local governments could carry out their functions within the federal system. Section IV summarizes and draws conclusions.

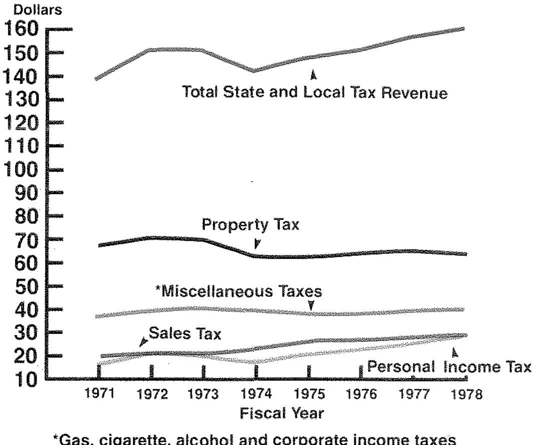
I. California Local Government Before 1978

During the early and middle 1970s, California was a high-tax state by almost any measure. For example, in Fiscal Year 1970-71 (hereafter "FY 1971"), California's total state and local tax burden of $137 for each $1,000 in personal income was high enough to rank eighth nationally. Of that total, 49 percent was generated by property taxes, 15 percent by sales taxes, and 10 percent by the state's personal income tax[1]. The income and sales tax burdens were not excessive relative to those in other states, as California ranked 18th in income taxes per $1,000 of personal income and 28th in sales taxes. Instead, the burden of property taxes appears to be largely responsible for the overall high tax burden borne by state residents. California ranked second in the nation in property taxes paid per $1,000 of personal income. Thus, Californians easily could have perceived their property tax burdens as excessive even before the years immediately preceding their tax revolt.

Chart 1 illustrates changes that occurred in California's combined state and local tax sources prior to the implementation of Propositon 13. By FY 1978, total tax collections for state and local governments had risen somewhat, to $158 per $1,000 personal income, from $137 in FY 1971. Meanwhile, property tax collections per $1,000 personal income actually fell slightly, from $67 in FY 1971 to $64 in FY 1978. As a result, the property tax became relatively less important. Its share in total tax collections fell from 49 percent in FY 1971 to 40 percent in FY 1978. At the same time, income and sales taxes became relatively more important revenue sources, providing 17 and 18 percent, respectively, of total state and local tax revenues in FY 1978.

The most striking observation regarding this period is that, despite the rapid increase in property values during the period and the widespread popular impression of mushrooming property tax burdens, property taxes as a proportion of personal income were in fact shrinking. Property tax rates themselves fell slightly, but much more surprising is the *drop* in assessed value per $1,000 of personal income from $2,487 in 1971 to $2,380 in 1978. Market values of property were increasing rapidly during this period but growth in *assessed* values did not exceed the rapid growth in personal income which also characterized the period. One explanation for the slower

Chart 1
State and Local Tax Revenues Per $1000 Personal Income
California excluding the City and County of San Francisco



*Gas, cigarette, alcohol and corporate income taxes accounted for 60 percent of miscellaneous revenues in 1977.

rise in assessed property values is that, at least in some areas, assessed valuations were kept below market value precisely to avoid the political consequences that would accompany sudden and sharp increases in property tax liabilities.

Calculations by Oakland (1981) suggest that assessed values of single-family homes were growing much more quickly during the immediate pre-Proposition 13 period than were property values more generally[2]. Therefore, individual homeowners are likely to have experienced substantial increases in their tax bills during this period.

Property Tax Structure

Prior to Proposition 13, property tax rates were not coordinated among the various fiscally independent local jurisdictions such as cities, counties, and school districts. As a result, voters who wished to reduce their property tax burdens had no clear local target to confront.

The governing body of each jurisdiction established its property tax rate annually to conform to its budget requirements, and forwarded the rate to the county assessor's office. The county then calculated tax bills for individual parcels by adding up the rates for all jurisdictions that serviced the area and multiplying by the assessor's office estimate of property value. Thus, an individual property's tax bill would likely include payments to a multitude of different jurisdictions, including the county, city, and school district as well as any number of special districts for services such as water, lighting, fire protection, streets, parks, flood control, cemeteries, or pest control.

In practice, a little over half of all California property tax revenues funded school districts. In FY 1972, school districts received 52 cents of each property tax dollar, while counties received 32 cents, cities 10 cents, and special districts 6 cents.

School Districts

Conversely, the property tax was a particularly important source of revenue for school districts. In FY 1972, for example, 54 percent of school district funds in California were derived from the property tax. Of the remainder, a majority (31 percent) came from state aid.

Starting in FY 1974, the school funding process changed considerably as a result of a 1971 ruling of the California Supreme Court (*Serrano v. Priest,* 5 Cal. 3rd 584). The Court ruled that it was unconstitutional for some school districts to provide inferior education because low wealth in their areas limited their property tax revenues. To implement the *Serrano* decision, the state legislature placed a cap on the amount of property tax revenues per pupil that school districts could raise. Each school district's allotment was increased annually, but to reduce the gap between high-wealth and low-wealth districts, those districts that had raised less money per pupil during FY 1973 received larger increases. By capping the principal revenue source available to school districts, *Serrano* effectively limited per-pupil spending as well.

At the same time, *Serrano* imposed a minimum per-pupil *spending* level on all school districts that raised the share of state aid in school funding. For those districts with low property tax revenues, state funds filled the gap between the 1973 property tax revenue base and the minimum per-pupil spending level. As a result, the share of property tax revenues in total school district funds fell to 46 percent in FY 1974 while the share of state aid rose to 39 percent. During subsequent years, the property tax share recovered somewhat, reaching 50 percent by FY 1978.

Counties

Counties also rely primarily on property taxes and intergovernmental grants for revenue. In California, counties provide a rather limited range of services, the most important of which are public assistance programs, judicial services, and health services. Most of these functions are mandated by the state and, as a result, the state traditionally has viewed counties more as administrative arms of state government than as autonomous local governments. A "rule of thumb" used in state government is that 75 to 85 percent of county expenditures are mandated by the state, leaving only 15 to 25 percent of county spending under local control. As a result, the level of differentiation among counties is limited[3].

Consistent with these characteristics, counties have little independent revenue-raising authority. They can impose property taxes and a few other taxes such as sales[4], real property transfer, and timber yield taxes, and they can charge fees for services they provide[5]. In 1978, before Proposition 13 was implemented, intergovernmental grants provided 50 percent of county revenues, while property taxes provided 33 percent, and user fees 9 percent. Nonproperty taxes and miscellaneous revenue sources accounted for the remainder.

Cities

Cities have considerably greater leeway both in terms of the services they provide and in terms of the financing instruments available to them. With few mandates from higher levels of goverment, cities are relatively free to spend their money as they see fit. Typical city services include fire and police protection, streets, parks, libraries, and museums.

Moreover, cities can impose a relatively wide range of taxes, including hotel, utility, and payroll taxes. In addition, the state sales tax law allows cities to receive sales tax revenue equal to 1.00 percent of sales[6]. In 1978, property taxes provided 14 percent of total city revenues, while nonproperty taxes provided 19 percent, current service charges 31 percent, state and federal grants 20 percent, and miscellaneous other sources the remainder.

In summary, in the years immediately before California voters passed Proposition 13, total state and local tax revenues were growing modestly relative to personal income. At the same time, income and sales taxes were becoming more important both absolutely and relative to property taxes. More surprisingly, property tax revenues actually were falling as a proportion of personal income.

II. Changes in Local Finance

In 1978, California voters passed Proposition 13, which placed a one percent ceiling on property tax rates and stipulated that a higher rate could not be imposed without a two-thirds majority of voters.[7] Since property tax rates in California had averaged 2.67 percent of assessed valuation in 1978 and were in some cases over 3 percent, Proposition 13 immediately cut property tax rates substantially. In addition, the initiative rolled back all assessed property values to their 1975-76 levels. Annual increases in assessed values could not exceed 2 percent or the inflation rate, whichever was lower. When a property was sold, however, it automatically would be reassessed at its market value[8].

In 1979, only a year after Proposition 13 was passed, California voters approved another government limitation initiative. The Gann initiative (Proposition 4) prohibited any state or local jurisdiction from spending more than it had spent during FY 1979, adjusted for increases in prices and local population. The Gann limit, like Proposition 13, included a voter override provision. By a simple majority, voters could approve spending in excess of the limit for a four-year period.

The Gann spending limit was more or less forgotten for several years because revenues were growing slowly enough that the limit was not binding for most jurisdictions[9]. More recently, however, it has become binding for many localities[10], and the number of jurisdictions constrained by the Gann spending limit is likely to mushroom over the next several years if revenues grow as rapidly as expected.

There are several possible reasons for the passage of these initiatives. Some argue that voters objected to increases in their property tax bills that were not the result of an explicit policy decision to increase property tax revenues, but rather resulted simply because property values rose. In fact, Oakland's calculations (1981) suggest that many owners of single-family homes did experience exorbitant increases in the assessed value of their homes, although the data presented in Section I suggest that such increases did not occur among property assessments more generally. Moreover, those for whom assessments did not rise dramatically still faced large potential reassessments and, in many areas, actual reassessments were not carried out uniformly or equitably.

Others argue that the tax revolt was a response to voters' perceptions that, despite a widespread demand for smaller governments, government was growing too large. This interpretation implies that voters thought that governments had become unresponsive to them, and that government was growing "too fast." The data suggest that state and local governments in California were imposing an increasing tax burden on state residents as tax revenues per $1,000 personal income grew from $137 in 1971 to $158 in 1978. Although the increase was not related to the property tax[11], the property tax provided a convenient target because it was the largest single source of state and local revenues, and its administration was so disjointed. Since the Gann initiative was unrelated to property taxation and applied to the state government as well as to local jurisdictions, its passage reinforces the idea that taxpayers had grown dissatisfied with the size and unresponsiveness of local governments.

Thus, two related explanations for the tax revolt are consistent with both the public finance climate during the middle 1970s and the passage of Propositions 13 and 4. Homeowners could have been rebelling against an increase in their property tax payments unrelated to an increase in their desire to pay for and receive local services. Alternatively, voters could have sought to reduce the absolute size of local (and state) governments by wresting control of local governments from politicians and bureaucrats. Under either interpretation, Proposition 13 is likely to have resulted in smaller local governments and in reduced property tax burdens.

If local governments have, in fact, become smaller, we would expect to see lower revenues and also less variation in the level of spending among jurisdictions. The imposition of a new upper revenue limit would restrain jurisdictions from spending larger than normal amounts of money and thereby reduce the degree of variation among them. Such a limit would compromise local governments' roles in the federal system by allowing voter-residents fewer choices of tax-spending combinations than they otherwise would have enjoyed.

III. The Impact of the Fiscal Changes

Changes in the fiscal environment for state and local governments were dramatic during the late 1970s and early 1980s, as Chart 2 illustrates. In FY 1978, just before Proposition 13 took effect, California's property tax revenue per $1,000 of personal income was $64, and property taxes provided 40 percent of all state and local tax revenue in California. In the following year, the property tax burden had fallen by more than half, to $30, providing only 25 percent of California's state and local tax revenues.

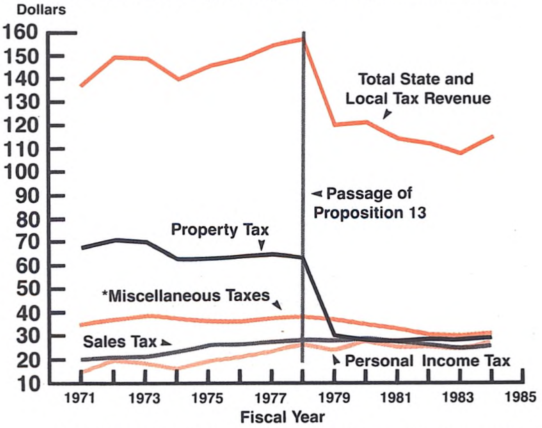
Sales and income tax burdens on California taxpayers remained relatively stable during the years after Proposition 13 took effect. Sales tax revenue per $1,000 personal income fell from $29 in FY 1978 to $26 in FY 1984, while individual income tax revenue per $1,000 personal income rose only from $27 to $28 during the same period. These two taxes taken together comprised 35 percent of total California state and local tax revenue in FY 1978, and that proportion rose to 43 percent in FY 1979. It rose slowly thereafter, reaching 47 percent in FY 1984.

As a group, these changes brought California's total state and local tax burden per $1,000 in personal income down from $158 in 1978 to $121 in 1979. By 1984, the burden had fallen still further, to $115. California generally ranked among the top five states in terms of total state and local taxes per $1,000 in personal income during the pre-Proposition 13 years, and was consistently in the top ten. The state's ranking fell to 25 immediately after Proposition 13 was imposed and averaged 20 during

Chart 2

State and Local Tax Revenues Per $1000 Personal Income

California excluding the City and County of San Francisco



*Gas, cigarette, alcohol and corporate income taxes accounted for 60 percent of miscellaneous revenues in 1977.

the five subsequent years. In FY 1984, California's total state and local tax revenue per $1,000 personal income was 99 percent of the national average, a substantial decrease from 121 percent in FY 1977.

School Districts

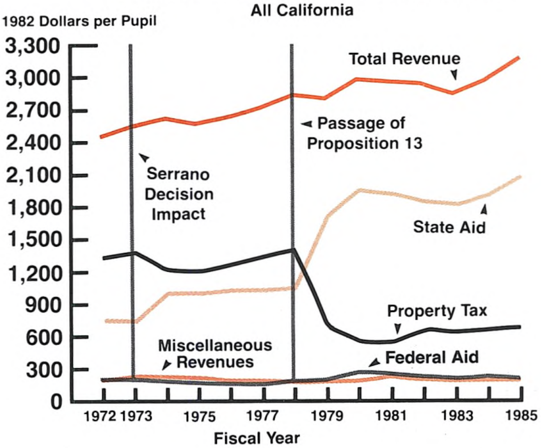
The discussion of the differences among jurisdiction types presented in Section I suggests that school districts should have borne the brunt of the impact from Proposition 13, since they previously had relied most heavily on property taxation and they had few alternative funding sources. However, the impact of Proposition 13 on school districts was greatly complicated by the *Serrano* decision.

Under the *Serrano* mechanism first implemented in 1974, state aid filled the gap between locally generated property tax revenues and minimum funding levels for those school districts that needed additional revenues. When Proposition 13 drastically reduced the amount of property tax money available, the contributions from state aid increased almost commensurately. Indeed, as Chart 3 shows, the time paths of state aid and property tax revenues (in real per-pupil terms) were almost perfect mirror images of each other. In FY 1985, the state provided 63 percent of funding for schools, and the property tax accounted for only 22 percent of school funds.

Total state funding for schools per $1,000 of personal income has fallen since Proposition 13 was passed, but school attendance has been falling rapidly enough that inflation-adjusted per-pupil revenues actually have increased. In 1982 dollars, per-pupil revenues fell from $2,841 in 1978 to $2,805 in 1979. Since then, however, they have increased every year except one[12], so that 1985 revenue per pupil was $3,176 (in 1982 dollars)[13].

Moreover, the evidence suggests that the degree of differentiation among school districts has declined during the period in which *Serrano* has been effective. For example, when unified school districts are ranked in order of their revenue limits, the dollar difference among districts in the middle 50 percent was $256 in FY 1974. By FY 1983 the difference had fallen to $129, despite a 92 percent increase in the price level during the period[14]. Because *Serrano* was aimed explicitly at limiting spending differences among school districts, it likely bears more direct responsibility for this change than does Proposition 13.

Chart 3
Sources of School District Revenues
All California



At the same time, Chart 3 suggests that, by changing the major funding source from property taxes to state aid[15], Proposition 13 may have reinforced the tendency toward greater uniformity among school districts. As a result, the Tiebout mechanism of "voting with the feet" probably is less important in education provision than it was before the *Serrano* and Proposition 13 changes took place.

Counties

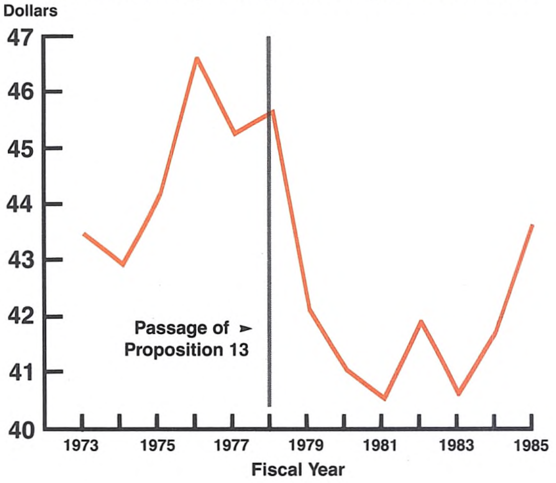
Because counties also derive a substantial proportion of their revenues from the property tax, one would expect Proposition 13 to have made a substantial impact on county finances. Indeed, county revenues per $1,000 of personal income fell by 16.5 percent between FYs 1978 and 1980[16]. However, as discussed in Section I, counties traditionally have had relatively little authority to differentiate themselves from one another because an average of 75 to 85 percent of their spending is mandated by the state.

As a result, the impact of Proposition 13 on counties was different from that on school districts. Overall, county funding fell whereas the same did not occur for schools. At the same time, counties' limited discretion before Proposition 13 suggests

Chart 4

Total City Revenues Per $1000 Personal Income

California excluding the City and County of San Francisco



that any impact on the degree of differentiation among counties would have been less dramatic than the combined effects of *Serrano* and Proposition 13 on school district differentiation.

Cities

The discussion in Section I suggests that Proposition 13 would have affected cities less than school districts or counties because cities relied less on the property tax for revenue. Nevertheless, cities in a sense had the most to lose from Proposition 13 because they had been most able to differentiate themselves from one another. Compared to other types of jurisdictions, cities had had many revenue sources and were not subject to as many state-mandated programs.

In addition to the property tax revenues they lost due to Proposition 13, cities lost a significant amount of federal grant money as the federal government rolled back its grants-in-aid programs. Between FYs 1978 and 1985, inflation-adjusted federal grants to cities[17] fell by 50 percent[18]. Whereas federal grants had provided 9.5 percent of total city revenues in 1978, they provided only 5.9 percent in 1984.

Chart 4 illustrates the combined effects of Proposition 13 and reduced federal funds on city finances. Total revenues fell substantially relative to estimated personal income[19] after Proposition 13 was implemented in 1979. They reached their nadir in FY 1981, but climbed thereafter; by 1985, total revenues stood significantly above their 1979 level.

Chart 5 breaks the change in total city revenues into its components. Service charges are extremely important sources of revenue. Because of the major infrastructure requirements of services such as water, utilities, and sewer, many cities must charge high fees to recoup the costs of providing these services. In addition, like counties, many cities generate "profits" from city services.

The total revenue picture for California cities changes, but not dramatically, when fee revenues are excluded from the total. Chart 6 gives an indication of the direction of changes, although a redefinition of data between 1981 and 1982 makes an accurate appraisal difficult[20]. The revenue data minus fees suggest, as do the total revenue data, that revenue growth resumed after a sharp decline in the late 1970s.

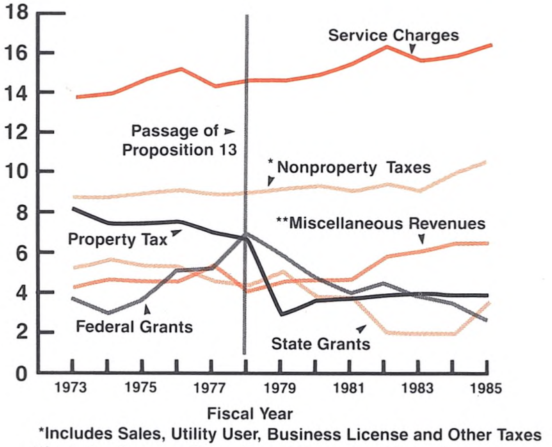
Table 1 sheds some light on how cities compensated for the declines in revenue from both property taxes and federal grants. It lists total revenues per $1,000 of personal income in 1978 and 1985 by source.

Nonproperty tax instruments provide the most obvious alternative revenue source. In addition to

Chart 5

City Revenues Per $1000 Personal Income

California excluding the City and County of San Francisco



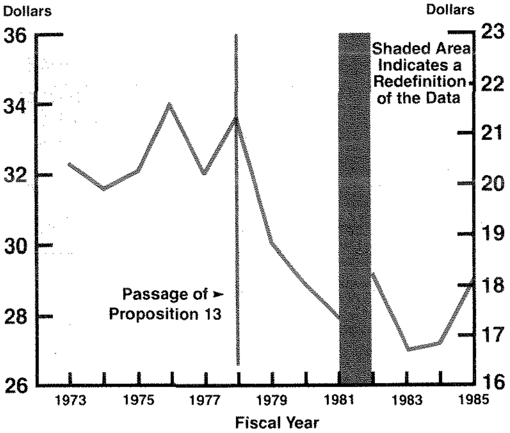
*Includes Sales, Utility User, Business License and Other Taxes

**Includes Permits, Fines, Use of Money and Property and Other Revenue

Chart 6

General City Revenues Per $1000 Personal Income

California excluding the City and County of San Francisco



sales taxes, many California cities impose taxes on utility use, real property transfers, transient lodging (hotels), and business licenses. These sources were tapped liberally with the result that revenues from nonproperty taxes rose 19.1 percent between 1978 and 1985[21]. While larger cities are in a better position to take advantage of such alternatives, nonproperty taxes can be important revenue sources for smaller cities as well. In FY 1985, nonproperty taxes accounted for 27 percent of total revenue for those cities with fewer than 25,000 residents, far more than the 10 percent that property taxes provided during that year.

Sales taxes are a particularly visible source of revenue for California taxpayers, but they were not responsible for the increase in nonproperty tax revenues. Sales tax revenues accruing to cities[22], as a proportion of total city personal income, actually fell 5 percent between 1978 and 1985, and their share of total revenues remained constant during that period.

Other revenue sources also grew substantially relative to personal income during this period. Although miscellaneous revenue sources provided less than 10 percent of total revenues in 1978, miscellaneous revenue grew by 62 percent between 1978 and 1985, and by 1985 actually provided 63 percent more money than did the property tax. Miscellaneous revenue sources include such items as investment and rental income, which cities began to tap much more aggressively when their more traditional revenue sources became scarcer.

User charges have provided another important source of alternative revenue for California cities. Service charge revenue as a proportion of personal income rose 13 percent between 1978 and 1985.

Table 1

Revenues per $1000 of Personal Income
(California Cities, excluding City and County of San Francisco)

	1978		1985	
	Amount	Percent of Total	Amount	Percent of Total
Total Revenue*	$45.55	100.0	$43.56	100.0
Property Taxes	6.73	14.8	3.94	9.0
Nonproperty Taxes	8.94	19.6	10.65	24.4
Sales Tax	5.59	12.3	5.31	12.2
Other Nonproperty Taxes	3.35	7.4	5.34	12.3
Service Charges	14.46	31.7	16.41	37.7
Intergovernmental Grants	11.45	25.1	6.12	14.0
Miscellaneous	3.97	8.7	6.43	14.8

*Components do not necessarily sum to totals due to rounding.

Some of the increase may have been associated with increased spending on the services for which cities levy charges, which is why this type of funding should not in principle be included as general revenues. When service charge funds are subtracted from total revenues, adjusted revenues in 1985 were 13 percent lower than their 1978 level.

Effect on Differentiation

One can get an idea as to whether the fiscal constraints on cities affected the extent to which they could differentiate themselves from each other by examining variances of revenue across cities. Calculating a consistent data set for all California cities is a prohibitively time-consuming task, so variances were calculated for two years using data from two counties only. These two counties, Alameda and Contra Costa, comprise the Oakland Metropolitan Statistical Area (MSA), and in 1985 included 31 cities[23] that varied substantially in terms of their sizes and the incomes and other characteristics of their residents.

Table 2 presents nonservice charge revenues per capita (in 1984 dollars) for these cities[24], with means and standard deviations. The table shows that total revenues on average fell about 10 percent between 1978 and 1985, although there are significant differences among the cities. Real per capita revenues rose for several cities, including Alameda, Oakland, Walnut Creek and Pleasant Hill. However, they fell substantially for several others, including Berkeley, Newark, Union City, Antioch, and Pittsburg.

While changes in per capita revenues went in both directions, the standard deviation fell slightly, from $240 to $235. The ratio of standard deviation to the mean rose slightly, from 0.617 to 0.671. These small changes suggest that cities' abilities to differentiate themselves from each other did not change significantly, and that Proposition 13 did not interfere with the variety of tax-spending combinations available to consumers when they "vote with their feet."

TABLE 2

Nonservice Charge Revenues, Oakland MSA Cities

(1984 dollars per capita)

	Fiscal Year 1978	Fiscal Year 1985
Alameda County		
Alameda	$ 312	$ 346
Albany	373	334
Berkeley	620	482
Dublin	—	422
Emeryville	1358	1342
Fremont	317	292
Hayward	462	395
Livermore	294	255
Newark	332	220
Oakland	682	793
Piedmont	419	426
Pleasanton	270	388
San Leandro	434	463
Union City	418	247
Contra Costa County		
Antioch	311	189
Brentwood	265	383
Clayton	189	165
Concord	323	326
Danville	—	79
El Cerrito	298	259
Lafayette	141	140
Martinez	319	256
Moraga	158	155
Pinole	196	315
Pittsburg	412	229
Pleasant Hill	188	272
Richmond	710	531
San Pablo	366	282
San Ramon	—	144
Walnut Creek	347	376
mean	$ 389	$ 350
standard deviation	$ 240	$ 235
ratio s.d./mean	0.617	0.671

IV. Summary and Conclusions

This paper has attempted to determine whether the initiatives passed during California's "tax revolt" had a significant impact on the level of spending by local jurisdictions or on the variation in spending levels among jurisdictions. The analysis presented suggests that, on the whole, Proposition 13 appears to have reduced the size of local governments since the tax burden, as measured by taxes as a proportion of personal income, has fallen since Proposition 13 was passed.

Other effects of Proposition 13 have differed considerably among the different types of jurisdictions, but for most localities, the changes brought about by Proposition 13's limitations on property tax revenue have been tempered by increased reliance on other revenue instruments. For example, school districts and counties now rely more heavily on state aid. For school districts, the increase in state aid was partly the result of the *Serrano* decision that called for increased homogeneity among school districts, and corresponded with an increase in state authority over education at the expense of local authority. For counties, increased state aid simply reflected the previously ignored reality that California counties exist largely to carry out state mandates. The extent of heterogeneity among counties was limited at the low end by these mandates before Proposition 13 was implemented, and now is limited at the high end as well, by revenue constraints.

Cities, in contrast, remain relatively autonomous local governments, and they continue to derive their revenues primarily from local sources. Proposition 13 did affect cities substantially, but its impact was mitigated by cities' initial limited reliance on property taxes and by their ability to increase revenues from alternative sources such as service charges and nonproperty taxes.

FOOTNOTES

[1] Most of the remaining tax revenues came from "selective" sales taxes on such items as gasoline, alcohol, and tobacco, which accounted for about 10 percent, and corporate income taxes, which accounted for about 5 percent. Motor vehicle licenses and other miscellaneous taxes accounted for the remainder.

[2] Oakland estimated that the share of single-family assessments in the total rose from 31.6 percent to 41.0 percent between FYs 1974 and 1978. Using his estimates together with state total assessed valuations reveals that single-family home values rose at an average annual rate of 20.4 percent over the five-year period, while assessed values of nonresidential property increased at an average annual rate of 12.7 percent during the same period.

[3] While counties could charge higher taxes and spend more money than state mandates called for, the existence of an effective lower limit on spending meant that the overall variation among counties would have been smaller than for jurisdiction types with no spending floor.

[4] All California counties receive revenues equal to 0.25 percent of taxable sales, but the funds must be spent for transportation. In addition, residents of some counties have voted additional sales taxes for transportation purposes, such as the 0.5 percent tax in the three counties served by the Bay Area Rapid Transit District.

[5] California state law prohibits local governments from charging "excessive" fees for services rendered. In principle, this means that they can recoup the costs of service provision but they cannot generate "profits" for use in other areas of government. In practice, however, it is widely acknowledged that many localities generate general revenues by selling such items as water and electricity at prices above true costs.

[6] Counties receive all or part of the 1.00 percent that cities do not opt for, and in addition receive the full 1.00 percent in unincorporated areas.

[7] Later the law was clarified to allow rates to exceed one percent if the additional proceeds were used to retire existing debt.

[8] Thus, by the 1980s, Proposition 13 had created a large gap in property tax bills between long-time owners and new owners.

[9] Some argue that in addition local jurisdictions inflated their spending during FY 1979 to make the Gann limit less restrictive.

[10] According to a recent study by the California Tax Foundation, 119 jurisdictions have sought to override the Gann spending limit at least once. Of the 60 elections held through 1986 in which increased spending authority did not require higher tax rates, only two resulted in a return of funds to taxpayers rather than an increase in spending authority. In contrast, tax-linked elections had passed 46 times and failed 33 times by early 1987.

[11] Of the $21 increase, $12.50 was due to an increased individual income tax burden and $8.54 was due to an increased sales tax burden.

[12] Per-pupil spending could increase from its 1979 level despite the Gann limit because the school districts' own limits include only locally generated funds. Since most of the increase in school funding has come from the state level, the funds are subject to the state's Gann limit, which has not yet become binding, rather than to the school districts'. Property tax revenues per pupil actually fell from $704 to $694 (in 1982 dollars) between FYs 1979 and 1985.

[13] The media have focussed considerable attention on California's declining commitment to funding education. The state's ranking in terms of per-pupil spending fell from 13 in 1974 to 26 in 1985. Correspondingly, California used

to spend more than the national average and now is spending less. It is worth noting, however, that throughout this period California has spent within five percent of the national average, which suggests that any decline in education spending relative to that in other states has not been particularly dramatic.

14. These figures are cited in Osman and Gemello, 1984.

15. There still is some room for school districts to exceed their *Serrano* limits. Since the *Serrano* case dealt only with the inequitable distribution of real property, it placed an upper limit on funding from property taxes only. School districts can raise additional funds using alternative taxes, and several school districts have passed parcel taxes that tax all homeowners an equal dollar amount. Piedmont has instituted a somewhat different parcel tax in which the tax liability is based on lot size. Barring further litigation, school districts may raise funds through taxes on anything except the value of real property, and may spend the proceeds as they wish.

16. Total revenues fell 16.5 percent. General revenues, which include only those that come with "no strings attached", fell 24.4 percent. The difference between the two is enterprise revenues, which consist of user fees. As property tax revenues fell, counties had an incentive to increase these fees (for water, sewer, utilities, etc.) in order to make up for lost property tax revenues, so general revenues fell more than did total revenues.

17. These figures exclude the City and County of San Francisco. The joint city-county nature of the government makes comparisons with other California cities misleading.

18. Counties and school districts did not suffer as much as cities did from reduced federal grants. For example, federal grants comprised 27 percent of county general funds in FY 1978 and 22 percent in FY 1985. The share of federal grants in school funding actually grew from 7.0 to 7.2 percent.

19. Personal income for California cities was estimated by multiplying California personal income by the proportion of state population residing in cities for each year. This accurately represents city personal income if incomes do not differ between incorporated and unincorporated areas, and accurately represents changes in personal incomes if income growth rates are equal in incorporated and unincorporated areas. Since neither of these conditions is likely to hold exactly, the measure can only approximate city personal incomes. During the period in question, between 70 and 75 percent of Californians resided in incorporated cities.

20. Specifically, during the fiscal years 1981 and earlier, cities reported revenues from certain functions inconsistently. For example, some reported "water" revenues under general revenues, while others reported a special "enterprise fund" for water services. The "general revenues" described in Chart 5 are those reported by cities so they include some service charge revenue. Beginning in 1982, reporting conventions were standardized and all revenues were reported either as "general" or as "functional". Under the new definitions, no service charge revenue is included as general. The data in the chart reflect total revenues (general plus functional) minus service charge revenues.

21. Proposition 13 did not limit local governments' authority to increase nonproperty tax rates (*Farrell v. San Francisco,* 32 Cal. 3rd 47, 1982). Later, in 1986, California voters passed Proposition 62, which stipulates that any tax increase is subject to approval of residents by a majority vote. However, Proposition 62 applies only to general law cities, and not to the 82 charter cities where most Californians reside.

22. In California, state law allows cities to receive revenues of up to 1.00 percent of taxable sales. This is part of the statewide 6 percent sales tax. Counties receive revenues not claimed by cities and revenues generated in unincorporated parts of the county, in addition to their own 0.25 percent transportation tax allocation.

23. In 1978 there were only 28 cities. Dublin in Alameda County and Danville and San Ramon in Contra Costa County were incorporated between 1978 and 1985.

24. Hercules (Contra Costa County) was excluded from the calculations. Hercules is unusual in that its per capita revenues in 1978 were over twice those of Emeryville, the second highest per capita spender, and were almost nine times the average of the other 27 cities. This was due to sales tax revenues that accounted for over half of total revenues. The excessive sales taxes ended in FY 1980. Since they had nothing to do with Proposition 13 and since including Hercules affected overall results substantially the city was omitted from calculations to make the sample more representative.

REFERENCES

Beebe, Jack. "Spirit of 13," *Weekly Letter,* Federal Reserve Bank of San Francisco, September 28, 1979.

California Department of Finance, *California Statistical Abstract,* Sacramento, 1986.

California Office of the Controller, *Annual Report of Financial Transactions Concerning Cities of California,* Sacramento, Fiscal Years 1973-74 through 1984-85.

California Office of the Controller, *Annual Report of Financial Transactions Concerning Counties of California,* Sacramento, Fiscal Years 1973-74 through 1984-85.

California Office of the Controller, *Annual Report of Financial Transactions Concerning School Districts of California,* Sacramento, Fiscal Years 1973-74 through 1984-85.

California Senate Local Government Committee, "Long-Term Local Government and School Financing," in *Implementation of Proposition 13,* volume II, Sacramento 1979.

California State Board of Equalization, *Annual Report,* Sacramento, Fiscal Years 1977-78 and 1984-85.

California Tax Foundation, "Up to the Limit: Article XIIIB Seven Years Later." Sacramento, CA, March 1987.

Oakland, William H. "Proposition 13: Genesis and Consequences," in George G. Kaufman and Kenneth T. Rosen, eds., *The Property Tax Revolt: The Case of Proposition 13.* Cambridge, MA: McGraw-Hill, 1981.

Osman, Jack W. and John M. Gemello. "California School Finance: Policy Perspectives," in John J. Kirlin and Donald R. Winkler, eds., *California Policy Choices.* Sacramento Public Affairs Center and USC School of Public Administration, 1984.

Tiebout, Charles M. "A Pure Theory of Local Expenditures," *Journal of Political Economy,* October 1956.

United States Department of Commerce, Bureau of the Census, *Census of Governments,* Volume 4 Number 5 (Compendium of Government Finances), Washington DC, Government Printing Office, 1967, 1972, 1977, 1982.