Choices in Banking Policy

J. Alfred Broaddus, Jr.

It is a great pleasure to have the opportunity to meet all of you and to share with you some of my thoughts on issues related to banking—what I like to refer to as "banking policy" as distinct from monetary policy. When I first joined the Federal Reserve way back in 1970, I did research on banking issues, and in fact my doctoral thesis had to do with banking policy. As time passed I was drawn ever more heavily into the monetary policy area. But I have never lost interest in banking issues, and my new role at our Bank obviously gives me ample reason, to put it gently, once again to give this broad and challenging area a very high priority in my personal work schedule.

My remarks this afternoon will summarize some of the conclusions I've reached regarding major issues currently facing banks, bankers, and regulators. First, I will say a little about the efficiency of bank regulation as it exists today, how it should be evaluated, and how it might be improved. Second, I will look at some of the trade-offs between the regulatory burden we'd all like to reduce, on the one hand, and the scope of the federal safety net, on the other. Finally, I will comment briefly on consumer and community reinvestment issues, which are receiving especially intense attention presently.

I begin with the fundamental idea that financial system arrangements are generally most efficient if left to private choice. This is merely a corollary to the well-known presumption in favor of unfettered competition in unconstrained markets. The unique characteristics of banking and finance sometimes cause people to lose sight of the applicability of this principle to these industries, but the extraordinary recent innovations in banking and financial markets should convince skeptics of the power of a competitive financial system, given the chance, to seek and find the most cost-effective means of intermediating between borrowers and lenders.

This article is adapted from an address by J. Alfred Broaddus, Jr., president of the Federal Reserve Bank of Richmond, at the Bank's Baltimore Branch on August 19, 1993. Jeffrey M. Lacker, research officer, contributed substantially to the article. The views expressed in the article are the author's and not necessarily those of the Federal Reserve System.

1. REGULATORY EFFICIENCY

Banking is one of the most heavily regulated of all industries. To determine the principles that ought to guide the design of banking regulation, we need first to ask why that regulation exists.

In my view, the strongest rationale for bank regulation derives from federal deposit insurance, discount window lending, and the Fed's involvement in the payments system. Together, these three activities are often referred to as the "federal safety net," but I find it most useful to think of them as credit enhancements provided by the federal government to the banking system. Deposit insurance is a third-party guaranty, analogous to standby letters of credit, private mortgage insurance, and other forms of credit insurance. Discount window lending is similar in many respects to a collateralized line of credit, and the credit extension inherent in the Fed's participation in the payments system is basically a clearinghouse overdraft facility.

Private providers of similar credit enhancements generally restrict the portfolio choice and risk-taking activities of recipients, since they recognize that third-party commitments often give rise to problems of "moral hazard"—to use a term coined in the insurance industry—which refers to the tendency of insured entities to take greater risks than they otherwise would. So, too, I think most people would agree that the government needs to constrain the portfolio choice and risk-taking activities of banks in order to protect the federal safety net from moral hazard.

A key point here, however, is that *competition among private credit enhancement providers forces them to minimize the burden of the restrictions they impose*. If one provider offers a guaranty with significantly more restrictive covenants than competitors offering the same guaranty, business is likely to be slow. On the other hand, a provider offering a guaranty with insufficient restrictions on borrower activity is likely to lose money steadily over time due to excessive risk taking by its customers. Competitive pressure ensures that constraints on the activities of recipients are *efficient* in the sense that they tend toward the minimum burden on borrowers consistent with actuarial soundness of the enhancement.

In my view, bank regulations should be efficient in exactly the same sense: that is, they should be just restrictive enough to protect the actuarial soundness of the federal safety net. Again, on the one side, insufficient restraint on bank activities could subsidize excessive bank risk taking and impose an unacceptable cost on taxpayers, the ultimate provider of federal credit enhancements. At the same time, however, excessive restraint on banking imposes needless costs on our financial system, increases the spread between borrowing costs and depositor returns, and ultimately risks reducing economic growth.

From this perspective, some aspects of current bank regulation clearly are flawed and in need of revision. For example, remaining restrictions on interstate

banking almost certainly could be eliminated without endangering the safety net. Indeed, a strong argument can be made that interstate banking would *reduce* risk to the safety net by allowing improved diversification of regional risks. Certainly, one of the fundamental banking lessons of the last decade ought to be the high risk of region-specific economic shocks to an industry that, for all of the structural changes that have occurred, is still dominated by local and regional institutions. It is regrettable that in the absence of federal legislation, we are forced to await a long and cumbersome process of statutory revision at the state level. Dismantling the existing barriers to interstate banking while preserving the competitiveness of banking at the local level ought to continue to be an important legislative priority.

Similarly, legislative restraints on bank entry into related financial markets are difficult to justify. The Glass-Steagall Act erects barriers between banking and commerce and between banking and securities markets. The barriers between banking and securities markets are said to be needed to prevent conflicts of interest, but our basic supervisory process seems quite capable of policing these—as it does now, for example, with trust departments—without the draconian prohibitions of Glass-Steagall. These barriers are often rationalized as risk-containment measures for the protection of the federal safety net, but, in the case of securities market activities, research has failed to support this claim. Fortunately, we have been able to ease some of these restraints at the regulatory level, but clearing away anachronistic federal statutory constraints in this area would make sense.

As I think you know, the Fed supports further relaxation of restrictions on interstate banking and bank powers, and these reforms, of course, were part of the Treasury Department's comprehensive proposals that led eventually to the Federal Deposit Insurance Corporation Improvement Act of 1991, or "FDICIA," as it is usually called. FDICIA was in large part a reaction by Congress to the perception that regulatory restrictions on bank risk taking were inadequate to protect the federal safety net. Regardless of whether that perception is fully justified or not, many of the Act's provisions, such as the requirement of prompt corrective action in the case of undercapitalized institutions, strike me as sound public policy and important steps forward. Other parts of the Act, however, failed to consider the costs of regulations that go well beyond what is required to protect the deposit insurance funds. Section 132 of the Act, in particular, which requires federal banking agencies to set so-called safety and soundness standards regarding operations, management, asset quality, earnings, stock values, and even employee compensation, seems clearly excessive. The actual language of this section is not much more specific than this, but it appears to envision rigid, predetermined rules for banks' internal management arrangements, irrespective of an individual bank's capital position. Such rules are not generally found in privately provided credit enhancements and, in my judgment, would constitute unnecessary, intrusive, and potentially

harmful micromanagement of any adequately capitalized bank. In short, Section 132 appears to raise the burden of bank regulation beyond the minimum level necessary to protect the federal safety net. The basic soundness of bank management has always been an important focus of the examination process. But it would be counterproductive to substitute mechanical formulas for the considered judgment of seasoned examiners, just as it would be undesirable to substitute mechanical credit approval rules for the considered judgment of seasoned loan officers.

With all of this in mind, let me just say that the Fed and the other federal bank regulators are striving to fulfill the intent of the law as efficiently as possible in implementing this section of the Act—in other words, with the smallest possible burden on the financial system. More generally—and somewhat ironically—another part of FDICIA, Section 221, directed the Federal Financial Institutions Examination Council (FFIEC) to review all banking regulations to determine whether they impose unnecessary burdens on regulated institutions and to make recommendations to reduce such burdens. The Fed and the other agencies that comprise the Council have completed this review and have already made a number of changes designed to reduce the burden of existing regulations. Beyond this, the Interagency Statement on Credit Availability issued in March, 1993, attempts to target exemptions from documentation requirements for better-capitalized institutions. This represents a step to build on in improving the efficiency of banking regulations by applying regulations more selectively to individual banks based on their capital.

All of these actions are constructive. It is important that regulations be refined on a continuing basis to improve their efficiency. There are limits, however, to the improvements that can be made in the context of the current statutory environment. In this regard, Federal Reserve Governor LaWare's suggestion—that an independent, nonpolitical commission be created and charged with developing a legislative agenda that would deal with regulatory burden in the broader context of the changing competitive condition of the banking industry—seems to merit greater attention than it has received to date. Some of you, recalling the legislative process that produced FDICIA, may reasonably wonder whether a broad banking reform effort can ever succeed. I don't have an easy answer to that question, but I do believe that the effort should be made and that an independent commission is a useful suggestion.

2. THE TRADE-OFF BETWEEN REGULATORY BURDEN AND THE SCOPE OF THE SAFETY NET

While we must constantly strive for the least costly and most efficient regulations to support the *existing* safety net, we also face broader choices in banking policy. Even if we were to achieve the least burdensome regulations consistent

with the actuarial soundness of the safety net as it exists today, as a society, we might still conclude that the costs exceed the benefits the safety net provides. Further reductions in bank regulations could then be sought by reducing the governmental credit enhancements the regulations are designed to protect—that is, by reducing the extent of the federal safety net. Private providers of credit enhancements typically allow less restrictive constraints for less extensive guaranties. For example, less onerous loan covenants are required of a borrower with lower leverage. Similarly, if the federal safety net were scaled back, we could reduce the regulatory burden on banks.

Two related and frequently overlooked provisions of FDICIA take important steps in this direction. First, FDICIA requires the FDIC to select the least-cost method of resolving failed depository institutions and to document its decision. This is important because least-cost failure resolution can reduce the extent to which uninsured depositors are implicitly insured at a higher cost to the insurance funds—in other words, it can limit the *implicit* scope of deposit insurance and the safety net. Second, FDICIA contains provisions designed to discourage Federal Reserve discount window lending to critically undercapitalized institutions and in some circumstances it imposes losses on the Fed in the event a borrower fails. These provisions seek to prevent discount window loans from artificially prolonging an institution's life and allowing uninsured claimants to continue withdrawing their funds at the expense of the FDIC. While these provisions have yet to be tested by the actual failure of a large institution, they should work to limit the scope of the "too-big-to-fail" doctrine and heighten the monitoring incentives of uninsured claimants, which would strengthen the case for easing bank regulation, especially the Section 132 variety.

In my opinion, perhaps the most disappointing aspect of FDICIA was the omission of any significant reduction in explicit deposit insurance coverage. A strong argument can be made that even apart from "too-big-to-fail," the coverage of federal deposit insurance is excessive from the standpoint of the incentives it creates (and doesn't create) among bank managers and bank customers and the risk it presents to the deposit insurance fund and ultimately the taxpayer. Reducing the extent of explicit deposit insurance coverage would provide a compelling reason for significant reductions in the regulatory burden on banks.

Many bankers and others naturally consider suggestions to reduce deposit insurance coverage dangerous because such a reduction might undermine public confidence in the banking system. Beyond this, many community bankers worry that it would weaken their competitive position in the industry if vestiges of "too-big-to-fail" remain in place.

These concerns are reasonable and understandable. After numerous increases in coverage over many years, capped by the sharp rise from $40,000 to $100,000 per account in 1982, reversing course might indeed reduce public

confidence in the short run. My own view, however, is that public confidence and the competitive positions of all banks would be strengthened over the longer pull, especially since the public is now much more conscious of the hazards and risks associated with deposit insurance in the wake of the savings and loan debacle. I believe the public is fully capable of understanding that reducing deposit insurance coverage would reduce risk in the banking industry by increasing (1) the degree to which depositors monitor the riskiness of individual banks and (2) self-regulation by the industry. I think it is very much in the longer-term interest of *all* bankers, whether from large banks or small ones, to help persuade the public of this view. The alternative is public demand for still more costly and burdensome legislation and regulation to protect the insurance fund. The latter seems to me to be clearly a bigger risk to the health of the industry than the immediate reaction to scaling deposit insurance back.

Other critics may claim that reducing explicit deposit insurance coverage would increase the risk of bank runs and panics like those of the 19th century. While 19th-century American banking lacked deposit insurance, it also lacked a central bank acting as lender of last resort. The Fed can prevent banking panics by supplying liquidity promptly and generously through the discount window and open market operations as the events surrounding the October 1987 stock market crash convincingly demonstrated. Scaling back deposit insurance would in no way diminish the ability of the Federal Reserve to stem financial panics.

The reason I am making so much of the need to reduce explicit deposit insurance coverage in one way or another is that I doubt very much that really meaningful regulatory relief—relief you can feel—will occur in the absence of such a reduction. Fortunately, some progress has been made in laying a foundation for reducing coverage in the future. For example, the FDIC, as mandated by FDICIA, recently completed a study of the feasibility of "tracking" the ownership of deposits by individuals across banks in order to gauge the feasibility of restricting coverage to one account per depositor. This is an important initiative, one I hope will be pursued. The FDIC has also studied the feasibility of partially privatizing federal deposit insurance. The FDIC would sell a portion of its deposit coverage in the private reinsurance market. This sale, in turn, would establish a market price for the insurance and indicate the restrictions private markets would impose on insured institutions. Also, private insurers are now offering supplemental deposit insurance directly to depositors. If such market arrangements prove viable, their availability might make reductions in FDIC deposit insurance coverage more palatable.

Before leaving the subject of the federal safety net, let me turn briefly to banking policy and the Fed's role in the payments system. Since its founding, the Federal Reserve has played a central role in the nation's payments system, and that role encompasses extensions of credit as well as transactional operations. Although it does not receive as much public attention as deposit insurance, there has been a growing awareness in recent years of the importance

of Federal Reserve credit and implicit guaranties to the payments system. This increased attention led initially to the introduction of specific regulatory constraints on payments system users, such as net debit caps for institutions participating in the Federal Reserve's Fedwire electronic funds transfer system. Further, under the Program for Payments System Risk Reduction, the Fed is reexamining the terms for such credit. As you know, the Fed will soon introduce a fee for daylight overdrafts in the reserve and clearing accounts of depository institutions, which is designed to increase the reliance on market forces to regulate the volume of intraday credit.

Payments system policy should continue to focus on the extent of the explicit and implicit guaranties the Fed provides and to strive to make the constraints on participants appropriate to the scale of the guaranties. As in the case of deposit insurance, financial market efficiency might well be improved by a more proscribed Fed credit exposure with consequently less encumbering regulatory constraints. The prospect of continued rapid technological advance in this area of banking lends weight to this view. It would be unfortunate indeed if the implementation of operationally more efficient payments system arrangements were stymied by regulatory schemes more appropriate to earlier technologies.

3. CONSUMER AND COMMUNITY AFFAIRS ISSUES

Let me turn finally to consumer and CRA issues. Obviously, no discussion of public policy toward banks would be complete without consideration of this increasingly important and, in some respects, contentious area. I can really only scratch the surface here. Consumer and CRA regulations may be viewed by some as a sort of quid pro quo for the benefits banks receive from deposit insurance and access to the discount window. Unlike basic safety and soundness regulation and supervision, however, community and CRA regulations play no direct role in protecting the safety net and therefore are not likely to be eased in response to scaling back the safety net.

As I see it, consumer and CRA laws and regulations have two basic purposes. First, consumer laws and regulations seek to ensure that lenders respect the basic legal rights of consumers in credit transactions—and most importantly that they not discriminate against particular prospective borrowers on the basis of sex, race, age, and so forth. Secondly, CRA regulations aim at encouraging and helping banks meet the credit needs of the communities in which they operate, especially for housing and community development purposes and always, of course, within basic safety and soundness constraints. These are not only reasonable but laudable objectives that reflect this nation's most cherished values. Few if any bankers dissent from these objectives.

There is, however, disagreement—and I think legitimate and understandable disagreement—regarding the detailed character of these regulations and the

way they are implemented in practice. Let me offer just a couple of comments in this regard.

First, credit markets, including markets for bank credit, generally allocate credit very efficiently among all creditworthy borrowers. With this in mind, regulators, and also consumer and community reinvestment activists and legislators, need to understand what you already understand all too well—that unduly burdensome, intrusive, and costly consumer and community reinvestment laws and regulations can well reduce the flow of credit and increase its costs unnecessarily to the very constituencies that activists, legislators, and regulators are trying to protect and assist. This is an instance of what Fed Governor Larry Lindsey calls the Law of Unintended Consequences, and unintended consequences are not at all unlikely in this area. The implication, of course, is a need for regulatory—and also legislative—restraint: adding new consumer and CRA laws and regulations only when there is a clear and compelling reason to do so, minimizing their intrusiveness, and continuously reviewing existing laws to find ways to reduce the burden they impose.

The second point I want to make is simply that we at the Federal Reserve Bank of Richmond want to do all we can to facilitate your compliance with consumer and CRA regulations and reduce the burden they impose on you. We see this as a fundamental regulatory obligation. I can guess how most of you react to someone who tells you he's from Washington and he's here to help you. At least I only have to say that I'm from Richmond and I'm here to help you. In any case, we have an active consumer and community affairs operation at our Bank that is separate from our examination staff. Our consumer and community affairs staff analyze local economic conditions across the District, with particular emphasis on the credit needs and development opportunities of moderate- and low-income households and communities. They offer specific and detailed information—both through conferences and in published form—designed to assist you in your compliance efforts. I hope you will take advantage of this assistance and let us know whenever we can be helpful to you in this area.

4. CONCLUSION

To quickly summarize the main points I've tried to make: First, regulations should be efficient, and since protecting the safety net is one of the central reasons for bank regulation, one way to promote regulatory efficiency is to try to aim for the minimum regulatory burden consistent with maintaining the actuarial soundness of the *existing* safety net. Second, there is a trade-off between the scope of the safety net and the burden of even the most efficient regulatory system. Consequently, beyond some point, reducing regulatory burden requires a reduction in the scope of the safety net and, in particular,

the coverage of the deposit insurance system. Finally, since consumer and CRA regulations have objectives other than protecting the safety net, they must be evaluated on different criteria. But activists who promote them, legislators who enact them, and regulators who implement them should be keenly aware of the Law of Unintended Consequences and the possibility that excessive zeal ultimately may be counterproductive. Attention to these points, I believe, can significantly enhance the contribution that necessary banking regulation can make to the economy's strength and its ability to grow.

Corporate Capital Structure: The Control Roles of Bank and Public Debt with Taxes and Costly Bankruptcy

Douglas W. Diamond

Corporate finance theory studies the way that firms choose to raise funds. Traditionally, this theory focused on the effect of capital structure on income tax payments and exogenously specified administrative costs of bankruptcy. More recently, this theory has emphasized the effect of capital structure on the control of subsequent investment decisions of the firm, in settings where managers' and investors' incentives are not perfectly aligned. Both the tax-oriented approach and the control-oriented approach capture important aspects of the decision that firms make when they choose a method of finance. To date, however, the insights from the two theories have not been integrated. Tax-oriented theories typically ignore issues of corporate control, while control-oriented theories typically ignore taxes. In addition, tax-oriented theories consider only a firm's choice between debt and equity, while some of the control-oriented theories study the importance of the source of debt finance: the choice between bank loans (privately placed debt) and bonds (publicly issued debt).

This article combines traditional tax-based capital structure theory with an analysis of the control and incentive effects of debt. It presents a model of both the firm's choice of the amount of debt and equity and its choice between bank loans and publicly traded debt. Following the traditional approach, capital structure choice is framed as a trade-off between tax savings of debt and costs of bankruptcy. Accounting for the control roles of bank loans and public debt

The author, the Theodore O. Yntema Professor of Finance at the University of Chicago, Graduate School of Business, is grateful for helpful comments from Peter Ireland, Thomas Humphrey, Jeffrey Lacker, Merton Miller, and John Weinberg. The views expressed are those of the author and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

emphasized in more recent work then allows for the endogenous determination of bankruptcy costs. The model shows how the costs of bankruptcy can sometimes be negative (so bankruptcy becomes a net benefit), when bankruptcy allows claim holders to prevent a borrower from undertaking an unprofitable investment.

Endogenous bankruptcy costs depend on the type of debt used and the characteristics of the borrower. One relevant borrower characteristic is the correlation between the return from past investments and the profitability of new investment. If this correlation is high, then the borrower will be unable to refinance debt only when its old and new investments are both unprofitable, so inability to refinance indicates that new investment is unprofitable and bankruptcy desirable. If the correlation is low, then the inability to refinance is not a clear indicator of poor prospects for new investment, and bankruptcy due to the inability to refinance will sometimes be quite costly.

Modigliani and Miller (1958) established the framework for studying capital structure by finding apparently reasonable conditions rendering a firm's capital structure irrelevant to its value. The earliest generalization was Modigliani and Miller (1963), which viewed capital structure as an attempt to reduce taxes. They studied the implications of a tax advantage to debt over equity that still exists in the United States. Corporate taxes are avoided for interest payments but not for dividends. If there are no other advantages to equity over debt, the conclusion is that firms should issue no equity and should issue debt with face value equal to the highest possible future value of the firm. Such all-debt firms would almost always default on their debt. Modigliani and Miller assumed that there was no cost associated with frequent default.

The next generalization in the literature assumes that there is an exogenous cost of default—a bankruptcy cost. This bankruptcy cost is a disadvantage of issuing too much debt that is traded off against the taxes saved. This forms the basis for the traditional "trade-off" approach to capital structure in Robichek and Myers (1965) and Kraus and Litzenberger (1973). This approach does not analyze the source of such bankruptcy costs or allow any non-tax benefits of debt. It identifies volatility of firm value as a force that limits debt. It predicts that firms with high-variance cash flow distributions will choose less debt and more equity than those with low variance. It also predicts that firms will be financed only with equity when there is no corporate income tax advantage to debt. Little empirical support for these implications exists, however.

Recent approaches to capital structure view the capital structure as influencing the investment decisions of the firm, either by providing incentives to management (see Jensen and Meckling [1976], Townsend [1979], Diamond [1984], and Gale and Hellwig [1985]) or by allocating some control of the firm to someone other than the management. When capital structure serves to transfer control from management to bondholders, one obtains a theory of the debt-to-equity ratio, as in Aghion and Bolton (1992), Bolton and Scharfstein

(1993), Diamond (1991a, 1993b), Hart and Moore (1989, 1990, 1991), Jensen (1986, 1989), Stulz (1990), and Titman (1984). Other research studies how financial contracts allocate control between management and bank lenders, as in Diamond (1984, 1991b, 1993a). Such work provides a theory of the characteristics of firms that use bank finance instead of issuing securities directly to the public. These recent approaches often ignore taxes and bankruptcy costs, however, because capital structure has important effects even without taxes or costs of bankruptcy. Since taxes and bankruptcy costs do exist, it is important to see how they interact with the phenomena described more recently.

This article integrates the bondholder control and bank control views into the tax savings versus bankruptcy cost approach to optimal capital structure. I allow for the effects of a debt default on the transfer of control of firm operating decisions. Default has different effects for publicly issued debt and bank debt. The costs of financial distress are specified as three separate components: the costs of restructuring defaulted public debt, the cost of ceasing a firm's operations, and the cost of lost going-concern value if a firm enters bankruptcy and then reorganizes. An optimal capital structure is determined by the interaction of these costs with the tax and restructuring advantages of equity, public debt, and bank debt. I use the term *bank debt* as a shorthand for privately placed debt, including debt held by insurance companies and other financial intermediaries.

The balance of this article is organized as follows. Section 1 describes both the tax savings from issuing debt rather than equity and the cost differences between bank debt and debt issued directly to the public. Section 2 outlines a model of capital structure choice. It begins by using the model to illustrate the results of traditional capital structure theory based on a trade-off of tax savings versus fixed bankruptcy costs. It describes the component costs of default on debt. The costs of defaulting on public debt and on private debt are analyzed in the two subsections under Section 2. Section 3 shows how the correlation between the cash from existing investments and the profitability of new investment influences the amount of debt and the type of debt a firm will choose to issue. Section 4 discusses the conclusions and implications that one can draw from the model.

1. THE TRADITIONAL THEORY

The older capital structure theories frame capital structure as a choice that balances the tax savings from debt against the exogenous bankruptcy costs incurred when there is default on debt. The model in this article is framed within this trade-off, in order to learn how the insights from the traditional approach interact with the newer, control-oriented approach. Before showing how to frame the newer approach in the context of the traditional approach, a simple capital structure model without control elements is presented.

Tax Savings Due to Debt

The tax advantage of debt over equity is due to the deductibility of interest payments from corporate income tax. Dividends and retained earnings are not deductible. If the firm's investors are not subject to different personal taxes for debt and equity, the corporate tax savings is the only tax effect of capital structure.[1] I assume that corporate taxes are a fraction t of corporate profits and that there are no personal taxes. A one dollar payment to equity costs the firm one dollar, and is worth one dollar to the investor. A one dollar payment of interest to a public debt holder costs the firm $1 - t$ dollars, because it reduces taxable income by one dollar. The interest payment is worth one dollar to the investor. Thus, there is an increase in the firm's after-tax profit of t when one dollar of payments to equity is replaced by one dollar of payments to debt. This increased profit makes debt a lower-cost form of capital than equity.

The model considers two types of debt: bank loans and public debt. Payments to the holders of either are deductible from corporate income. There are cost-of-capital differences, however, because the bank incurs operating costs and corporate taxes of its own. In addition, banks are subject to expenses that are equivalent to taxes, such as reserve requirements and Federal Deposit Insurance Corporation (FDIC) premiums in excess of the value of deposit insurance. Reserve requirements are a tax because no interest is paid on reserves, and FDIC premiums in excess of the value of deposit insurance increases a bank's cost of funding itself with deposits. Let the sum of the bank's added costs and taxes be denoted by z, per dollar of its income. A one dollar payment of bank interest saves t in corporate tax for the firm, but incurs bank taxes and costs of $z \geq 0$. The net savings from replacing a one dollar payment to corporate equity with a one dollar payment on a bank loan is then $t - z$. Bank loans are more costly than public debt, but have a lower cost of default, which is described later. Bank debt is, on balance, less costly than equity: I assume that $t > z$.

To keep the notation simple, I will overstate the tax advantage of debt by assuming that principal as well as interest payments are deductible from corporate tax. No qualitative results depend on this simplification.

The Model

On date 0, the firm chooses a capital structure. On date 1, several events occur. The cash flows from the firm's previous investments arrive. The firm faces new

[1] If investors are subject to differential individual taxes, there is a tax advantage to debt if the sum of individual and corporate tax is lowest for debt payments (Miller 1977). The personal tax advantages of equity are due to low taxation of capital gains and deferral of unrealized capital gains. I will formally introduce only corporate tax savings and assume that the investors are tax-exempt, but the corporate tax rate can be interpreted as the net corporate and personal tax saving of payments to debt over payments to equity.

investment opportunities and chooses a new investment. Finally, both public and bank debt contracts mature. The firm can pay its debts with the cash from its investments and from the proceeds obtained from issuing new securities. If the firm continues operations after date 1, it is liquidated at date 2, with residual claims going to equity owners in proportion to their ownership.

The firm chooses a date-0 capital structure to maximize its market value. The firm can issue either public or bank debt. Let the face value of public debt be *R*. Public debt must be fully repaid or there will be bankruptcy. The United States Federal Trust Indenture Act makes it difficult to restructure out of court because a vote to forgive or extend the debt requires unanimous consent (see Roe [1987] and Gertner and Scharfstein [1991]). While there are methods of restructuring public debt to avoid a default, these are costly and sometimes unsuccessful.

Instead of public debt, the firm can issue bank debt (get a bank loan), with face value denoted by *r*. Bank debt can be renegotiated, with the possibility of avoiding bankruptcy. I do not allow combinations of the two types of debt. Focusing on the choice between the two types of debt simplifies the analysis without producing misleading results. A bank's incentive to extend maturity and restructure debt is removed when combined with a large amount of public debt (see Bulow and Shoven [1979], Gertner and Scharfstein [1991], and Diamond [1993a, 1993b]). In addition to either type of debt, the firm can issue equity, a claim that requires no fixed date-1 payment. Equity is a proportional claim on any and all dividends the firm may declare, but the firm has no legal obligation to pay dividends in any period that it is not being liquidated. I assume that the firm will not be liquidated until date 2, absent outside intervention. The date-2 value depends on the firm's manager's decisions on date 1, as well as on past decisions.

The market value on date 0 of a date-1 cash flow is its discounted present value. I assume, for simplicity, that all investors are risk-neutral and that interest rates are zero, implying that the discounted present value is just the expected value of the cash flow distribution.[2]

The next two subsections review traditional tax-oriented capital structure theory where the control role of debt is absent. To illustrate the added implications of the control role of debt, I will review traditional capital structure theory, which allows no control role. This will provide a framework for understanding the control role of debt.

[2] Alternatively I could assume that there are complete Arrow-Debreu markets, implying that there is a market price today for every risk. This allows market prices to provide appropriate discount rates for any risk. In this case one replaces the probability of a given cash flow with the market price of one dollar delivered in the situation in which the cash flow is equal to that amount.

Review of Traditional Capital Structure Theory Without Bankruptcy Costs

The traditional approach to capital structure abstracts from issues related to the control of the firm's future investment decisions. Thus, consider the simple case in which the firm is liquidated at date 1 because it has no new investment opportunities. Assume that all debt is public debt and bankruptcy has no cost. The only role for capital structure is to minimize taxes.

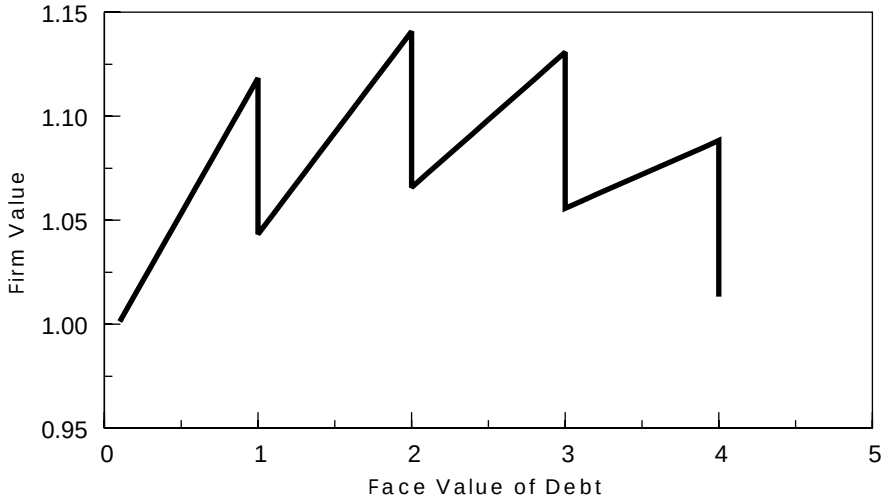
The date-0 value of the firm if unlevered (all equity) is then the discounted value of the after-tax profits. The pre-tax value of the firm on date 1, c, has possible realizations $c \in \{1, 2, 3, 4\}$. Each realization has equal probability, $P = ¼$. The market value of the unlevered firm is $(P \cdot 4 + P \cdot 3 + P \cdot 2 + P \cdot 1)(1 - t) = (2.5)(1-t) \equiv V^u$. With debt of face value $R \leq 1$, the firm can always deduct the payment from its corporate taxes, saving Rt, and firm value is $V^u + Rt$. Define τ_R as the date-0 present value of tax savings from increasing debt to R from the largest integer value less than R. This means that $\tau_1 \equiv t$ is the value of taxes saved with debt equal to one. Further increasing debt to a face value $R \in (1, 2]$, increases the value to $V^u + \tau_1 + (R - 1)(3Pt)$. The added taxes are saved only when the firm is worth more than one, because only payments made are tax-deductible. Therefore, increasing debt from one to two saves $3Pt = ¾t \equiv \tau_2$. Similarly, $\tau_3 = t/2$ and $\tau_4 = t/4$. The increase in value from a unit increase in R decreases for higher values of R. Further increases in R save more taxes until the firm's value is maximized with $R = 4$ and firm value is 2.5.

Fixed Bankruptcy Costs

Suppose that there is a fixed cost ϕ that is incurred whenever the firm cannot fully repay its public debt (see Robichek and Myers [1965] and Kraus and Litzenberger [1973]). Think of ϕ as an unavoidable legal fee. The cost of bankruptcy trades off against tax savings to determine the value-maximizing capital structure. There is no risk of bankruptcy for debt with face value $R \leq 1$. Value increases to $V^u + \tau_1$ with $R = 1$. Further increasing the face value from $R = 1$ to $R = 2$ increases date-0 firm value by $\tau_2 - P\phi$. Increasing leverage beyond one decreases firm value if the present value of tax savings is less than that of bankruptcy costs. Because taxes are only saved for payments actually made, the marginal value of tax saving per unit of debt is reduced as debt climbs ($\tau_4 < \tau_3 < \tau_2 < \tau_1$). If $\tau_4 < P\phi$, then eventually tax savings are smaller than bankruptcy costs, and there is a limit to desired leverage. Figure 1 shows the effect of leverage on firm value under the traditional capital structure theory.[3] The firm value drops by the present value of bankruptcy costs at each positive integer value. Bankruptcy costs are sufficiently large in Figure 1 to imply that the optimal value of public debt is $R = 2$.

[3] The example assumes that $t = .13$ and $\phi = .3$.

Figure 1 Traditional Capital Structure Theory



Note: Firm value drops by the bankruptcy cost at each positive integer value. Bankruptcy costs are sufficiently large to imply that the optimal value of public debt is $R = 2$.

If bankruptcy costs are nontrivial, traditional capital structure theory implies that firms with high variance of value will have low leverage. Without corporate tax, the model predicts that there will be no debt issued. The crucial assumptions are that there are no effects of capital structure on the firm's decisions and that the cost of bankruptcy is the same for all bankruptcies. In what follows, future decisions are introduced by allowing the firm an investment choice at date 1. Profitable investment is a source of firm value in addition to its cash from previous investments. The firm will be in default only when the sum of the cash from old operations and the net present value of new investment is less than the amount of debt to be repaid. Before providing these details, the next section describes the costs and benefits of bankruptcy.

2. CONTROL AND THE BENEFITS OF DEBT

There are conflicting interests between the management of the firm and its outside investors. The management derives more benefits than do outsiders from the firm's growth and its continued operations. Some reasons for this conflict include the costs of a manager's immediate lost reputation if operations are closed and the increase in the manager's incremental value to the company once a project is undertaken (the manager's information is needed to most

profitably continue the project, even if the ex-ante net present value is negative). These control benefits imply that management will continue to invest even if investment prospects are bleak. The prospects of future investments cannot be costlessly observed by a court, but the prospects are observed by investors at date 1; the manager has no private information. A management incentive contract that required a court to determine the profitability of each investment would be expensive to enforce. Because outside investors observe profitability, they can prevent the manager from making a bad investment if, and only if, they have control of the firm. Investors have control only if the firm defaults on its debt. Default on public debt will require the use of bankruptcy court, but default on bank debt need not. Equity contracts have no terms that can trigger a transfer of control (I assume that a takeover is not a possibility). If the firm is financed exclusively with equity, outsiders never have control and the firm will always invest. If the firm cannot fully repay its debt obligation, then the firm cannot avoid a default and the owners of the debt can take control of the firm. The details of this process are described in the next two subsections.

The firm's net present value of new investment at date 1, N, will be one of two possible values: $N = N_G > 0$, a good investment, or $N = N_B < 0$, a bad investment. Management will prefer to invest in either case. There is a gross benefit of $-N_B$ from defaulting on debt and preventing investment decision when investment is unprofitable and $N = N_B$. The firm ought to be liquidated when $N = N_B$, but this can only be done in bankruptcy. There are no gross benefits of defaulting on debt and controlling investment when $N = N_G$. There are also costs of using bankruptcy court, described below.

The net costs of using bankruptcy court depend on the type of reorganization that is needed and the type of debt that the firm has. The administrative costs are as follows:

1. Entering into formal bankruptcy proceedings reduces the going-concern value of the firm's future investments. These are lost reputation and physical costs. These costs are only relevant if the firm reorganizes after filing for bankruptcy. This cost is denoted by γ and is incurred under bank debt or public debt.

2. There are costs of closing and quitting operations. These costs must be incurred if the firm ceases to operate and do not depend on the type of financial contracts the firm has. These are the costs of breaking other contracts, such as leases, if the firm ceases to operate. This quitting cost is denoted by q and is incurred under bank debt or public debt.

3. There are legal costs of restructuring or renegotiating public debt issues. These costs are incurred if the firm gets into formal bankruptcy proceedings without fully repaying its public debt. The costs also can be interpreted as costs of restructuring public debt outside formal

bankruptcy. The magnitude of the cost can depend on whether the firm reorganizes or quits operations; the costs are denoted by k_g and k_q, respectively. No such costs are incurred in restructuring bank debt.

The Costs and Implications of Bankruptcy Initiated by Default on Public Debt

A default on public debt incurs administrative bankruptcy costs of $\gamma + k_g$ if the firm continues as a going concern and $q + k_q$ if there is liquidation. Liquidation then yields $c - q - k_q$, whereas reorganizing as a going concern yields $c + N - \gamma - k_g$. The U.S. Bankruptcy Law requires a vote of the lenders to choose the reorganization plan, suggesting that the more valuable option will be selected. I assume that the bad investment is sufficiently unprofitable that it is worth incurring bankruptcy costs to avoid it: $N_B < -(q + k_q)$. This implies that net bankruptcy costs of public debt ($q + k_q + N_B \equiv B$) are negative when $N = N_B$, on account of the control role of debt. Net bankruptcy costs are $\gamma + k_g \equiv G$ when the firm is reorganized and continues operations. I assume that the good investment is sufficiently profitable that it pays to reorganize to undertake it, i.e., $N_G > G$, and that the firm will restructure. Unlike public debt, bank debt can be restructured outside bankruptcy. The restructuring of bank debt is analyzed next.

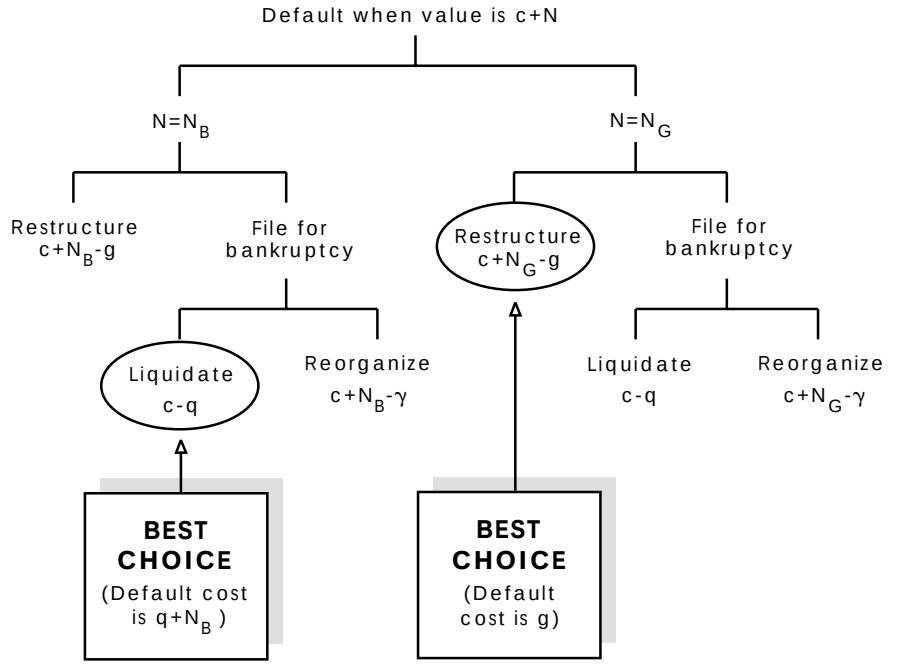
Default on Bank Debt: Bankruptcy Versus Renegotiation

If a default is on bank debt, the bank can choose to renegotiate rather than force bankruptcy. The bank will renegotiate rather than force bankruptcy when its payoff from renegotiating exceeds what it will get in bankruptcy. When the firm is worth more as a going concern because $N = N_G$, the bank will renegotiate and save the costs of bankruptcy. When liquidation is desired because $N = N_B$, the bank will initiate bankruptcy and liquidate. A decision tree that illustrates this choice by the bank is given in Figure 2. It illustrates the bank's decision process which is described in the next two paragraphs.

The bank's payoff in bankruptcy is as follows. The value of the firm if it is reorganized as a going concern is $c + N - \gamma$ (saving k_g compared to public debt). If the firm is liquidated, the value is $c - q$ (saving k_q compared to public debt). In bankruptcy, the bank chooses the option with the largest value. Therefore, bankruptcy yields the bank the larger of $c + N - \gamma$ and $c - q$. Given a firm in bankruptcy, the bank reorganizes the firm if and only if $N \geq \gamma - q$. This implies that the bank reorganizes a bankrupt firm when it has good investments, $N = N_G$, achieving a payoff of $c + N_G - \gamma$. When a bankrupt firm has unprofitable investments, $N = N_B$, the bank liquidates, achieving a payoff of $c - q$.

I assume that the bank has substantial bargaining power. The bank can make a take-it-or-leave-it offer to the borrower to reschedule outside bankruptcy when the borrower does not pay in full. I assume that the borrower

Figure 2 The Cost of Default on Bank Debt



gets nothing in bankruptcy and will accept any offer that deters the bank from forcing bankruptcy when it otherwise would choose to file.[4] It is possible that the bank's rescheduling is costly; let g denote this cost. I assume that $g < \gamma$ so that it is cheaper to reschedule a going concern outside bankruptcy. The rescheduling cost includes the cost of rewriting and renegotiating contracts. The bank's payoff if it reschedules the debt is $c + N - g$. It reschedules if this payoff exceeds the larger of $c + N - \gamma$ and $c - q$. Since $c + N - g$ exceeds $c + N - \gamma$, this implies that the bank will restructure whenever $N > g - q$. The bank will restructure when $N = N_G$, but will force bankruptcy when $N = N_B$.[5] The savings from having bank debt instead of public debt in the event of a potential default are then k_q when $N = N_B$ (because the bank also

[4] The results of the model are robust to giving the borrower some bargaining power and thus a positive payoff in bankruptcy. If the borrower gets a payoff of Δ in bankruptcy, the bank's take-it-or-leave-it offer must provide that borrower a payoff of Δ outside bankruptcy. One can then reinterpret N as net of the claim Δ that the borrower can appropriate.

[5] This follows from the assumption in the previous subsection that $N_B < -(q + k_q) < g - q$ and $N_G > \gamma + k_g > g - q$.

uses bankruptcy) and $G-g$ when $N = N_G$ (because the bank then avoids bankruptcy). One expects that the major savings are due to avoiding bankruptcy for a going concern and that bank rescheduling costs are low.

One can easily extend the model to cases where there are more general managerial incentive problems in the firm. Suppose that instead of just continuing to invest when only poor investments are available, when $N = N_B$, management's objectives differ from outsiders in other ways. Management might choose an investment that is not the most profitable (absent outside intervention). The model can be reinterpreted in such a way that the transfer of control from a default leads to a change in the chosen investment instead of a liquidation.

When the bank reschedules with $N = N_G$, bank debt serves a role in avoiding bankruptcy costs that is similar to equity (which has no fixed claim that can lead to a default). When the bank forces bankruptcy with $N = N_B$, it removes cash from management's control, similar to the role of public debt described in Townsend (1979), Diamond (1984), Gale and Hellwig (1985), Lacker and Weinberg (1989), and Jensen (1986, 1989), at lower ex-post cost than does public debt. Both types of debt have the advantage over equity of blocking undesirable investment by the firm.

3. THE LINK BETWEEN CASH FLOW AND THE NET PRESENT VALUE OF NEW INVESTMENT

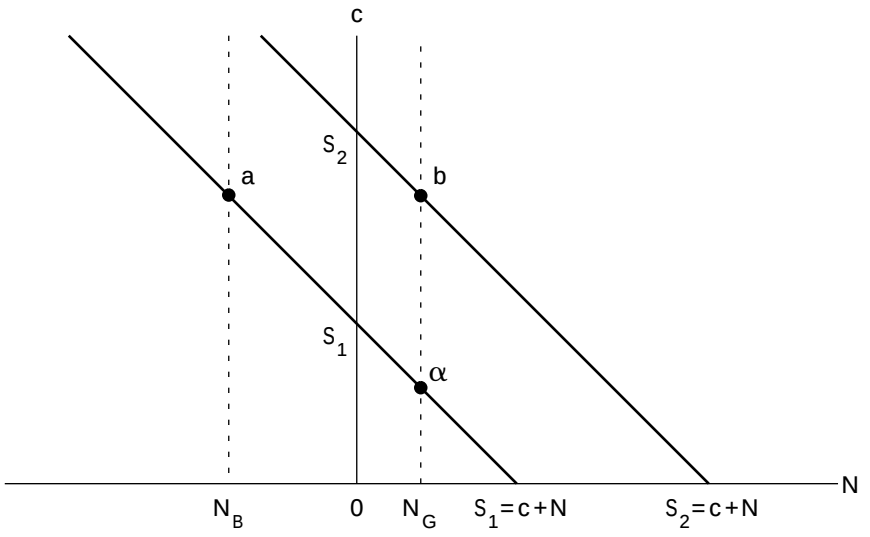
I now consider a more general model in which at date 1 the firm will have cash of c and will face new investment opportunities with net present value of N. The sum of these, $c + N \equiv S$, is the total date-1 value of the firm if it continues operations. Note that the firm is able to borrow to finance its future investments and, if these are sufficiently profitable, use the proceeds to pay off old debt. Therefore, S is the maximum that the firm is able to pay to claimants on date 1. If the firm cannot raise enough to pay off old debt, then lenders have control and can block the firm from continuing to invest. The firm will be able to fully repay its debt when S equals or exceeds the face value of the maturing debt. If S is less than the amount of debt due, there will be bankruptcy if the debt is public. If, instead, the debt is a bank loan, there will be bankruptcy if investment prospects are bad ($N = N_B$) and restructuring outside bankruptcy if investment prospects are good ($N = N_G$).

Default on public debt is desirable (the net bankruptcy cost, B, is negative) when investment prospects are bad. Public debt leads to only desirable defaults if the firm can choose a debt level such that there is default if and only if the prospects of future investment are bad. Default if and only if prospects are bad requires that total firm value, S, falls below the face value of debt, R, if and only if investment prospects are bad. Such a public debt level exists if the correlation between N and $S = N + c$ is perfect. For example, if N varies, but not c, the correlation between value, S, and prospects, N, is perfect and the

ability to refinance reveals just investment prospects. In this case one can choose an amount of debt below $c + N_G$ but above $c + N_B$ that will control investment decisions without generating any defaults with $N = N_G$. This relationship is illustrated in Figure 3. A line $c + N = S$ shows those combinations of c and N that imply the ability to refinance debt of face value S. Suppose that only N varies and the possible realizations of c and N are denoted by two horizontally aligned points such as the points marked a and b. An amount of public debt less than $S_1 = c + N_B$ never leads to default and thus fails to prevent bad investment. An amount of public debt exceeding $S_2 = c + N_G$ leads to costly defaults when $N = N_G$. Therefore, public debt exceeding $S_2 = c + N_G$ would be selected only for its tax advantages.

In the general case where cash flow is random, it might be impossible to avoid undesirable defaults when $N = N_G$ without some failure to default when $N = N_B$. Referring to Figure 3, if the point marked α is a possible realization (along with points a and b), then a face value of public debt greater than S_1 results in desirable default for realization a but undesirable default for realization α. Any face value less than or equal to S_1 implies failure to trigger default for realization a, allowing the firm to make bad investments. In general,

Figure 3 The Ability to Refinance Depends on S ($S = c + N$)



Note: The ability to avoid default on debt depends on S, the sum of cash (c) and net present value of future investment (N). The correlation between S and N determines how much information about N is revealed by a default. If all three points a, b, and α are possible, then there exists no amount of debt such that there is default if and only if $N = N_B$. If only a and b are possible and the point marked α is impossible, then for debt with face value between S_1 and S_2, a default reveals a low value of N.

it will be impossible to find an amount of public debt that results in default if and only if the firm's investment prospect has a negative net present value. The cost of using public debt to control the firm's investment choice depends on the correlation between cash flow from previous investments and the value of future investments. Public debt is a very good control device if prospects, N, and cash from old investments, c, are both stochastic but N is much more variable than c and they have a nonnegative correlation. A related condition that makes public debt a low-cost control device is if both c and N are variable and if they are sufficiently positively correlated. For sufficiently high correlation between c and N, one can choose a face value of debt, R, such that $S \geq R$ if the probability that $N \geq N_B$ is arbitrarily high, and if $S < R$ the probability that $N = N_G$ is arbitrarily low. On the other hand, public debt is an expensive control device if c is quite variable and c and N are uncorrelated or c is negatively correlated with N. Under either condition, there is a low correlation between S and N, and many low realizations of S imply good investment prospects ($N = N_G$) while many high realizations of S imply bad investment prospects. A low correlation between S and N implies that a high probability of costly bankruptcy is required to obtain a high probability of beneficial bankruptcy that controls unprofitable investment. Referring back to Figure 3, it will be more expensive to use default on public debt to stop investment when $N = N_B$ if the point marked α is a possible realization along with a and b.

To examine the effects of the correlation between the total firm value, $S = c + N$, and the profitability of new investment, N, suppose that there are four possible date-1 realizations of S. The values of S are one, two, three, and four. Each realization has equal probability, $P = ¼$. There is a positive, but possibly imperfect, correlation between S and N. Table 1 describes the conditional distribution of total firm value, S, given the net present value of new investment, N. When firm value is very low ($S = 1$), investment prospects are bad ($N = N_B$). When firm value is very high ($S = 4$), investment prospects are good ($N = N_G$). For intermediate values of S, either value of N is possible. A correlation parameter, u, a number between zero and one, describes how uncorrelated are S and N. Increased values of u reduce the correlation between S and N. The probability that new investment is profitable ($N = N_G$) when firm value is somewhat low ($S = 2$) is u. The probability that new investment is profitable when firm value is somewhat high ($S = 3$) is $1 - u$. When $u = 0$, S and N are perfectly correlated.[6]

[6] The discussion in the text, combined with the definition $S = c + N$, implies the following about the value of c given each value of S. When $S = 1$, $N = N_B$ and $c = 1 - N_B$. When $S = 4$, $N = N_G$ and $c = 4 - N_G$. When $S = 2$ or $S = 3$, either value of N is possible. When $S = 2$, the pair $N = N_G$, $c = 2 - N_G$ occurs with probability u, and the pair $N = N_B$, $c = 2 - N_B$ occurs with probability $1 - u$. When $S = 3$, the pair $N = N_G$, $c = 3 - N_G$ occurs with probability $1 - u$, and the pair $N = N_B$, $c = 3 - N_B$ occurs with probability u.

Table 1 The Conditional Distribution of N Given $S \equiv c + N$

	u=0	$u \in$ (0,1)	u=1
$S = 1$	$N = N_B$	$N = N_B$	$N = N_B$
$S = 2$	$N = N_B$	$N = N_B$ with probability $= 1 - u$ $N = N_G$ with probability $= u$	$N = N_G$
$S = 3$	$N = N_G$	$N = N_B$ with probability $= u$ $N = N_G$ with probability $= 1 - u$	$N = N_B$
$S = 4$	$N = N_G$	$N = N_G$	$N = N_G$

The date-0 value of the firm's cash flows is independent of u, but the correlation between S and N is decreasing in u. Increasing u decreases the correlation between cash flow and the profitability of new investment (because reducing the correlation between $S = c + N$ implies reduced correlation between c and N). Many of the implications of changing the level of the correlation parameter, u, can be seen by comparing the case of $u = 1$ with $u = 0$. The next subsection explores these implications in the case in which the firm makes use of public debt.

The Optimal Quantity of Public Debt

The value of the firm with public debt R depends on the net bankruptcy costs and tax savings of the chosen capital structure. The possible values of total firm value, $S = N + c$, are denoted by i, and $i \in \{1, 2, 3, 4\}$. Let X_i denote the net, non-tax bankruptcy cost from defaulting on public debt when $S = i$. This is a real cost from debt with face value R exceeding i. Recall that the (negative) cost of bankruptcy when investment prospects are bad ($N = N_B$) is $N_B + q + k_q \equiv B$. The (positive) cost of bankruptcy when investment prospects are good ($N = N_G$) is $\gamma + k_g = G$. The probability distribution of N given S described in Table 1 implies that the bankruptcy costs for each value of S are as follows: $X_1 = B$, $X_2 = uG + (1 - u)B$, $X_3 = uB + (1 - u)G$, $X_4 = G$.

Let $\Pi(R)$ denote the total date-0 value of a firm with public debt of face value R. The date-0 firm value, Π, depends on the value of tax savings from debt with face value R and the (possibly negative) net costs of bankruptcy, X_i. Since bankruptcy costs are incurred only if R exceeds one, $\Pi(1) = V^u + \tau_1$. Increasing R from one to two garners an incremental tax benefit of $\tau_2 = \frac{3}{4}t$ and incurs a (negative) bankruptcy cost of $\frac{1}{4}X_1$. Thus $\Pi(2) = \Pi(1) + \tau_2 - \frac{1}{4}X_1$.[7]

[7] Similarly, $\Pi(3) = \Pi(2) + \tau_3 - \frac{1}{4}X_2$ and $\Pi(4) = \Pi(3) + \tau_4 - \frac{1}{4}X_3$.

On account of the tax savings from debt, firm value $\Pi(R)$ is increasing in R whenever incremental bankruptcy costs are non-positive ($X_R \leq 0$). Because bankruptcy is desirable when $S = 1$ ($X_1 < 0$), the optimal value of R exceeds one and the minimum optimal value of R is two (because there is no effect on the probability of bankruptcy of increasing debt between one and two). The optimal face value of public debt is equal to either two, three, or four, because increasing R in between these values saves taxes and has no effect on bankruptcy costs. Proposition 1 characterizes the optimal amount of public debt, the amount that maximizes the date-0 value of the firm.

Proposition 1 The value-maximizing face value of public debt, R^*, is given as follows:

$$R^* = 2 \text{ if } t < \min\left\{\frac{G+B}{3}, \frac{uG+(1-u)B}{2}\right\}.$$

$$R^* = 3 \text{ if } t > \frac{uG+(1-u)B}{2} \text{ and } t < uB+(1-u)G.$$

$$R^* = 4 \text{ if } t > \max\left\{\frac{G+B}{3},\ uB+(1-u)G\right\}.$$

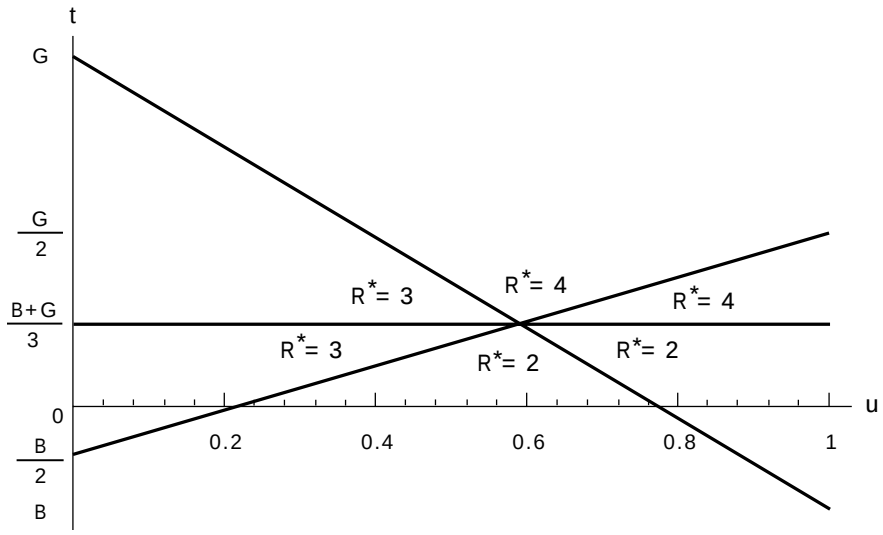
Proof: See Appendix.

Figure 4 shows an example of the results of Proposition 1 by giving the optimal face value of public debt for the possible values of the tax saving from debt, t, and the correlation parameter, u.[8]

A way to give a simple interpretation of Proposition 1 is to consider the characterization of the optimal level of public debt when tax effects dominate and when they do not dominate. One way to describe capital structure choice when taxes do not dominate is to examine the debt quantities that are selected when tax savings from debt are absent ($t = 0$). In this case, Proposition 1 implies that there is a critical value of the correlation parameter, $u = u'$, that determines whether debt with face value $R = 3$ is optimal. Debt with face value three is optimal if $u \leq u'$, and another value is best for $u > u'$. If $G + B > 0$ (default costs given good prospects are bigger than the benefits of default when prospects are bad), then the level of debt when $u > u'$ is $R = 2$ and the critical value, u', is given by $u' = -B/(G - B)$. If, instead, $G + B < 0$ (default costs given good prospects are less than the benefits of default when prospects are bad), then the best level of debt for $u > u'$ is $R = 4$ and $u' = G/(G - B)$. The value of the firm is weakly decreasing in u in either case, because firm value given debt of $R = 3$ is decreasing in u and is independent of u for other values of R.

[8] The example assumes that $G = .65$ and $B = -.2$.

Figure 4 Optimal Face Value of Public Debt, R^*



Note: t is the tax rate on corporate profits. The parameter u describes how uncorrelated are total firm value, S, and the net present value of new investment, N. Increased values of u reduce the correlation between S and N. See Table 1.

If taxes are sufficiently large, $t > G$, then high leverage ($R = 4$) dominates for all values of u.[9] The tax savings then dominate default costs regardless of the correlation structure of value and investment prospects. I assume that $t < G$, which implies that the magnitude of the correlation between total firm value, S, and the net present value of new investment, N, influences the optimal amount of public debt and the cost of using public debt as a control device.

The higher cost of using public debt as a control device when u is high and the correlation of S and N is low is illustrated in Figure 5. Figure 5 plots date-0 firm value, Π, as a function of R (the face value of public debt) for the cases of $u = 0$ (high correlation) and $u = 1$ (low correlation). The example in the figure assumes a high cost of going bankrupt when investment prospects are good, relative to tax savings of debt ($\frac{1}{3}[G + B] > t$), so that a capital structure of all public debt, $R = 4$, is not the optimum.[10]

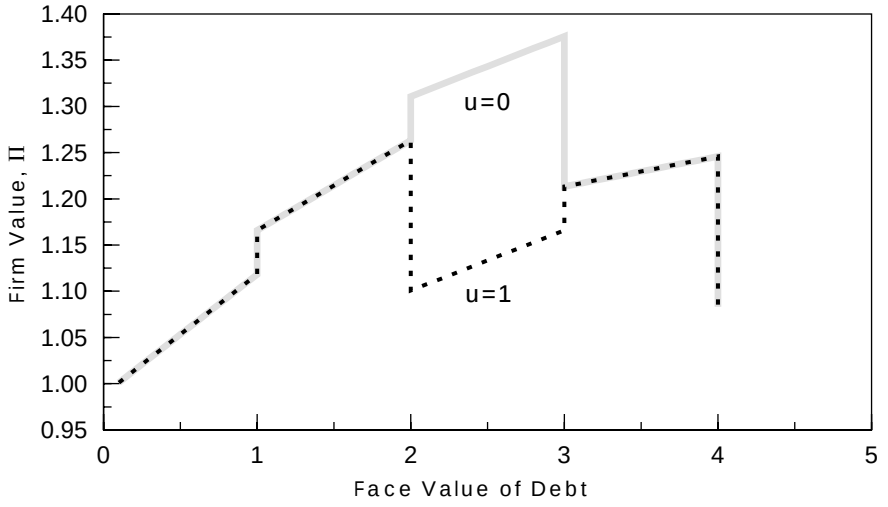
[9] Note that because $G > 0 > B$, $[uB + (1 - u)G] \leq G$ and $(G + B)/3 \leq G$. Thus $t > G$ implies $R^* = 4$: tax savings from increasing debt from three to four are greater than the maximum bankruptcy cost, G.

[10] The example assumes that $t = .13$, $\gamma = .65$, $k_g = 0$ (implying $G = .65$), $q = .2$, $k_q = 0.01$, $N_B = -.4$ (implying $B = -.2$), and $N_G = .8$.

Figure 5 shows that when $u = 0$ (high correlation between S and N), date-0 firm value is maximized with debt of three. With public debt having a face value of $R = 3$, the firm defaults if its date-1 value, S, is one or two. Defaulting when S is equal to one or two is beneficial because investment prospects are bad ($N = N_B$) in either case, and the control benefit of stopping a bad investment exceeds the administrative costs of using bankruptcy court. Default is avoided when S is equal to three or four, and for both values of S investment prospects are good ($N = N_G$).

Figure 5 also shows that when $u = 1$ (low correlation between S and N), date-0 value is maximized with debt of face value two. When $S = 2$, the firm has good investment prospects ($N = N_G$) and bankruptcy is costly. When $S = 3$, the firm has bad investment prospects and bankruptcy is beneficial. However, the costs of bankruptcy when $S = 2$ are sufficiently large that they outweigh the benefits of bankruptcy when $S = 3$. Debt with face value two avoids bankruptcy when S is either two or three. This results in higher date-0 firm value than debt with face value of four (which would result in bankruptcy both for $S = 2$ and $S = 3$).

Figure 5 Firm Value Given Public Debt When $u = 0$ and $u = 1$



Note: When $u = 0$ (high correlation between S and N), date-0 firm value is maximized with debt of three. This leads to default only if the firm has bad future investments; in addition, it saves taxes. When $u = 1$ (low correlation between S and N), date-0 value is maximized with debt of face value two. When the firm can repay exactly two, it has good investment prospects ($N = N_G$) and bankruptcy would be costly. This cost exceeds the benefits of debt with face value exceeding three (which would lead to bankruptcy when the firm can repay exactly three and has bad investment prospects).

To control the investment decision with debt implying default when $S = 3$, the debt must also be in default for all lower values of S. When the correlation between S and N is high (u is nearly zero), low firm value (S below three) implies the need to control investment and high firm value (S above three) implies no need for control. When the correlation between S and N is lower (u is nearly one), then there is need to control investment for the relatively high firm value, $S = 3$, but little or no need to control investment decisions for the relatively low firm value, $S = 2$. Low correlation implies that it is costly to default when $S = 2$, but default when $S = 2$ is necessary in order to induce a beneficial default when $S = 3$. Decreased correlation (an increase in u) between firm value, S, and the net present value of new investment, N, increases the cost of using public debt to control investment when $S = 3$ and decreases the benefits.

When $u = 0$ (high correlation between S and N), the optimal amount of public debt is $R = 3$. When $u = 1$ (low correlation between S and N), the optimal amount of public debt is $R = 2$ (because the example assumes $t < [G + B]/3$). This implies that there exists a value of u, denoted by $\hat{u} = (2t - B)/(G - B) \in (0, 1)$, such that $R = 3$ is optimal for all $u < \hat{u}$, and $R = 2$ for all $u > \hat{u}$.

Increasing the correlation between cash flow from old investment and the profitability of new investment will generally increase the optimal amount of public debt and will increase the date-0 value of the firm.[11] When the correlation is low, public debt is an expensive control device. If the tax benefits of debt are not extremely high, firms with low correlation will choose low debt when given a choice between public debt and equity. The next subsection examines the cost of the alternative of bank debt.

The Optimal Quantity of Bank Debt

Resolving default is less costly with bank debt than with public debt. The discussion in Section 2 entitled "Default on Bank Debt: Bankruptcy Versus Renegotiation" establishes that the total cost of resolving a default when investment prospects are good ($N = N_G$) is g for bank debt, a saving of $G - g$ over the resolution cost given public debt. This is a large saving because the bank avoids bankruptcy court when the firm is worth more as a going concern. The total cost of resolving a default when investment prospects are bad (and $N = N_B$) is $N_B + q$ for bank debt, because a bad investment with net present value of N_B is avoided, but unavoidable administrative costs of q are incurred.

[11] Cases in which decreasing u (increasing the correlation between S and N) decreases debt occur as described immediately after Proposition 1. This requires that control aspects of debt are very valuable ($B << 0$) relative to the cost of bankruptcy when investment prospects are good, and $B + G < 3t$. The debt decrease from four to three for low values of u occurs because there is then little need for control when $S = 3$.

Define this cost of resolving default on bank debt when investment prospects are bad as $b \equiv N_B + q$. A default on bank debt when investment prospects are bad saves $B - b = k_q$ compared to public debt. The saving, k_q, represents the bank's comparative advantage in bankruptcy court. This cost saving is probably smaller than the saving when investment prospects are good, because bankruptcy court is not avoided.

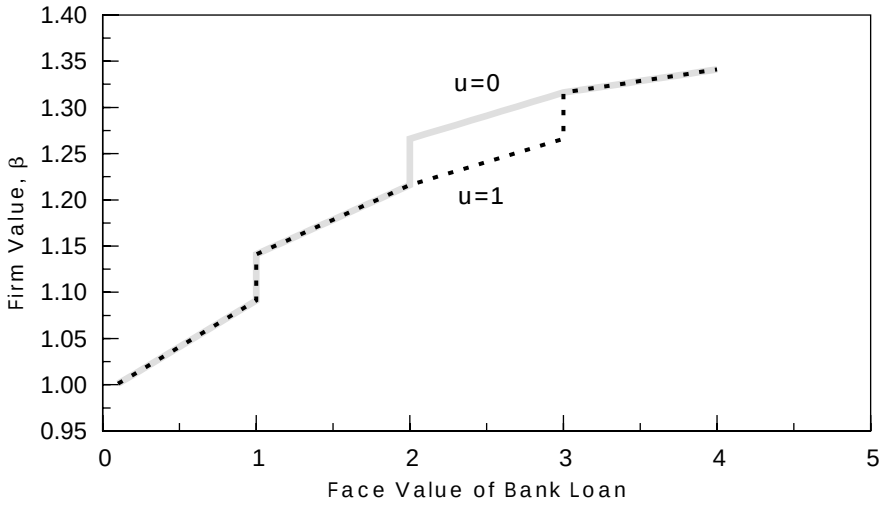
The analysis of the optimal quantity of bank debt is similar to that of the optimal quantity of public debt, except the costs of default are those described in the previous paragraph and the bank has a higher cost of capital, as discussed in Section 1. Let x_i denote the net cost of defaulting on bank debt when date-1 firm value, S, is equal to i. Because investment prospects are bad when $S = 1$ ($N = N_B$), the cost of default on bank debt when $S = 1$, x_1, is equal to b. Because investment prospects are good when $S = 4$ ($N = N_G$), the cost of default on bank debt when $S = 4$, x_4, is equal to g. Recall that given total date-1 firm value, S, equal to two, the probability that $N = N_G$ is u (and the probability that $N = N_B$ is $1 - u$). The cost of default when $S = 2$ is then $x_2 = u \cdot g + (1 - u)b$. Given total date-1 firm value, S, equal to three, the probability that $N = N_G$ is $1 - u$ (and the probability that $N = N_B$ is u). The cost of default when $S = 3$ is then $x_3 = u \cdot b + (1 - u)g$. The results of Proposition 1, which describes the optimal amount of public debt, can be used to determine the optimal amount of bank debt. Substitute for the bankruptcy costs X_i the bank's costs x_i, for $i = 1, 2, 3, 4$, and instead of tax savings t, use $t - z$ to take account of the bank operating costs.

The optimal level of bank debt is less sensitive than is public debt to the correlation between firm value and the prospects for new investment. Banks have default cost advantages over public debt when investment prospects are good, which implies that undesirable defaults have a smaller effect on the value of the firm than with public debt. Figure 6 shows an example where the cost of reorganizing bank debt when prospects are good is less than the tax savings (net of bank costs) from added debt, or $g < t - z$. This assumption implies that optimal bank leverage is $r = 4$, independent of u. In this case, the optimal level of bank debt is $r = 4$ both for $u = 1$ and $u = 0$.[12] Similarly, $r = 4$ is the optimal bank debt level for all values of correlation between total date-1 firm value and net present value of new date-1 investment (all u between zero and one).

Let the value of the firm with bank debt, as a function of the amount of debt, r, be given by the function $\beta(r)$. Because I make the simplifying assumption of sufficiently low costs of reorganizing bank debt when investment prospects are good, the optimal value of bank debt is $r = 4$, and the value of the firm if it chooses the optimal bank debt is $\beta(4)$. The choice between bank and public

[12] The figure assumes $g = 0$, $z = .1$, and all of the parameters defined in footnote 10.

Figure 6 Firm Value Given Bank Debt When $u = 0$ and $u = 1$



Note: The optimal level of bank debt is $r = 4$ both for $u = 0$ and $u = 1$ because the costs of reorganizing a default on bank debt are less than the tax savings from additional debt. (The figure, but not the analysis in the article, assumes that there is no cost to reorganizing a default on bank debt.)

debt is equivalent to comparing this firm value, $\beta(4)$, to the date-0 firm value with the optimal level of public debt and choosing the form of debt leading to higher firm value. This comparison is discussed in the next subsection.

Bank Debt Versus Public Debt

If bank operating costs are too high, then public debt will dominate even if bank debt has default cost advantages. Similarly, if banks' default cost advantages are large, then bank debt will dominate even for rather large operating costs. For moderate levels of bank operating costs and bank debt default cost advantages, the optimal choice will depend on the characteristics of the borrower. In particular, the choice can depend on the correlation between future firm value, S, and future investment prospects, N.

Proposition 2 gives conditions where one type of debt dominates the other for all values of the correlation between S and N and then characterizes debt choice in the intermediate case where neither type of debt dominates the other. In this case, the choice depends on the parameter u, the degree to which S and N are uncorrelated.

Proposition 2 Bank debt is preferred to public debt if and only if $\theta < \min\{3t - B, t + u(G - B), G\}$, where θ is the value of bank operating costs minus the savings in default costs of bank debt of $r = 4$ versus public debt of $R = 4$ (θ is given by $\theta = 10z + g - 2k_q$).

If $\theta < \min\{t, G\}$, then bank debt is best for all values of u. If instead $\theta > \min\{3t - B, G\}$, then public debt is preferred for all values of u.

The choice between public and bank debt depends on u if $\theta < \min\{G, 3t - B\}$ and θ satisfies $t + G - B > \theta > t$ (this last condition is equivalent to $[\theta - t]/[G - B] \in (0, 1)$). In this case, bank debt is preferred if $u \geq u^*$ and public debt preferred for $u < u^*$, where the value of u^* is given by

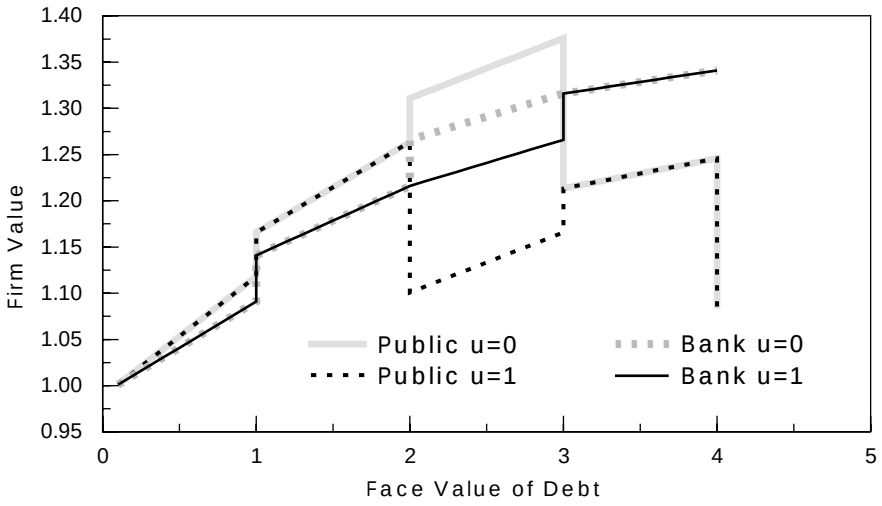
$$u^* = \frac{\theta - t}{G - B} = \frac{10z + g - 2k_q - t}{G - B}.$$

Proof: See Appendix.

Public debt dominates bank debt for all u and t if the added operating cost of bank funding minus the reorganization cost savings over those with public debt of face value $R = 4$ ($\theta = 10z + g - 2k_q$) exceeds G, the cost of a public debt default when investment prospects are good. If the cost of defaulting on public debt when prospects are good is sufficiently high, then bank debt is preferred when t is high and there are large tax savings of debt. For high tax savings, borrowers are driven to choose a debt structure where all types of default are likely, and banks' reorganization cost savings dominate independent of the level of u. At these high tax rates, the optimal level of public debt would be $R = 4$. If tax savings are lower, then bank debt is preferred if controlling investment when prospects are bad is important (B is low) and the prospects of new investment are not too correlated with total firm value (u is high). In this case, it is important to have high debt to control investment when prospects are bad, but it is not possible to do this with public debt in a way that avoids a significant chance of bankruptcy when investment prospects are good. Figure 7 superimposes Figures 5 and 6 to show an example where public debt is best when $u = 0$ but bank debt is best when $u = 1$.

Bank debt is more expensive, per unit, than public debt, because of bank costs and bank taxes. However, the borrowers that rely on bank debt are those that use large quantities of debt. This is because the non-tax cost disadvantages of public debt are most pronounced for very high levels of debt. If only low fractions of capital are raised with bank debt, it has few control advantages over public debt because bankruptcies that occur after very poor performance are not, on balance, costly.

Figure 7 Firm Value for Both Bank and Public Debt When $u = 0$ and $u = 1$



Note: Figure 7 superimposes Figures 5 and 6 to show an example where public debt is best when $u = 0$ but bank debt is best when $u = 1$.

4. IMPLICATIONS AND CONCLUSIONS

Traditional capital structure theory obtains strong results by framing the choice as a trade-off between tax savings and exogenous costs of bankruptcy. When there are no costs of bankruptcy, an all-debt firm is optimal; when there are bankruptcy costs but no tax savings, an all-equity firm is optimal. This article begins by reviewing these results and illustrates the intermediate case where the trade-off yields a capital structure containing both debt and equity. In this case, firms with more variable cash flows choose less debt.

The more recent control-based theories of capital structure have not been framed as representing a trade-off of tax savings against bankruptcy cost. In some cases both taxes and bankruptcy costs have been ignored for simplicity. This article draws on elements of this control-based theory to determine the costs of bankruptcy endogenously. This reveals that the costs of bankruptcy can sometimes be negative; there are situations when bankruptcy is beneficial to prevent management from initiating a bad investment project. Interpreting bankruptcy costs as sometimes including this control benefit of debt allows integration of many of the ideas in control-based theories and the ideas in the traditional theories.

Considering the control role of debt allows a comparison of bank debt and publicly issued debt. Relative to publicly issued debt, bank debt is more expensive because banks must cover many variable operating costs. But banks enjoy an offsetting cost advantage: they can restructure outside bankruptcy those firms that default but have access to viable investment projects. In light of this second cost trade-off, the correlation between cash flow and the net present value of future investment becomes another key determinant of optimal capital structure. If this correlation is low, a firm will often default on its debt when it has viable future investments, which leads bank debt to have a net cost advantage over public debt. If this correlation is high, then a firm will only rarely default on debt when it has good new investments because these two components that determine its ability to refinance its debt move together. In this case, public debt has a cost advantage over bank debt.

If there are large tax advantages of debt over equity, then firms will be induced to issue mainly debt. A firm that issues very large amounts of debt will default on its debt even when its cash flow is fairly high and its new investments are reasonably good. In that situation, the ability to restructure bank debt outside bankruptcy is beneficial. This implies that large tax advantages of debt lead firms to substitute toward bank debt and away from both equity and public debt.

In countries that have small tax advantages to debt finance, the model predicts that those that use bank debt will have a lower correlation between total firm value and the profitability of new investment than those that use public debt. Equivalently, firms that use bank debt will be those with a lower correlation between cash flow from previous investment and the profitability of new investment. This will show up empirically as a lower correlation between the cash flows from old investment and the amount on new investment they undertake, as compared with firms that rely on public debt. Hoshi, Kashyap, and Scharfstein (1990) find exactly this correlation structure in Japanese data comparing bank borrowers with firms that rely on public debt. They explain the higher correlation between cash from old investments and the level of new investment among firms that rely on public debt as evidence that firms are sometimes unable to raise funds when they have good prospects, forcing them to rely on internal funds. Their explanation is not inconsistent with the model in this article (firms with good prospects but low total value experience financial distress). However, firms choose between the two sources of finance based on the correlation between cash flow and the optimal amount of new investment. The correlation observed in the data might be generated not only by the financing constraints of those who rely on public debt, but also by the more informative signal that lagged cash flow provides about the profitability of new investment for firms that choose public debt.

Increasing the tax advantage of debt makes more borrowers prefer bank debt. Firms with higher correlation between total value and prospects for new

investment are induced to choose high leverage with bank debt where they would have chosen low leverage with public debt at lower tax advantages. One implication of this result is that in countries with large tax advantages to debt, bank lending will be pervasive. If all firms face high costs of reorganization with public debt, banks will attract customers who need debt for control but want to save reorganization costs, plus others who do not need debt for control purposes but just for its tax savings. In these countries where banks are predicted to dominate the debt market, a bank's average customer will have a stronger correlation between cash flow and the quantity of new investment, because the firms for which cash flow is strongly correlated with the profitability of new investment opportunities are included in the set of bank customers. I am not aware of empirical evidence on these implications. This type of implication shows the importance of simultaneously considering the tax, bankruptcy, and control roles of debt. Studying the interaction of the various roles of debt yields fresh interpretations of existing empirical evidence as well as entirely new implications.

APPENDIX

Proof of Proposition 1:

Define the function τ_R as the marginal value of taxes saved by increasing debt to face value R from a debt equal to the largest integer $i < R$. For example, if the face value R is less than one, τ_R is just the total tax saving. Similarly, if the face value R is between one and two, τ_R is the total tax saving minus τ_1. Let $I(R)$ denote the greatest integer less than or equal to R. The function τ_R is given by $\tau_R \equiv \sum_{i \geq R} P_i\{[R - I(R)] \cdot t\}$. The total value of tax benefits from debt with face R is then $\tau_R + \sum_{i<R} \tau_i$.

The date-0 value of a levered firm with public debt level R is the value of the unlevered firm, V^u, plus the tax savings, minus the bankruptcy costs. Let $\Pi(R)$ denote the total date-0 value of a firm with public debt of face value R. Recall that $P_i = P = ¼$. Firm value is given by $\Pi(R) = \tau_R + \sum_{i<R}[\tau_i - (¼X_i)] + V^u$, where $G = \gamma + k_g > 0$, $B = N_B + q + k_q < 0$, $X_1 = B$, $X_2 = (u \cdot G) + (1 - u) \cdot B$, $X_3 = [(1 - u)G] + (u) \cdot B$, $\tau_1 = t$, $\tau_2 = ¾t$, $\tau_3 = ½t$, and $\tau_4 = ¼t$.

The optimal face value is at least two, because $t > 0$ and $B < 0$ imply that $\Pi(R)$ is strictly increasing up to $R = 2$. The optimal value, $R^* \in \{2, 3, 4\}$, because $t > 0$ implies that $\Pi(R)$ is strictly increasing for $R \in (2, 3]$ and $R \in (3, 4]$. Finding the optimal value then involves comparing date-0 firm value, $\Pi(R)$, at these three values. The comparisons are as follows:

$\Pi(2) \geq \Pi(4)$ iff $G \geq 3t - B$, or $t \leq (G + B)/3$.

$\Pi(2) \geq \Pi(3)$ iff $t \leq \frac{1}{2}[u \cdot G + (1 - u) \cdot B]$.

$\Pi(3) \geq \Pi(4)$ iff $t \leq u \cdot B + (1 - u) \cdot G$.

The optimal value is $R^* = 2$ if $\Pi(2) \geq \Pi(4)$ and $\Pi(2) \geq \Pi(3)$, or $t \leq \min\{(G + B)/3, \frac{1}{2}[u \cdot G + (1 - u) \cdot B]\}$.

The optimal value is $R^* = 3$ if $\Pi(2) \leq \Pi(3)$ and $\Pi(3) \geq \Pi(4)$, or $t \geq \frac{1}{2}[u \cdot G + (1 - u) \cdot B]$ and $t \leq u \cdot B + (1 - u) \cdot G$.

The optimal value is $R^* = 4$ if $\Pi(2) \leq \Pi(4)$ and $\Pi(3) \leq \Pi(4)$, or $t \geq \max\{(G + B)/3, u \cdot B + (1 - u) \cdot G\}$.

Q.E.D.

Proof of Proposition 2:

Because the optimal value of bank debt is $r = 4$, public debt results in higher firm value if firm value, Π, with public debt of two, three, or four exceeds $\beta(4)$. Note that $\beta(r) = \tau_r - z_r + \sum_{i<r}[\tau_i - z_i - (P_i x_i)] + V^u$, where the τ_r functions are given in the proof of Proposition 1 and the other terms are as follows: $z_1 = z$, $z_2 = \frac{3}{4}z$, $z_3 = \frac{1}{2}z$, $z_4 = \frac{1}{4}z$, $b = N_B + q$, $x_1 = b$, $x_4 = g$, $x_2 = u \cdot g + (1 - u) \cdot b$, and $x_3 = u \cdot b + (1 - u)g$.

Firm value given bank debt is $\beta(4)$, given by: $\beta(4) = V^u + \frac{1}{4}[10(t - z) - 2(N_B + q) - g] = V^u + \frac{1}{4}[10(t - z) - 2(B - k_q) - g]$. Define $\theta = 10z + g - 2k_q$. The condition for $\Pi(2) \leq \beta(4)$ is $\theta < 3t - B$. The condition for $\Pi(4) \leq \beta(4)$ is $\theta < G$, which is independent of u or t. The condition for $\Pi(3) \leq \beta(4)$ is $\theta < t + u(G - B)$. Bank debt is preferred if and only if all three of these conditions are true, or $\theta < \min\{3t - B, t + u(G - B), G\}$. Bank loans are thus preferred for all $u \in [0, 1]$ if and only if this condition is true for $u = 0$, implying $\theta < \min\{t, G\}$, because $3t - B > t$. Public debt is preferred for all $u \in [0, 1]$ if and only if it is preferred for $u = 1$, implying public debt dominates if $\theta > \min\{3t - B, G\}$, because $t + G - B > G$. If neither of the two inequalities hold for θ, then the choice of lender depends on u; this requires that θ satisfy $t + G - B > \theta > t$, because $t < G$. This condition is equivalent to $1 > (\theta - t)/(G - B) > 0$. The critical value of $u = u^*$ satisfies $\theta = t + u^*(G - B)$, or $u^* = (\theta - t)/(G - B)$.

Q.E.D.

REFERENCES

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

Bolton, P., and D. Scharfstein. "Optimal Debt Structure with Multiple Creditors," Working Paper. Cambridge Mass.: MIT Sloan School, June, 1993.

Bulow, J., and J. Shoven. "The Bankruptcy Decision," *Bell Journal of Economics,* vol. 9 (Spring 1979), pp. 436–45.

Diamond, D. W. "Bank Loan Maturity and Priority when Borrowers Can Refinance," in C. Mayer and X. Vives, eds., *Capital Markets and Financial Intermediation*. Cambridge, England: Cambridge University Press, 1993a.

_____. "Seniority and Maturity of Debt Contracts," *Journal of Financial Economics,* vol. 33 (June 1993b), pp. 341–68.

_____. "Debt Maturity Structure and Liquidity Risk," *Quarterly Journal of Economics,* vol. 106 (August 1991a), pp. 709–37.

_____. "Monitoring and Reputation: The Choice Between Bank Loans and Directly Placed Debt," *Journal of Political Economy,* vol. 99 (August 1991b), pp. 689–721.

_____. "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies,* vol. 51 (July 1984), pp. 393–414.

Gale, D., and M. Hellwig. "Incentive Compatible Debt Contracts: The One-Period Problem," *Review of Economic Studies,* vol. 52 (October 1985), pp. 647–64.

Gertner, R., and D. Scharfstein. "A Theory of Workouts and the Effects of Reorganization Law," *Journal of Finance,* vol. 46 (September 1991), pp. 1189–1222.

Hart, Oliver, and John Moore. "A Theory of Debt Based on the Inalienability of Human Capital," Working Paper 3906. Cambridge, Mass.: National Bureau of Economic Research, 1991.

_____. "A Theory of Corporate Financial Structure Based on the Priority of Claims," Working Paper. Cambridge, Mass.: Massachusetts Institute of Technology, 1990.

_____. "Default and Renegotiation: A Dynamic Model of Debt," Discussion Paper 57. London: School of Economics, June 1989.

Hoshi, T., A. Kashyap, and D. Scharfstein. "The Role of Banks in Reducing the Costs of Financial Distress in Japan," *Journal of Financial Economics,* vol. 27 (September 1990), pp. 67–88.

Jensen, M. "The Eclipse of the Public Corporation," *Harvard Business Review,* vol. 67 (September–October 1989), pp. 61–74.

_____. "Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers," *American Economic Review,* vol. 76 (May 1986), pp. 323–29.

_____, and W. Meckling. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics,* vol. 3 (October 1976), pp. 305–60.

Kraus, A., and R. H. Litzenberger. "A State Preference Model of Optimal Financial Leverage," *Journal of Finance,* vol. 28 (September 1973), pp. 911–22.

Lacker, J., and J. Weinberg. "Optimal Contracts Under Costly State Falsification," *Journal of Political Economy,* vol. 97 (December 1989), pp. 1345–63.

Miller, M. H. "Debt and Taxes," *Journal of Finance,* vol. 32 (May 1977), pp. 261–76.

Modigliani, F., and M. H. Miller. "Corporate Income Taxes and the Cost of Capital: A Correction," *American Economic Review,* vol. 53 (June 1963), pp. 433–43.

_____. "The Cost of Capital, Corporation Finance and the Theory of Investment," *American Economic Review,* vol. 48 (June 1958), pp. 261–97.

Robichek, A. A., and S. C. Myers. *Optimal Financing Decisions*. New York: Prentice-Hall, 1965.

Roe, M. "The Voting Prohibition in Bond Workouts," *Yale Law Journal,* vol. 97 (December 1987), pp. 232–79.

Stulz, R. "Managerial Discretion and Optimal Financing Policies," *Journal of Financial Economics,* vol. 26 (July 1990), pp. 3–28.

Titman, S. "The Effect of Capital Structure on a Firm's Liquidation Decision," *Journal of Financial Economics,* vol. 13 (March 1984), pp. 137–52.

Townsend, R. M. "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory,* vol. 21 (October 1979), pp. 265–93.

The Free Trade Debate: The Illusion of Security Versus Growth

Robert L. Hetzel

The debate over Nafta, the North American Free Trade Agreement, exposed deep divisions within American society. *The New York Times* (11/16/93) commented on the results of a poll over Nafta:

> Support for the accord has broken down along lines of social class rather than on the traditional party divisions that typically define policy debates. College graduates, people with annual household incomes above $75,000 . . . supported the agreement. But those with a high school degree or less . . . blue-collar workers and those with union members in their households . . . opposed Nafta. (P. B12)

A picture on the same page as this article showed a worker demonstrating against Nafta with a sign reading, "Don't send my job to Mexico." The Nafta debate was so emotional because it crystallized underlying concerns about job insecurity and about the erosion of real wages of unskilled labor. Nafta became a symbol for these concerns. Critics of Nafta assume that the government can provide economic security by restricting competition.

I make the case for free trade. After Section 1, which provides some economic background to the current debate, I make the classical economic arguments for free trade. Free trade allocates resources to their most efficient use. As part of this process, it redistributes jobs to the most productive industries, without affecting the total number of available jobs. I also make the newer argument that free trade increases the growth rate of per-capita income. The world needs U.S. leadership to maintain an open trading system so that poor countries can grow their way out of poverty through integration into the world economy.

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

In addition to discussing arguments about economic efficiency, I discuss protectionism as a fiscal policy of taxes and transfers. Viewed from this perspective, protectionism is a fraud. It cannot achieve the avowed aim of its proponents to help the poor. The cost of using protectionism to preserve jobs in obsolescent industries is too high, and the income transfers more often go to the well-off than to the poor. Finally, protectionism exercises a deleterious effect on the nature of democratic government. By removing fiscal transfers from a recorded budget, it subverts the constitutional mechanisms in place that give content to the idea that sovereignty resides with the people. Protectionism encourages government dominated by special interests.

1. RESTRUCTURING OF THE U.S. ECONOMY

Political pressure for protectionism will always arise from producers desirous of limiting foreign competition. The current political pressure for protectionism, however, is more widespread. Much of the current impetus toward protectionism represents a belief that limiting foreign competition can stop a restructuring of the U.S. economy that is working to the disadvantage of the unskilled. What are the forces that are producing this restructuring and is protectionism a desirable response to them?

Three great forces are causing a profound restructuring of the U.S. economy. First, the telecommunications revolution, aided by the computer, is reducing the need for production to be organized by people in the same physical location (Jensen 1993). As a result, firms are becoming smaller and more specialized. Often, part of a production process that formerly was completely domestic is performed abroad.

Second, many less-developed countries (LDCs) and formerly communist countries are ending their isolation from the world economy. To obtain technologically sophisticated capital goods from Western countries, these countries will have to offer in trade the kinds of goods they have an advantage in producing, namely, goods whose production requires large amounts of relatively unskilled labor. As a consequence of this change in the composition of the supply of goods to the U.S. and the demand for goods from it, production in the U.S. will increasingly emphasize the high-technology goods that require an educated labor force.[1]

Finally, the technology that made possible mass production is no longer the special province of the Western world. The spread of knowledge has eliminated

[1] Indirect evidence for this statement can be found in the increasing return to education. In 1988, earnings of male college graduates exceeded those of male high school graduates by about 60 percent, up from 30 percent in 1980 (Kosters 1992). It is difficult, however, to separate the effects on the return to education of an increasingly open world economy from general technological progress.

the formerly high returns to use of this technology. As a consequence, manufacturing can no longer provide middle-class incomes for unskilled laborers. Toward the end of the 19th century, the United States became the world's preeminent industrial power because of its ability to produce huge quantities of standardized products. Especially after World War II, the United States had no rivals in manufacturing. U.S. workers profited because of the U.S.'s near monopoly on the technology of mass production and because of the escape of the U.S. capital stock from wartime destruction. The spread of technological knowledge, however, has ended the days when U.S. workers could make high wages for performing repetitive tasks. Protectionism cannot restore America's unique position in the post-World War II period. It can only retard an inevitable adjustment to fundamental economic forces.

These three forces are remaking the U.S. economy into a collection of service industries that require a highly educated labor force. Today, anything as complicated as a bicycle is made from a combination of components from numerous countries around the world. The highly skilled jobs are in organizing production rather than in making the components. Robert Reich (1991) surely had the U.S. in mind when he argued the following:

> What's traded between nations is less often finished goods than specialized research, design, fabrication, management, marketing, advertising, consulting, financial and legal services, as well as components and materials. . . . [W]hich nation's workers are responsible for the high value-added activities—such as research, design, manufacturing engineering, complex fabrication and strategy? . . . A nation whose work force is largely in [this] camp will achieve a high standard of living overall. (P. 6)

2. THE BASIC ISSUES

The core argument of Nafta critics was simple. Because U.S. workers earn more than Mexican workers, U.S. companies will move production to Mexico. The United States, Nafta critics reasoned, will lose jobs. This argument is appealing because it seems to encapsulate recent experience. In the 1950s, for example, most televisions sold in the U.S. were produced in the U.S. Now they are produced abroad in low-wage countries. If the U.S. had prohibited the importation of televisions, there would at present be more workers in the U.S. producing televisions. Left unsaid, however, is the fact that there would also be fewer workers in U.S. export industries. In addition, as the English economists David Ricardo and John Stuart Mill demonstrated almost two centuries ago, U.S. workers overall would be producing less valuable goods than they are producing now.

In order to think about the effect of free trade on jobs, it is useful to imagine two countries, East and West, initially prevented from trading with each other. What happens if they begin to trade? Can East lose jobs to West?

To be more precise, assume that each country produces the same two goods, widgets and creakles. Given the different natural resources of each country, East will be better suited for production of one good, say widgets, than the other. That good will be plentiful and will sell for a relatively low price. West will probably be in the opposite situation. It will be good at making creakles, which will be plentiful and sell for a relatively low price.

All that is required for trade to be mutually beneficial is that the goods sell for different prices in the absence of trade. With the advent of trade, both countries become better off by exporting their relatively abundant good in return for the other good. As a result, each country produces relatively more of the good in which it possesses a comparative advantage in production. East neither gains nor loses jobs, although free trade distributes some workers to more productive occupations. After all, the only reason West exports goods to East is that it wants goods produced by workers in East. (For a brief history of how economists have developed these ideas formally, see Humphrey [1988].)

Adam Smith (1937) pointed out that the wealth of a nation increases as its economy becomes large enough for individuals to specialize in production. He extended this common sense argument to free trade:

> It is the maxim of every prudent master of a family never to attempt to make at home what it will cost him more to make than to buy.... What is prudence in the conduct of every private family can scarce be folly in that of a great kingdom. If a foreign country can supply us with a commodity cheaper than we ourselves can make it, better buy it of them with some part of the produce of our own industry, employed in a way in which we have some advantage. (Pp. 424–26)

Rephrasing Smith, the U.S. should welcome cheap foreign goods and devote the resources those imports liberate to more productive uses.

3. PROTECTIONIST ARGUMENTS AND FALSE ANALOGIES

Fallacies about free trade arise because of incorrect generalization from individual experience. Consider an individual who works for a firm losing out to foreign competition. When the firm closes, the worker will have to find a new job. He will receive no income, apart from unemployment insurance, while job hunting. Because the worker will have learned skills that are particular to his old company, he will probably start a new job at a lower wage. Anthony P. Carnivale, chief economist of the American Society for Training and Development, reports that studies show that the wages of laid-off workers are lower initially, "by 10 percent on average for service workers, 20 percent for manufacturing workers, and 30 percent for automobile and steel workers" (*New York Times*, 10/3/93b, p. 28). (See also Jacobson, LaLonde, and Sullivan [1993].)

The laid-off worker is likely to generalize from his experience and conclude that protectionism would make workers better off. He is not likely to understand the consequences of protectionism for the workings of the economy, however. Workers who complain about foreign competition take for granted that they can walk into a Wal-Mart and have before them a huge variety of inexpensive goods. Many of those goods are produced abroad. Just as important, the goods produced domestically are of a higher quality and are cheaper when they face foreign competition. If the government prevents the marketplace from distributing resources to their most productive use, the Wal-Mart of today would look like the five and ten of the 1950s.

Fallacies are especially easy to propagate when they concern international rather than domestic trade. With free enterprise, groups of individuals compete to furnish goods and services for particular markets. Some groups win and others lose in this competition. A country's citizens gain collectively, however, because free entry and its concomitant free exit allocate resources to their most productive use. Free trade is an international extension of the free entry and exit that makes a market economy work domestically. With international trade, however, it is easy to spread the fallacy that one group's loss in a particular market is a loss for the country when, in fact, markets are working to distribute resources to their most productive use.

Countries' enthusiasm for exports and antipathy toward imports is an example of generalizing incorrectly. Countries frequently promote exports while discouraging imports. Exports and imports, however, are opposite sides of a single transaction. Collectively, the citizens of a country export goods and assets only because they want to import. They do not export as a matter of charity. The fallacy that a country can discourage one side of a transaction (imports) without discouraging the other side (exports) arises because particular exports are not associated with particular imports.

Protectionists use the analogy to national power and prestige to argue that there are winners and losers in international trade. It is true that military power is relative. One country becomes stronger than another country. The analogy does not hold for trade, however. Countries trade because it is mutually advantageous.

The intellectual ancestor of protectionism is mercantilism (Sowell 1978). Under mercantilism, governments intervened in the economy to prevent imports of final goods with the intention of running a trade surplus and accumulating gold. Today, protectionists argue that government should prevent imports to increase the job security of workers. The analogue to the mercantilist idea that the world possesses a fixed stock of wealth (gold), which governments should try to gain at the expense of their neighbors, is the idea that the world possesses a fixed stock of jobs, which governments should try to gain at the expense of their neighbors. This point of view is reflected in a reference to the U.S. merchandise trade deficit by an anti-Nafta critic: "If we just stopped trading with the rest of

the world, we'd be $100 billion ahead" (*Wall Street Journal*, 10/20/93, p. A9). Like mercantilists, who did not see the contradiction between their measures to accumulate gold and individual well-being, modern-day protectionists see no contradiction between their measures to limit competition and individual well-being.

Adam Smith (1937) commented incisively on the fallacy that international trade produces winners and losers:

> By such maxims as these, however, nations have been taught that their interest consisted in beggaring all their neighbors. Each nation has been made to look with an invidious eye upon the prosperity of all the nations with which it trades, and to consider their gain as its own loss. Commerce, which ought naturally to be, among nations, as among individuals, a bond of union and friendship, has become the most fertile source of discord and animosity. (P. 460)

4. INNOVATION AND FREE TRADE

Growth is integrally linked with the open competition of free markets. It is the competition among different groups wanting to bring goods to a market that furnishes the incentive to innovate and reduce costs. The competition produced by free entry yields a quest for the profits that come from being the first to market a new and attractive good or the first to reduce costs of producing an existing good. This search for high profits yields only brief success. Yesterday's winner in the competition to build the best personal computer is hardly likely to be today's winner. The search for evanescent profits, however, drives the innovation that spurs growth.

Free trade is a major source of the competition that drives innovation. This insight has been documented recently by the McKinsey Global Institute in Washington, D.C. The Institute compared productivity for the United States, Germany, and Japan in selected sectors: car assembly, motor parts, metalworking, steel, consumer electronics, food manufacturing, and brewing. For each country, the Institute found that sectors facing foreign competition were highly productive, while protected sectors were unproductive. For example, in Japan, food manufacturing and brewing are protected from foreign competition. In these sectors, output per man hour is only a third of that in the U.S. The director of the Institute summarized the results of the study as follows: "[T]he more open you are, the more productive you become" (*New York Times*, 10/22/93, p. D1).

Consider also the explanation offered in *The New York Times* (11/21/93) for why Japan lags the U.S. in the technology of wireless communication. "The Ministry of Posts and Telecommunications has ruled over the industry with a heavy hand and has been slow to authorize new services. Such tight regulation

might have helped protect Japan's market from foreign competition, but it has also stifled the innovation spurred by the more open market in the United States" (p. D1). By limiting competition, protectionism reduces incentives to increase productivity. In practice, protectionism also limits productivity growth by preserving industries that fail to remain competitive. Examples in Western countries are shipbuilding, steel, mining, and coal (Ford and Suyker 1990, p. 49).

One at times hears the comment that arguments for free trade are "academic" or "theoretical." That comment reflects a failure to understand the forces shaping international events. The most momentous event of the last part of the 20th century was the collapse of societies that attempted to isolate themselves from the world economy. Communist countries, with their ponderously inefficient command economies, were perpetually frozen into yesterday's technology. The LDCs, with their pervasive system of state controls and governmental monopolies, watched the rest of the world leave them in a time warp. The economies of these countries stagnated because the protectionism required to preserve their internal monopolies isolated them from the world economy and deprived them of the competition that spurs technological innovation and growth.

In the last several years, economists have expended considerable effort investigating the sources of economic growth. The importance of trade for growth has been documented by studies showing why some non-Western countries, but not others, grew rapidly in the last several decades. (See, for example, Moreno [1993]; Roubini and Sala-i-Martin [1991]; and Gould, Ruffin, and Woodbridge [1993].) Free trade and its counterpart, the free flow of capital, spread the knowledge that powers technological advance. Brazil, which until recently has been highly protectionist, is a negative example. For instance, for many years Brazil prohibited the import of computers or foreign software. As a result, Brazilian computers were both outmoded and more expensive than foreign computers. The inability of Brazil to make use of modern computer technology dampened innovation throughout its economy.

America, which has maintained a fairly open economy since World War II, is a positive example. At the same time that American firms are investing abroad, foreign firms are investing in the U.S. For example, the German automotive firms BMW and Daimler-Benz are now building plants in the U.S. Many of the new production techniques that are enhancing the productivity of American workers came from Japan. Toyota originated "lean production," which emphasizes just-in-time inventory control, quality control, and multitasking among workers who work and solve problems in small groups. The international organization of economic activity provides the practical way in which innovation from one part of the world is made available to another. As *Business Week* (11/8/93) wrote of a multinational corporation:

> GE is telegraphing the message that for the company to remain competitive and profitable, it has to establish deep manufacturing, technological, and financial roots elsewhere. . . . "The modern company has to spread its brains, its centers of excellence," says Fresco [GE vice-chairman]. It really is a citizen of many countries rather than a citizen of one. (P. 70)

5. U.S. WORLD LEADERSHIP

After World War II, the U.S. provided the leadership for the creation of an open world trading system. Much of the motivation came from a desire to provide a healthy economic environment in which free countries could flourish. Free trade was the economic counterpart to the Kennan-Truman doctrine of containing the expansion of Communism. The American policy of free trade deserves as much credit as containment for the collapse of Communism.

Today, free trade remains just as important. It is essential to elimination of poverty in the LDCs.[2] The specialization that free trade makes possible raises living standards, especially for small countries, which lack a large internal market. Also, if a poor country does integrate into the world economy, it can grow rapidly by drawing on the stock of technological and organizational knowledge that developed countries have acquired. Korea, for example, doubled its output per capita in an 11-year period, 1966 to 1977. Specialization, however, creates an interdependence among the countries of the world. That interdependence in turn creates the possibility of a trade war that could cause a world depression. U.S. leadership has been an important reason why the world has been able to avoid trade wars in the post-World War II period.

The United States can contribute to an increase in LDC living standards, especially in Latin America, by allowing its entrepreneurs to use their management skills to organize the labor force in these countries. It can play there the same role as Hong Kong and Taiwan are playing in Guandong province in China. If the U.S. does not exercise world leadership by promoting an open trading system, that leadership will pass to other countries. Technological leadership could go to countries like the Asian Little Dragons, such as Taiwan, Korea, Hong Kong, and Singapore. The LDCs of the world want technologically sophisticated capital goods. To get those goods, they will supply developed countries with goods whose production favors large amounts of semiskilled labor. If the U.S. closes its markets to such goods, it will also close down much of its own high-tech industry.

---

[2] Ironically, in the U.S., some of the same organizations that seek to alleviate poverty overseas also opposed Nafta. "The United Methodist Church, for instance, is opposed because it believes Nafta would throw people out of work and wreck the environment" (*Wall Street Journal*, 12/23/92, p. 1). The author's own Methodist church has supported a clinic in Matamoros, Mexico. The higher incomes of Mexican workers that would be produced by Nafta would allow them to purchase better health care.

An article in *The Washington Post* (11/7/93) explains where the jobs in U.S. high-tech industries will go if the U.S. closes its borders to imports from low-wage countries:

> The South Korean and Taiwanese economies are being transformed to more advanced industrial bases, spurred in part by a surge in exports to China. . . . The industries losing investment and jobs to China require large numbers of workers sweating over routine tasks. . . . But the explosive growth of China's economy is stoking demand for Korean and Taiwanese products that involve higher technology. (P. H1)

6. THE COST OF PROTECTIONISM

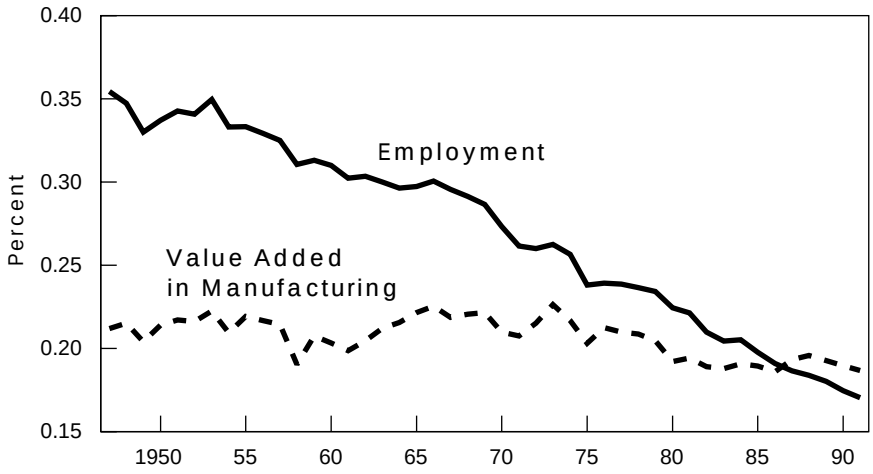
Some idea of the cost of using protectionism in an attempt to preserve jobs can be gained by observing government intervention in agriculture. In agriculture, gains in productivity outstrip gains in product demand. That is, productivity gains shift the supply schedule of agricultural goods outward faster than rising income shifts the demand schedule. Only the sustained exodus of farmers keeps the price of agricultural products from falling.

Most Western governments have intervened heavily in agricultural markets to preserve agricultural employment. What do the results from this intervention suggest for government intervention to limit job loss in manufacturing? First, no government has been able to reduce the secular decline in agricultural employment. In 1900, more than a third of the U.S. labor force was employed in agriculture. Today, only about 3 percent of the population depends upon agriculture for its livelihood. Second, government intervention is extremely expensive. For example, the Organisation for Economic Co-operation and Development puts the per-capita cost of government support for agriculture in 1992 at $360 in the U.S., at $450 in the European Community, and at $600 in Japan (*Financial Times,* 8/16/93).

Similarly, in manufacturing, the rate of growth of productivity is so high that employment in manufacturing falls over time, while the share of manufacturing in U.S. output remains steady. Figure 1 shows the shares of manufacturing employment and output in total employment and output from 1947 through 1991. Over the entire period, the share of manufacturing output has remained fairly steady at around 20 percent. The share of manufacturing employment in total payroll employment, however, has fallen steadily from 35 percent to about 17 percent. In the second quarter of 1993, manufacturing employment was somewhat less than 18 million, only moderately higher than average employment in the 1950s. Manufacturing output, however, has almost quadrupled.

The steel industry, which has been one of the most highly protected industries in the U.S., furnishes another example of ineffective but expensive government policies for protecting jobs. Employment in the steel industry fell

Figure 1 Manufacturing Employment and Output as a Share of Total Employment and Output



Notes: Employment is wage and salary workers in manufacturing divided by the total of non-agricultural wage and salary workers as measured by the Bureau of Labor Statistics payroll employment survey. Value added in manufacturing is real GDP originating in manufacturing (deflated using the manufacturing value-added price deflator) divided by real GDP (*National Income and Product Accounts of the United States, 1929–1982*, 1986; Yuskavage 1990).

by 30 percent between 1982 and 1990 even though steel production rose by 45 percent. Although import quotas and tariffs prevented any increase in imports over this period, the growth of mini-mills, which use fewer workers, increased competition, raised productivity, and reduced employment (*Washington Post*, 10/8/93). In a recent study, Gary Hufbauer and Kim Elliott found that the net cost per year to the U.S. for each job saved in 21 protected industries was $54,000. The cost per job saved ranged from a high of $115,000 per year in luggage to $4,000 in corn brooms (*New York Times*, 11/12/93).

The changing character of world trade renders modern-day protectionism especially costly. Much of the growth in international trade is in services. Computers and new communications technology make it possible to perform data processing and other kinds of back-office record keeping abroad.

> Barbados, Jamaica, the Philippines, Singapore and Ireland have emerged as the most popular "back office" locations. The jobs range from simple data entry to accounting, medical transcription, telemarketing, and technical support for high technology products. . . . In the years ahead, some analysts say, tens of thousands of clerical and technical jobs could migrate abroad.
>
> (*Wall Street Journal*, 8/14/91, p. 1)

The importation of labor services made possible by advances in telecommunications, however, cannot be prevented without isolating the U.S. from the free flow of information.

7. KEEPING GOVERNMENT TRANSPARENT

Keeping Government Accountable

U.S. constitutional democracy is based on the concept of limited government, which reduces the ability of officials to exercise power arbitrarily. A significant means of limiting the power of government is to organize economic activity through the voluntary exchange of a free enterprise economy with its separation of competition for control over resources from competition for control over political power. That separation limits the incentives to compete for power because gaining political power does not confer unlimited power to control. Talented, aggressive individuals thus have incentives to organize economic activity as well as to compete for political power. The resulting distribution of competitive individuals between the private and government sectors is part of limiting the power of government.

Because protectionism results in an organization of economic activity through government control rather than markets, it contributes to a system of incentives that promotes the social importance of political power. One result is to reduce the ability of government to function by encouraging the formation of lobbies to influence government. These lobbies become single-issue blocs. For example, congressmen in districts where economic activity is devoted significantly to textile or sugar production, which benefit from quotas on foreign imports, often require support for these quotas as a condition for joining coalitions for passage of legislation unrelated to trade. The separation of powers that characterizes U.S. government, however, creates the need for coalitions to pass legislation. Achieving political consensus then becomes harder because of the difficulty in forming coalitions out of many single-issue voting blocs. Forming the coalitions necessary to conduct the business of government requires perpetual promises of special favors. By giving government control over the distribution of income, protectionism encourages the formation of the single voter blocs that produce legislative gridlock.

Economic progress inevitably produces winning and losing producers. (Everyone gains as a consumer.) With free enterprise, the winners do not compensate the losers. When the government organizes economic activity, the necessity of governing through coalitions means that often for change to occur the winners must provide some compensation to the losers. The difficulty of arranging such compensation limits the pace of economic progress. An example is the difficulty governments in some countries are having closing their inefficient steel mills. These governments fear the political repercussions from job losses that would come with ending government subsidies. The political

difficulty of compensating the steel workers who would lose their jobs induces governments to resist economic change.

The politically corrosive effects of protectionism can be seen most clearly in countries where it has been pursued vigorously. Argentina, for example, is currently dismantling the legacy of Peronism. Peronism only differed in degree from the protectionist program of Nafta critics. It carried to the logical limit the protectionist idea that government can provide job security by limiting competition. Juan Peron promised job security to urban factory workers by protecting Argentine firms from competition. Tariffs and quotas prevented foreign competition, and cartelization and price fixing prevented internal competition. Jobs at firms threatened with bankruptcy were protected through nationalization. As of 1990, more than a third of urban workers worked for the government (*New York Times,* 5/14/90).

Wealth was not gained through entrepreneurial effort, but rather through acquiring government sanction to operate a monopoly. By making government the arbiter of the distribution of income, Argentina encouraged the organization of economic activity into large blocs powerful enough to lobby government or to threaten the government with disruptive strikes. Those who could not organize went into the underground economy. When the prices of Argentina's agricultural exports stopped rising in the post-World War II period and when industrial productivity stagnated, the only forces capable of holding Argentina together were militarism or strident nationalism.

In the absence of competition, Argentina's monopolies became notoriously inefficient. The state oil company drilled wells just to keep its employees busy (*Wall Street Journal,* 7/9/91). Customers had to wait several years to get a telephone from the state phone company (*New York Times,* 4/23/90). State-owned enterprises ran deficits, and the government financed those deficits by printing money. In 1989, inflation was close to 3,000 percent. In that year, rioters looted supermarkets.

To borrow the vocabulary of Nafta critics, there was nothing academic or theoretical about the consequences of protectionism. For the first part of the 20th century, Argentines possessed a standard of living roughly the same as the United States. Argentina purchased a short-lived job security for some workers, but at the price of poverty for many of those excluded from the government's system of worker welfare. In an article aptly entitled, "Argentines Count the Cost of Politics," the *Financial Times* (4/20/89) reported:

> Government figures estimate that 30 percent of households are now classifiable as poor, lacking sufficient income to cover basic necessities of clothing, diet, and education. In 1988, the United Nations Children's Fund estimated that 20,000 Argentine children annually died prematurely from diseases directly related to malnutrition. Some 2m live in slums around Buenos Aires in conditions familiar to countries lacking a tenth of the country's natural resources.
>
> (P. 6)

Fortunately, Argentina has now undertaken a vast program of free market reforms including privatization and drastic reduction in trade protection. By the end of 1994, it plans to be part of a tariff-free common market, known as Mercosur, which includes Brazil, Paraguay, and Uruguay.

Monitoring Government

Limited government makes it feasible for citizens to monitor the state's activities. That monitoring gives content to the premise of American constitutional democracy that sovereignty resides with the people. A key way in which the Constitution provided for the monitoring of government was the assignment of fiscal policy to Congress. Congress, in turn, with its two houses and large number of members, was designed to ensure open debate. It was no accident that fiscal policy was assigned to the "world's greatest deliberative body."

Protectionism constitutes a shadow fiscal system of taxes and subsidies. Tariffs and quotas allow Congress to impose taxes and grant subsidies that would not be feasible if they had to be openly debated. When government imposes a tariff or quota, it imposes a hidden tax. That tax is paid by consumers in the form of higher prices. Nowhere does the tax paid by consumers appear on any recorded budget.

Consider comments in *The New York Times* (10/3/93a) about the Canadian experience under its recent free trade policies:

> Old manufacturing industries have been clobbered, but new high technology industries like precision instruments, telecommunications, computer parts and specialized machinery are starting to flourish. . . . The losers—old line businesses like food processing and makers of furniture, appliances and clothing—tended to be labor intensive. The winners are high-technology companies that pay more because of higher skills that add greater value to the end product.
>
> (P. 1)

If Canada had resisted change by raising tariffs to protect its threatened "old line businesses," those businesses would have been the recipients of the expenditures of the resulting shadow fiscal system. Consumers and the individuals who would have gone into the new high-technology industries would have paid the taxes. The appeal of protectionism is that these fiscal transfers are off budget. While the recipients of the benefits are aware of the benefits they receive, those who pay the tax are usually unaware of the burden imposed on them.

8. DISTRIBUTIONAL ISSUES

Tariffs Are Regressive

Protectionism is driven by the easy identifiability of its benefits and the diffuse, hidden nature of its costs. The incentives it creates to organize

politically virtually ensure that wealth is in practice redistributed to politically influential groups and away from the politically powerless. Because the wealth transfers created by protectionism go unrecorded on the government's regular budget, open debate cannot offer protection against perverse wealth transfers. Opponents of Nafta asserted that free trade hurts the disadvantaged. While it is true that the changing U.S. comparative advantage in world trade favors those with an education, it is wrong to conclude that free trade hurts other groups. The taxes that tariffs and quotas impose are often regressive. Consider the case of textiles. The U.S. imposes quotas on more than 3,000 kinds of textile products (Bovard 1991). These quotas impose a tax in the form of higher prices. The U.S. International Trade Commission has estimated that without tariffs and quotas on textiles, the price of clothing would drop by 11.4 percent (*New York Times,* 11/29/93). According to a study by William Cline of the International Institute of Economics, that tax amounted to $260 per household in 1991 (Jones 1991). The tax is a small fraction of the income of a wealthy family, but a large fraction of the income of a poor family. Import quotas on automobiles, shoes, beef, and sugar impose the same kind of regressive tax. For example, U.S. import quotas on beef raise the price of hamburgers, a common part of the diet of lower-income Americans (Sheehan 1993). In general, quotas hurt the poor disproportionately because they cause foreign producers to alter the mix of their exports in favor of high-priced goods.

Similarly, the benefits from trade restrictions often affect the distribution of income perversely. New tariffs and quotas produce a windfall for the existing stockholders of corporations while offering no increase in wages to low-wage workers who, unlike the favored stockholders, continue to offer their labor services in a competitive market. Sugar offers an example. The government keeps the domestic price of sugar at about twice the world level through import quotas. The Commerce Department estimated that for 1988 import quotas added around $3 billion a year to the grocery bills of consumers (Gatt 1993, p. 6). *The Wall Street Journal* (6/26/90) reported:

> The General Agreement on Tariffs and Trade has decreed U.S. sugar import quotas illegal. . . . Opposing change is Big Sugar's lobby and its phalanx of political action committees, long fabled on Capitol Hill for their generosity. From 1983 through mid-1989, sugar and corn sweetener lobbyists supported their pitches to Congress with $3.3 million in campaign contributions. . . . That's a lot of money from about 10,000 beet growers in the Midwest and the West; 1,000 cane producers, dominated by a few big sugar planters and corporations. . . . But they can afford it. Two of the biggest beneficiaries of the sugar program . . . collected what the sweetener users group calls a "windfall" of $180 million in sugar benefits last year. (P. 1)

Needless to say, none of that windfall goes to the workers in the fields cutting the sugar cane.

Unskilled Workers

The deterioration in the economic well-being of less well educated workers since the early 1970s has made the issue of free trade with low-wage countries highly emotional. Economic reasoning (formalized in the Stolper-Samuelson theorem) suggests a tendency toward the equalization of wage rates across countries. The importance of this influence on wages, however, is easily exaggerated. Wage rates should tend to equalize for particular skill levels. On average, U.S. workers have considerable education and training, so very few are in direct competition with the uneducated, manual laborers of the LDCs. Also, trade with low-wage countries cannot be the major reason for the deterioration in relative wages of low-wage workers in the United States because U.S. foreign trade with low-wage countries is relatively unimportant. As Krugman and Lawrence (1993) point out, the average U.S. trading partner in 1990 had a manufacturing wage rate 88 percent of the U.S. level. Imports from countries with wage rates less than half the U.S. level amounted to only 2.8 percent of GDP, a fraction unchanged since 1960.

Nevertheless, changes in the world economy will make the U.S. labor market more inhospitable in the future to unskilled workers. The integration into the world economy of the formerly Communist countries and the LDCs in Latin America and Asia will add to the world labor market a huge number of unskilled workers. China and India each have populations near one billion. The increased competition from those workers will reinforce the erosion in the real wages of unskilled and blue-collar workers in the U.S.

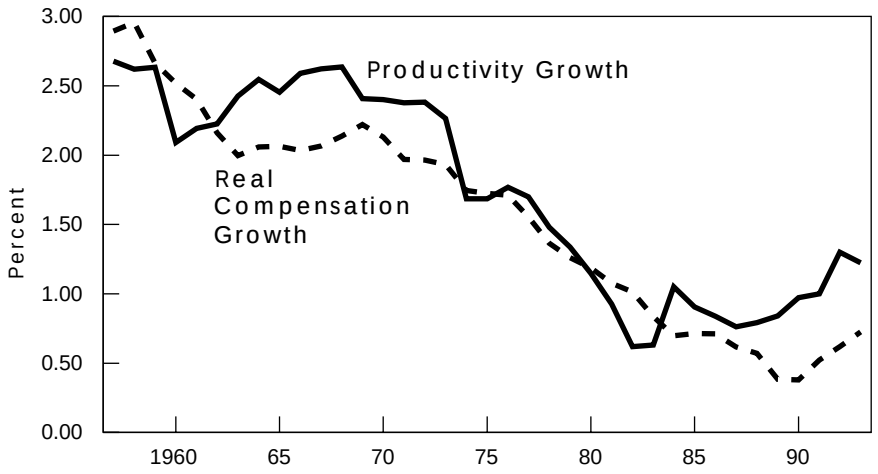
What can the U.S. do to help its disadvantaged workers? In considering the effects of changes in world comparative advantage on the distribution of income, one should keep in mind two characteristics of income distribution—the *inequality* among income groups and the *mobility* among income groups. The integration of the world economy may increase income inequality, but it can also offer increased mobility by increasing returns to investing in education. The income ladder in the U.S. may start with a low rung, but access to education makes the higher rungs widely accessible. Such education includes on-the-job training. In a study of job-related education, Alan Eck (1993) found that high school graduates who had taken jobs requiring both qualifying training and subsequent on-the-job training earned slightly more than college graduates with neither kind of training.

It is important to avoid policies that reduce wage inequality by limiting income mobility. Most European countries, for example, established high minimum wages in an effort to prevent the erosion in wages at the bottom of the pay skill that occurred in the U.S. in the 1980s. One consequence was to price many workers out of the market and to raise the unemployment rate. Moreover, those who become unemployed remain unemployed for long periods. Almost half of Europe's unemployed workers have been unemployed for more than a

year (*The Economist*, 10/9/93). Those workers lose some of the job skills they already possess, thereby limiting the possibility of a good job in the future. In November 1993, only 22 percent of unemployed workers in the U.S. had been unemployed for more than half a year. The mean duration of unemployment was 19.3 weeks and the median was 8.7 weeks (Bureau of Labor Statistics and Employment and Training Administration).

Finally, it is important to avoid all kinds of government interference in markets such as protectionism that reduce productivity. Productivity growth is the engine that pulls up all real wages, low and high, over time. Figure 2 (suggested by Prudential Insurance *Economic Review*, October 1993) plots the growth rate of labor productivity and of real worker compensation per hour. The series are plotted as ten-year moving averages to eliminate cyclical variation. As shown in Figure 2, productivity growth is the key determinant of real wage growth.

Figure 2 Growth Rates of Labor Productivity and Real Compensation per Hour



Notes: Observations are ten-year moving averages of annual growth rates. Productivity is labor output per hour. Real compensation is compensation per hour converted to constant dollars using the GDP deflator. Both series are published by the Bureau of Labor Statistics.

9. HELPING THE DISADVANTAGED

Protectionism imposes a tax on consumers and on the workers who otherwise would have worked in an expanded export sector. As a tax, protectionism is grossly inefficient in transferring income. Its income transfers often

hurt the economically disadvantaged. The economically disadvantaged can be helped with greater assurance of success through the fiscal system consisting of on-budget congressional appropriations and taxes rather than through protectionism. There are ample opportunities to use the existing fiscal system to redistribute income toward the disadvantaged in legitimate on-budget ways. A redesign of the current fiscal system to help the disadvantaged would be much more likely to achieve a desirable distribution of income than the capricious intervention in particular markets recommended by protectionists.

One step the U.S. government could take to soften the economic hardships of the less fortunate would be to tilt the income transfers it controls in their favor. Between 1965 and 1980, the government increased dramatically its control over the distribution of income. In 1965, transfer payments to individuals made by local, state, and federal government were 5.5 percent of GDP. By 1993, this figure had risen to 14 percent.[3] Also, tax expenditures are a significant factor in the U.S. fiscal system. (Tax expenditures measure the revenue loss due to tax breaks for special groups.) As a percent of gross national product, they are about 7 percent (Table 3-16 in Peterson [1991], p. 90).

It is not hard to imagine ways to redesign the current fiscal system to lessen the inequality of income and to help those who are hurting because of an increasingly competitive marketplace. The immediate response to specific suggestions, however, is likely to be that they are politically painful. But does not that response explain much of the political appeal of protectionism? Protectionism, by allowing its proponents to argue that they are addressing the problems of the disadvantaged, makes it possible to avoid discussion of genuine, but politically difficult, responses to the problems of the disadvantaged.

World economic integration and technological innovation will all make the labor market increasingly inhospitable for the unskilled and uneducated. Government cannot protect this group through protectionism and other kinds of direct intervention that reduce economic efficiency. Government could, however, alter the taxes and transfers of the modern welfare state in ways that promote the economic well-being of the least fortunate.

10. CONCLUDING COMMMENTS

The costs imposed by tariffs, quotas, and other forms of trade discrimination appear on no budget. For this reason, the public loses an important protection against wealth transfers from the less fortunate to the politically well organized. At least in the case of protectionism, direct government intervention in markets

[3] Figures on transfers are from “Transfer Payments to Persons” in tables of Federal Government Receipts and Expenditures and State and Local Government Receipts and Expenditures in *Economic Report of the President.*

to redistribute income in practice has often redistributed income perversely. The appropriate way to help disadvantaged workers is to make certain that the overall effect of the fiscal system is to redistribute income to low-income individuals.

The 20th century began as an optimistic era of free trade, free movement of capital and peoples, and the free flow of ideas across national boundaries. The internationalism of that era ended with World War I. The totalitarianism and nationalism of the ensuing period and the murderous wars they spawned came close to extinguishing the human freedom valued by Western civilization. Fortunately, after World War II, the U.S. became a leader in recreating a liberal world order characterized by the free international movement of goods and ideas. Free trade means open borders and the free flow of ideas across national boundaries. The free flow of ideas is the essential condition for the creation of a democratic and prosperous world. U.S. leadership will determine the kind of world the 21st century will be. The weather vane of that leadership is the commitment to free trade.

REFERENCES

Bovard, James. "High Cost of Textile Protection," *The Journal of Commerce,* December 10, 1991, p. 12A.

Business Week. "GE's Brave New World," November 8, 1993, pp. 64–70.

Eck, Alan. "Job-Related Education and Training: Their Impact on Earnings," *Monthly Labor Review,* vol. 116 (October 1993), pp. 21–48.

The Economist. "Doleful," October 9, 1993, p. 17.

Financial Times. "Counting the Cost of Trade Impasse," August 16, 1993, p. 3.

_____. "Argentines Count the Cost of Politics," April 20, 1989, p. 6.

Ford, Robert, and Wim Suyker. "Industrial Subsidies in the OECD Countries," *OECD Economic Studies,* Autumn 1990, pp. 37–81.

Gatt Secretariat. "Trade, the Uruguay Round and the Consumer." Information and Media Relations Division of Gatt, August 11, 1993.

Gould, David M., Roy J. Ruffin, and Graeme L. Woodbridge. "The Theory and Practice of Free Trade," Federal Reserve Bank of Dallas *Economic Review,* Fourth Quarter 1993, pp. 1–16.

Humphrey, Thomas M. "The Trade Theorist's Sacred Diagram: Its Origin and Early Development," Federal Reserve Bank of Richmond *Economic Review,* vol. 74 (January/February 1988), pp. 3–15.

Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan. "Long-Term Earnings Losses of High-Seniority Displaced Workers," *Economic Perspective*, November/December 1993, pp. 2–20.

Jensen, Michael C. "The Modern Industrial Revolution, Exit, and the Failure of Internal Control Systems," *The Journal of Finance*, vol. 48 (July 1993), pp. 831–80.

Jones, Kent. "Real Victims of Textile Quotas," *The Journal of Commerce*, June 26, 1991, p. 12A.

Kosters, Marvin H. "The Rise in Income Inequality," *The American Enterprise*, November/December 1992, pp. 29–37.

Krugman, Paul, and Robert Lawrence. "Trade, Jobs, and Wages," Working Paper 4478. Cambridge, Mass.: National Bureau of Economic Research, September 1993.

Moreno, Ramon. "Are World Incomes Converging?" Federal Reserve Bank of San Francisco *Weekly Letter*, November 26, 1993.

The New York Times. "Government Says Federal Tariffs Cost Shoppers $19 Billion Yearly," November 29, 1993, p. A14.

_____. "Now It's Japan's Turn to Play Catch-Up," November 21, 1993, p. D1.

_____. "Americans Are Closely Split on Trade Pact, Poll Shows," November 16, 1993, p. B12.

_____. "The High Costs of Protectionism," November 12, 1993, p. D1.

_____. "Why U.S. Is Indeed Productive," October 22, 1993, p. D1.

_____. "Canada's U.S. Trade Experience Fuels Opposition to the New Pact," October 3, 1993a, p. 1.

_____. "Policy on Jobs Holds Trade Pact's Fate," October 3, 1993b, p. 28.

_____. "Argentina's Painful Path to Efficiency," May 14, 1990, p. D1.

_____. "Argentina Tries to Sell Its Shaky Phone System," April 23, 1990, p. D8.

Peterson, Wallace C. *Transfer Spending, Taxes, and the American Welfare State*. Boston: Kluwer Academic Publishers, 1991.

Reich, Robert. "The Myth of 'Made in the U.S.A.,' " *The Wall Street Journal*, July 5, 1991, p. 6.

Roubini, Nouriel, and Xavier Sala-i-Martin. "Financial Development, the Trade Regime, and Economic Growth," Working Paper 3876. Cambridge, Mass.: National Bureau of Economic Research, October 1991.

Sheehan, James M. "Sacred Cows of Protectionism," *The Journal of Commerce*, November 19, 1993, p. 6A.

Smith, Adam. *The Wealth of Nations,* Edwin Cannan, ed. New York: The Modern Library, 1937.

Sowell, Thomas. "Adam Smith in Theory and Practice," in Fred. R. Glahe, ed., *Adam Smith and the Wealth of Nations: Bicentennial Essays.* Boulder, Colo.: Colorado Associated University Press, 1978, pp. 149–72.

The Wall Street Journal. "In Debate over Nafta, Many See Global Trade as Symbol of Hardship," October 20, 1993, pp. 1 and A9.

_____. "Free-Trade Pact Spurs a Diverse Coalition of Grass-Root Foes," December 23, 1992, p. 1.

_____. "American Firms Send Office Work Abroad to Use Cheaper Labor," August 14, 1991, p. 1.

_____. "South Americans Push Sales of State Assets in Swing to Capitalism," July 9, 1991, p. 1.

_____. "Small Minnesota Town Is Divided by Rancor over Sugar Policies," June 26, 1990, p. 1.

The Washington Post. "Asia's 'Dragons' Accept Trade's Pains and Gains," November 7, 1993, p. H1.

_____. "NAFTA and Trade Are Tiny Parts of a Job Revolution," October 8, 1993, p. G1.

Yuskavage, Robert E. "Gross Product by Industry, 1988–91," *Survey of Current Business,* November 1993.

Delivering Deposit Services: ATMs Versus Branches

David B. Humphrey

Over the past 20 years (1973–1992), the total number of banking offices has grown from 40,600 to 63,900, an expansion of 57 percent. This exceeded the 21 percent growth in the adult (age 18 and older) population. The number of automated teller machines (ATMs) has grown even more rapidly, from fewer than 2,000 to more than 90,000 over the same period. As a total, there was one banking office or ATM for 3,700 people in 1973. In 1992, there were three banking offices or ATMs for the same number of people. This increase effectively tripled the accessibility and convenience of bank-provided deposit services. In addition, ATMs are typically "open" 24 hours a day, providing even more convenience than a traditional banking office.

Ever since ATMs were first introduced in 1971, they have been touted as a potentially lower-cost alternative to the traditional branch banking office. The presumption of cost savings from expanded ATM use has in the past focused on scale economies. Substantial scale economies were indeed estimated for ATMs using special FDIC survey data for 1975 (Walker 1978, 1980). This early analysis is augmented here with a new estimate of ATM scale economies using survey data for 1984. The two scale estimates are similar but suggest that ATM technology has improved over time, leading to greater scale economies.

While ATM scale economies appear to be substantial, they may not translate into reductions in bank costs or increases in bank profits. This can occur if, for the same set of "free" or below-cost deposit services, consumers use ATMs more intensively than they had previously used a traditional banking office. Similarly, the scale economy benefits of ATMs can be dissipated if ATMs are

The author, the F. W. Smith Eminent Scholar in Banking and Professor of Finance at Florida State University, gratefully acknowledges comments by Allen Berger, Bill Cullison, Mike Dotsey, Tony Kuprianov, Larry Pulley, and John Walter. Larry Pulley at the College of William and Mary provided the scope estimates using the composite functional form, and Caroline Kreimer, Chris Otrok, and Floyd Tyler provided research assistance.

"oversupplied" to consumers primarily to enhance or maintain deposit market shares. Thus the existence of ATM scale economies may or may not lead to lower bank costs or increases in profits.

The primary purpose of this article is to determine the impact of an increase in ATM use on bank costs and profits. This is obtained by estimating separate multi-output banking cost and profit functions using cross-section data for 161 banks during 1991 and 1992. In brief, there appears to be no significant reduction in costs when ATMs are substituted for banking offices in the delivery of deposit services. On balance, while consumers have clearly benefited from the increased availability and convenience of an expansion of banking offices and ATMs over the last 20 years, banks today realize no net cost savings from these developments. Indeed, deposit delivery costs are higher, not lower. However, because of revenue effects, net income (profit) is marginally higher and represents a small net benefit to banks.

1. ATM USE, SCALE ECONOMIES, AND TRANSACTION COST

The Structure of U.S. Payments

Table 1 shows the percentage volume and values of the various methods of making payments in the U.S. economy. As in most countries, cash is the most frequently used payment instrument. Cash is estimated to account for 83 percent of all U.S. payment transactions.[1] The next most important instrument in terms of transaction volume is the check at 14 percent. Thus cash and checks account for over 97 percent of transaction *volume*. All other payment instruments—credit cards, automated clearing house (ACH) "electronic checks," traveler's checks, money orders, point of sale (POS) debit cards, and wire transfers—account for less than 3 percent of total transactions. The ordering for transaction *value* is a different story. Wire transfers, which average $3.3 million per transaction, account for 82 percent of total payment value, while checks comprise 16 percent. Thus over 98 percent of payment values are shouldered by wire transfers and checks. The value of cash transactions is less than one-half of 1 percent of the total.

While surveys show that cash is the most frequently used payment method, the overall value of cash transactions is small because cash is used primarily for small-value transactions. ATMs fit into the U.S. payment structure in two

[1] Cash has been estimated to account for 86 percent of all payment transactions in Germany, 78 percent in the Netherlands, and 90 percent in the United Kingdom (Boeschoten 1992, pp. 73–74). The procedures used to estimate U.S. payment volumes and values are quite complex and are contained in Humphrey and Berger (1990), Table 2-A1.

Table 1 The Structure of the U.S. Payment System

Type of Payment Instrument	Volume Composition (percent)	Value Composition (percent)	Average Value (dollars)
Nonelectronic			
Cash	83.4	0.4	5
Check	14.1	16.3	1,188
Electronic			
Credit Card	2.1	0.1	62
ACH	0.3	1.1	3,882
Wire Transfer	0.1	82.1	3,300,000

Source: Humphrey and Berger (1990), Table 2-A1.

ways. First, ATMs are an increasingly important source of cash to deposit holders for cash transactions. Second, the greater convenience of ATMs has lowered the transactions cost of using cash as a means of payment (Boeschoten 1992; Daniels and Murphy 1993).

Prior to the 1940s, most cash was obtained at the workplace; employees were commonly paid in cash, usually on a weekly basis. After employers converted to payroll checks, the main sources of cash acquisition shifted to cashing one's entire paycheck, writing checks for cash at one's bank, or writing a check at the supermarket or other retail establishment for a value larger than the purchase amount. Now, with easy access to ATMs, cash is substituting for checks written solely to obtain cash—previously 8 percent of all checks (Bank Administration Institute 1979).

What Do ATMs Do?

ATMs provide many of the most demanded deposit services. In order of importance, as shown in Table 2, these services include cash withdrawals, cash or check deposits, transfers among deposit accounts, and bill payments.[2] Surveys suggest that cash withdrawal accounted for 77 percent or more of all ATM transactions in 1991, 1984, and 1975. Since only 1 percent of ATM transactions represent bill payments, it would be incorrect to conclude (as some have) that ATMs represent a move to electronic payments. In fact, ATMs have promoted an increased use of cash to the detriment of checks and potential electronic

[2] Since separate balance inquiry transactions are commonly made prior to withdrawing cash to see if the balance is sufficient for the withdrawal, these transactions have not been included in the breakdown in Table 2.

Table 2 Use of ATM Services

	1975	1984	1991
Cash Withdrawal	77%	77%	86%
Cash or Check Deposit	20	19	10
Account Transfer	2	3	3
Bill Payment	1	1	1

Source: Walker (1978); van der Velde (1985); and Board of Governors of the Federal Reserve System (1991).

payments such as point of sale (POS) debit cards. Furthermore, ATMs are also a partial substitute for nationwide bank branching because they enable depositors to obtain cash from their deposit account while traveling out of state.

ATM Scale Economies

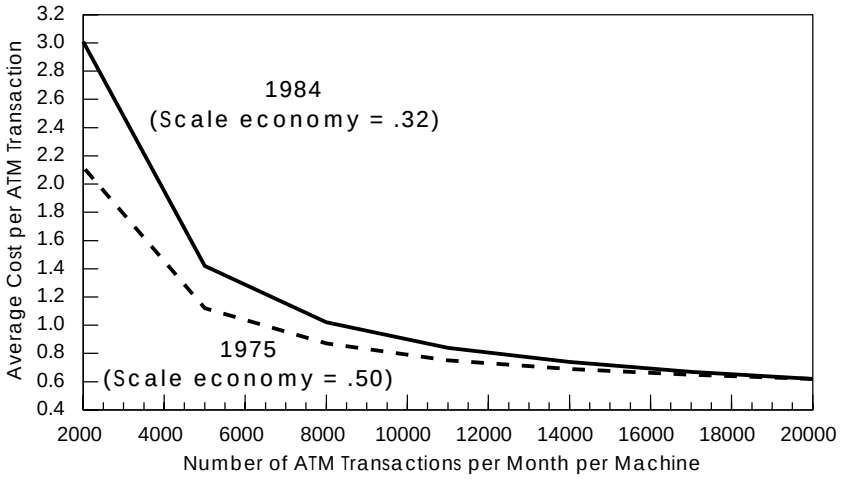
Because of scale economies, the early promise of ATMs was that the cost of an ATM transaction at mature volumes would be considerably below the cost of the same transaction at a standard branch office. Early analysis by Walker (1978) found large scale economies associated with increases in ATM transaction volume. Scale economies for 1975 are illustrated by the dashed line in Figure 1. In this year, the average total cost per ATM transaction rises only by 5 percent for each 10 percent increase in total monthly transactions volume, giving a scale economy measure of .50.[3] The solid line in Figure 1 shows ATM scale economies for 1984. There is some improvement because the cost per transaction is estimated to rise by only 3.2 percent for each 10 percent increase in transaction volume, giving a scale measure of .32.[4]

At a monthly transaction volume close to 5,000, Walker found that the cost per ATM transaction was substantially less than that of a transaction in a traditional banking office. By 1992, average transaction volume per ATM per month was over 6,000 (Barthel 1993a). Therefore, if Walker's analysis was correct, scale economies are being realized and ATM costs per transaction should be less than at a traditional banking office. A later detailed study by

[3] The scale economy value of .50 was derived from a simple log-linear equation relating ATM total costs to ATM transactions (Walker 1978).

[4] The scale economy value of .32 is derived from an in-depth cost analysis (van der Velde 1985, Figures 2 and 4) that gave $.36 as the mean average variable cost per ATM transaction (assumed to remain constant) and $1.22 as the mean average fixed cost per ATM transaction (which will fall as volume is raised above the mean). These values were determined at a monthly mean per machine transaction volume of 4,343. From these data, the implied total cost associated with different ATM transaction volumes was constructed and used in ln (total cost) $= a + b$ ln (transaction volume); estimation gave $b = .32$—the constant scale elasticity.

Figure 1 Relationship Between the Average Cost and Volume of ATM Transactions



Note: Computed from Walker (1978) and van der Velde (1985); see footnote 4 in the text.

Berger (1985) supported this conclusion and found that the fully allocated cost of a cash withdrawal transaction using an ATM was about one-half the cost of the same transaction using a human teller in a bank branch office. These studies therefore support the early historical presumption for cost savings by substituting ATMs for banking offices.

Lower ATM Cost per Transaction Offset by Higher Usage

While it is thus clear that the average transaction cost of an ATM is considerably below the cost of using a standard banking office, this lower unit cost has not translated into much overall cost savings for banks. The problem has been that the greater convenience of ATMs has led users to withdraw less cash per transaction from ATMs than they did from a branch office. This response is consistent with the inventory theory of demand for idle cash balances (Baumol 1952). The greater convenience of ATMs reduced the cash acquisition transaction cost for depositors, leading to a greater frequency of these transactions and a corresponding reduction in the average amount of idle cash balances held by the public.[5]

[5] Reductions in average idle cash balances may occur even if there is increased use of cash in payment transactions, as noted above. Although the reduction in idle cash balances likely has affected the monetary aggregates, this influence was in all probability smaller than two other important events—the rise in cash management and money market mutual funds—that occurred at the same time.

Although an ATM transaction costs as little as one-half as much as a teller transaction at a branch, ATMs are being used up to twice as often as was a teller. As a result, the cost savings per ATM transaction expected by banks has been largely offset by the unexpected increase in use (Berger 1985).

Until recently, about the only way most banks have obtained revenues on their ATM investment has been through fees charged when one bank's ATM is used by a customer of another bank.[6] When a customer uses another bank's ATM—a "foreign ATM"—for cash withdrawal, an interchange fee of about $1.00 is commonly assessed. In contrast, a cash withdrawal from an ATM owned by one's own bank is usually, but not always, free.[7] Although the foreign ATM fee may seem relatively small, it generates the majority of revenues associated with ATM use. As ATMs have expanded, the number of foreign (cash withdrawal, etc.) transactions has risen from 15 percent of all transactions in the mid-1980s, to 40 percent in 1989, to around 50 percent today (McAndrews 1991).

2. GROWTH IN ATMS OVER TIME

Availability of ATMs: 1973–1992

An estimate of the total number of ATMs in the United States is shown by the solid line in Figure 2.[8] When first introduced in 1971, ATMs expanded at an increasing rate until 1984–85, at which point the yearly expansion fell off markedly as the market became increasingly saturated. This pattern of growth—increasing at an increasing rate, reaching an inflection point, and then growing at a decreasing rate—is standard for new innovations.

Have ATMs Replaced Bank Branches?

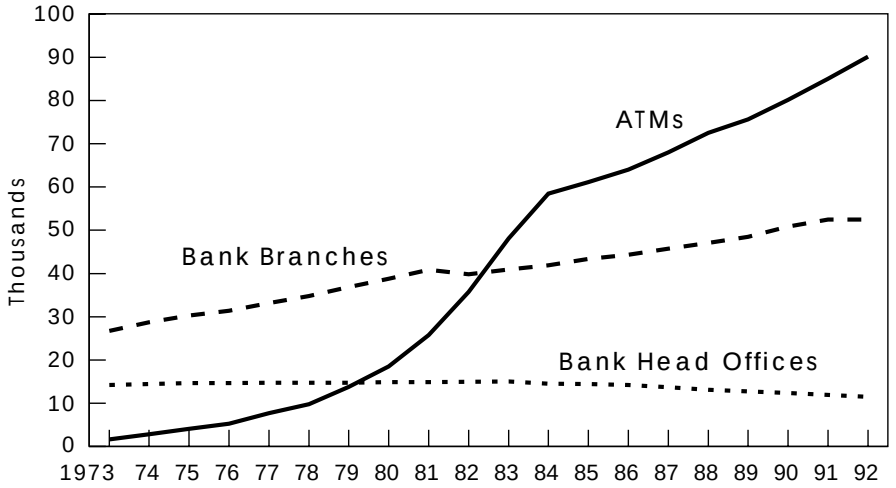
Since ATMs represent an alternative delivery method for deposit services, their rapid expansion suggests that they may have substituted for the traditional banking office in providing deposit services to the public. The growth in banking offices is also shown in Figure 2 and is divided between head offices (dotted

[6] In some cases, banks have provided ATMs not because of a strong expectation of reducing costs but rather as a defensive measure to preserve deposit market share as competitors introduced this new service for their customers.

[7] Only about one-fourth of banks charge their own customers for using the bank's own ATMs. This fee was about $.40 per transaction in 1992 (Barthel 1993a). The $1 fee for use of a foreign ATM is cost-effective, compared to a traveler's check, if more than $100 is withdrawn. Traveler's checks typically carry a fee of 1 percent of the dollar value obtained.

[8] Eugene Snyder of the Division of Federal Reserve Bank Operations at the Federal Reserve Board in Washington, D.C., provided these estimates based on industry benchmark figures and interpolation for years with missing values. These estimates are very similar to those of Laderman (1990) who obtained her primary estimates from the same source—*Bank Network News*.

Figure 2 Number of ATMs, Bank Branches, and Head Offices

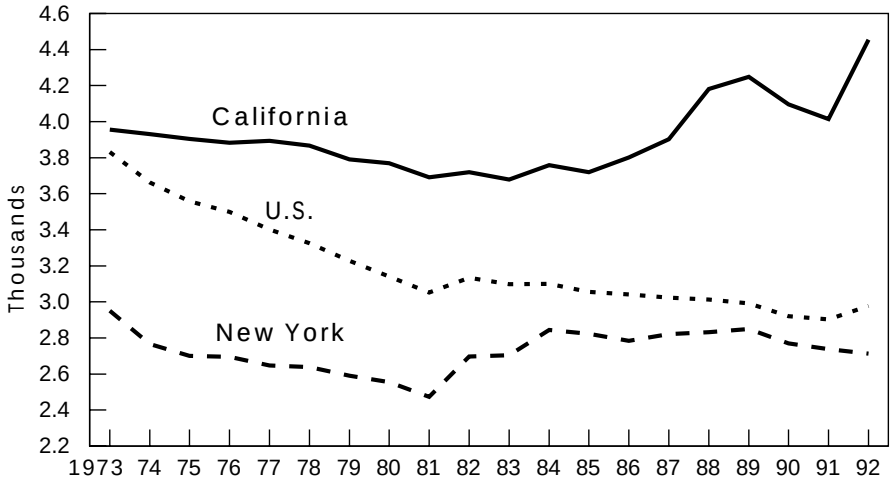


Source: FDIC, *Statistics in Banking*, various issues, and Eugene Snyder, Federal Reserve Board, Washington, D.C. (estimates of ATMs).

line) and branches (dashed line). The main point of this comparison is the more rapid growth of ATMs. While there were fewer than 2,000 ATMs and 26,700 branches in 1973, there were 90,000 ATMs and 52,400 branches in 1992. Over this same time period, the number of head offices—which equals the number of banks—fell slightly from 14,200 to 11,500. It is estimated that in 1993 around 40 percent of depositor transactions at financial institutions will be performed by ATMs rather than by tellers at branch offices (Barthel 1993b). In addition, consultant analysis suggests that by the end of the decade the number of branch offices could fall by 20 percent as bank customers are increasingly directed toward self-service activities (Tracey 1993).

One crude measure of banking office convenience would be the population served per banking office (specifically, the number of individuals 18 years and older per branch plus head office). This relationship is shown for the entire country by the dotted line in Figure 3. There is a downward trend in the number of individuals per banking office, falling from 3,800 per office in 1973 to 3,000 per office in 1992. Thus banking offices expanded more rapidly than the population being served. If ATMs replaced banking offices, we might have expected that the population/office ratio would have risen, not fallen as the aggregate data in Figure 3 indicates.

Figure 3 Population Served per Banking Office



Source: FDIC, *Statistics in Banking*, various issues, and U.S. population data on individuals age 18 and older.

The aggregate U.S. data, however, is biased by the fact that over the 1973–1992 period, 13 states removed restrictions on intra-state branching (Amel 1993). By 1992 all states allowed limited or statewide branching. The removal of branching or "unit banking" restrictions in various states has in the past led to increases in the number of banking offices in these states (Savage and Humphrey 1979). Thus the aggregate population/office ratio would fall for this reason alone.

Two large states, California (solid line) and New York (dashed line), however, had no restrictions on intra-state branching during the 1973–1992 period. These two states account for 28 percent of total domestic deposits and are the home states of the largest banks in the United States. In both states the population/office ratio first fell and then rose over 1973–1992. This result is consistent with ATMs substituting for offices after the early 1980s when the growth in offices did not keep pace with the growth in population in these two states. Anecdotal information also suggests that the increased focus on reducing bank operating costs after the early 1980s, along with the opportunity given management through mergers of banks in overlapping market areas to close underutilized branch offices and rely instead on ATMs, facilitated a substitution of ATMs for banking offices and personnel (Barthel 1992).

3. ARE ATMS ASSOCIATED WITH HIGHER OFFICE PRODUCTIVITY, LOWER AVERAGE COST, OR HIGHER AVERAGE PROFITS?

The Core Deposit/Office Ratio and ATM Use

A simple but approximate measure of the "productivity" of a bank's branch office network commonly used in the banking industry is the core deposit/office ratio. The value of core deposits—demand, savings and small-denomination time deposits—represents an important banking "output," while the number of banking offices reflects an important banking "input." Indeed, the production of deposit services accounted for 49 percent of all bank value added during the 1980s, as measured by the allocated costs for physical capital, labor, materials, and other noninterest expenses, while loans accounted for only 28 percent.[9] The question addressed here is how this simple "productivity" measure—output per unit of input—varies with increases in ATM use.

Figure 4 shows a plot and the fitted regression line of the relationship between the log of the core deposit/office ratio and the intensity of ATM use, as reflected in the log of the ATM/office ratio.[10] For all of the 161 banks sampled both in 1991 and 1992, there is a positive (and statistically significant) relationship; that is, the simple productivity measure rises as the intensity of ATM use increases.[11]

The positive relationship shown is consistent with the contention that increases in ATM use allow the number of branches to decline while supporting the same level of deposit services. Based on the regression results, a 300 percent increase in the intensity of ATM use—moving from one ATM for every two banking offices to two ATMs per banking office—is associated with a 120 percent increase in deposits, from $20 million to $44 million per average office.[12]

---

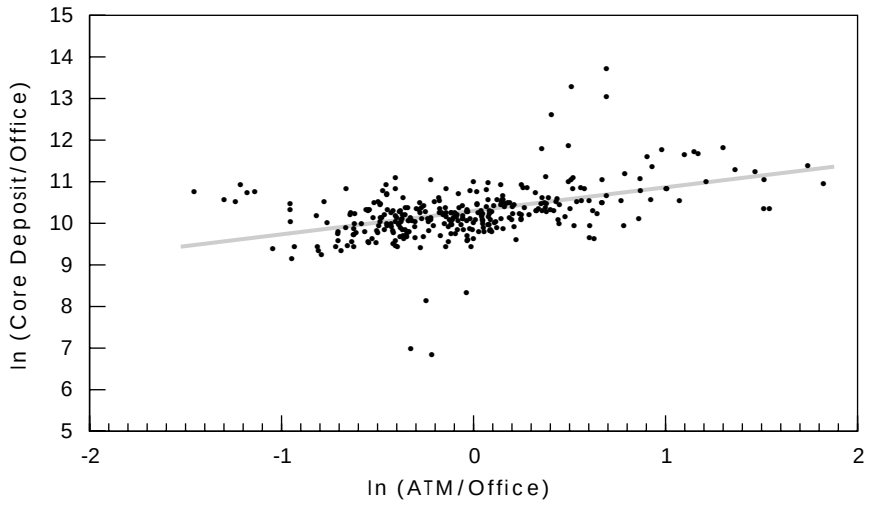
[9] These cost allocations are from the Federal Reserve's survey of financial institutions reported annually in *Functional Cost Analysis* and refer to banks with $200 million to $1 billion in deposits.

[10] The relationship shown and fitted is the following: ln (core deposit/office) $= 10.30 + .56$ ln (ATM/office). Both estimated parameters were significantly different from zero at the .05 level; the adjusted $R^2 = .20$. The double log specification was used to reduce the possible effects of heteroscedasticity as the variance of the dependent variable appeared to become larger for greater values of the independent variable. A quadratic specification gave similar results.

[11] Strictly speaking, we would expect banking output to rise if we increase inputs, such as increasing the use of ATMs. Thus our focus is on how much this single factor productivity measure rises, rather than if it rises at all.

[12] Referring to footnote 10, when ATM/office $= .5$, the predicted core deposit/office ratio is $\exp[10.30 + .56(\ln .5)] =$ $20 million. When ATM/office $= 2$, the predicted core deposit/office ratio is $\exp[10.30 + .56(\ln 2)] =$ $44 million.

Figure 4 Relationship Between Core Deposit/Office and ATM/Office Ratios, 1991–1992



Source: See the appendix.

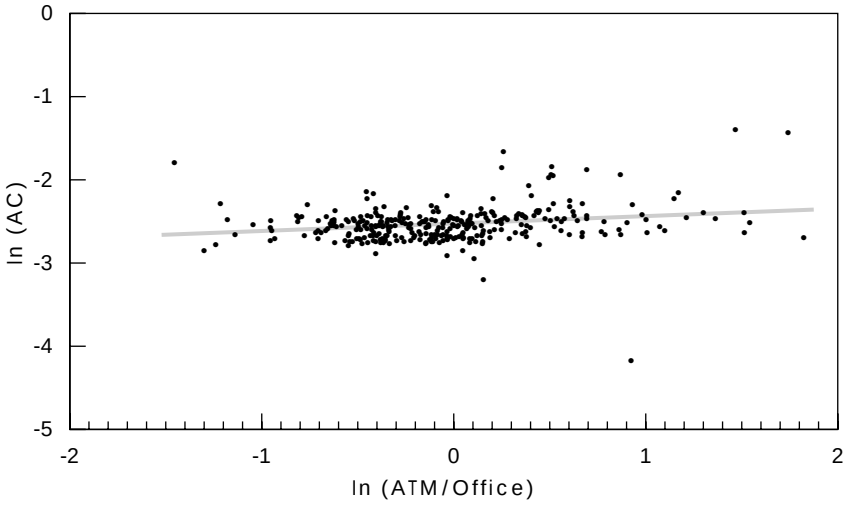
Average Cost and ATM Use

Since deposits are generally a cheaper source of loanable funds than purchased money, the result that the core deposit/branch ratio seems to rise with more intensive use of ATMs may also translate into a lower average cost of banking activity as the intensity of ATM use increases. The average cost (AC) of banking activity is measured here as the total operating plus interest cost per dollar of assets, or the total cost/total asset ratio.

Figure 5 shows how the measure of average cost varies with the ATM/office ratio for the same set of banks. The fitted relationship is slightly positive, suggesting that greater intensity of ATM use may be associated with a higher total cost/total asset ratio. The estimated relationship is exceedingly weak, however, since ATMs are only a small component of total cost. Although we do not rely on these estimates, due to a very low R^2, they weakly suggest that average cost may rise by 13 percent with a 300 percent increase in ATM intensity—from 7.56 cents per dollar of assets with one ATM for every two offices to 8.56 cents with two ATMs per office.[13]

[13] The estimated relationship is $\ln(AC) = -2.52 + .09\ \ln(ATM/office)$ and both parameters are significant at the .05 level. However, the adjusted R^2 is only .05. The predicted values of AC associated with ATM use are derived from $\exp[-2.56 + .09 \ln(ATM/office)]$, where ATM/office ranges from .5 to 2 (as in the previous footnote). A quadratic specification yields similar results. When operating cost per dollar of assets was used as the dependent variable, average operating cost also rose with the increase in the ATM/office ratio (not shown).

Figure 5 Relationship Between Average Cost (AC) and ATM/Office Ratios, 1991–1992



Source: See the appendix.

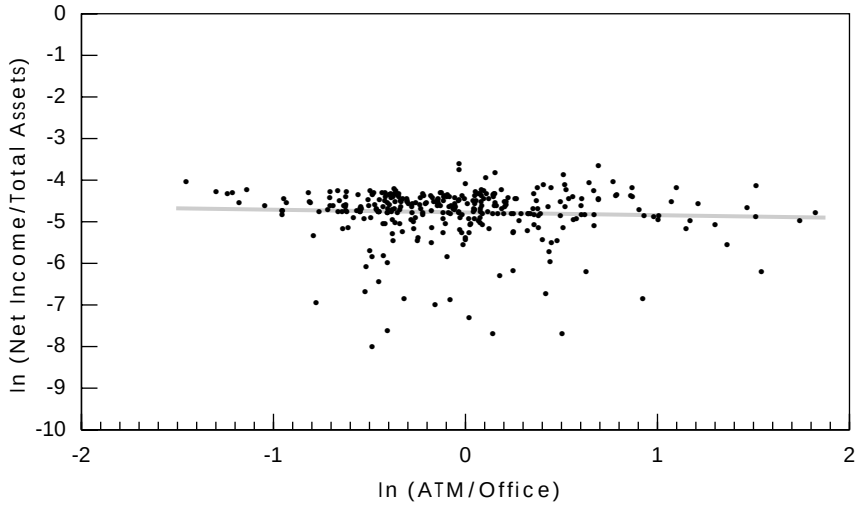
Average Profits and ATM Use

Although average costs do not appear to fall as ATMs are more intensively used, it is possible that bank revenues may be higher when more ATMs are provided. First, revenues are generated directly when a foreign ATM is used. In addition, about one-fourth of banks charge their own customers for using the bank's own ATMs. Second, the expanded convenience of ATMs may enable a bank to retain a more profitable customer base than would otherwise be possible. This may raise revenues from non-deposit services and/or permit a bank to pay a lower deposit interest rate or assess a higher monthly minimum balance on deposit accounts. All of these influences, if they are significant, could lead to higher bank profits.

Figure 6 shows a plot and the fitted relationship between the log of the ratio of net income (a common measure of bank profits) to total assets and the log of the ATM/office ratio.[14] As shown, it appears that bank profits—here measured by the return on assets (ROA)—fall slightly as the intensity of ATM use rises. However, this reduction in ROA is not significant (and the $R^2 = .00$).

[14] The relationship is the following: ln(net income/total assets) $= -4.78 - .06$ ln(ATM/office). Only the intercept was significantly different from zero at the .05 level. The same results were obtained from a quadratic specification.

Figure 6 Relationship Between Return on Assets and ATM/Office Ratios, 1991–1992



Source: See the appendix.

Therefore, ROA is apparently not affected by a 300 percent increase in ATM intensity—from one ATM for every two offices to two ATMs per office.

From this simple analysis, it appears that while there is an improvement in the core deposit/office ratio with increases in ATMs, average costs do not appear to fall. In addition, profits—as measured by ROA—are neither reduced or increased as ATMs are substituted for traditional branch offices in the delivery of deposit services. However, the analysis presented does not control for the many other factors that are known to influence bank costs and profits. To address this issue, and also to provide a more direct measure of the effects of substituting ATMs for banking offices, one needs a more comprehensive analysis.

4. COST EFFECTS OF SUBSTITUTING ATMS FOR BRANCHES

A Cost Function Model

Our approach is to specify separate multi-output cost and profit functions where the quantities of ATMs and banking offices enter directly as substitute deposit delivery methods. In such a model, the variation in total cost or total profit associated with different banks' use of ATMs versus bank offices can be determined

while holding constant the many other influences that affect cost and profit differences among banks. The benefits from joint use of ATMs and branches will be reflected in a scope economy measure. This measure compares the cost or profit of providing deposit services using ATMs and branches jointly versus the cost or profit of using each delivery method separately. Scope economies exist when the cost (profit) of using both delivery methods jointly is lower (higher) than when used separately.[15]

The cost function used expresses total bank operating plus interest costs (C) as being determined by the total deposit, loan, and security output services a bank provides (q_i); the number of banking offices maintained (B); the number of ATMs a bank owns (ATM); and the labor, physical capital, and deposit input prices a bank faces (r_k). More formally, the cost function $C(q_i, B, ATM, r_k)$ is specified using a composite functional form. Developed by Carroll and Ruppert (1984, 1988), the composite form has been shown to provide stable estimates of scope economies, in contrast to other functional forms (Pulley and Braunstein 1992; Pulley and Humphrey 1993). This form has been simplified[16] and can be expressed as:

$$\begin{aligned} C^{(\phi)} = \{[\alpha_0 + \Sigma\alpha_i q_i + \tfrac{1}{2}\Sigma\Sigma\alpha_{ij} q_i q_j + \delta_B B + \tfrac{1}{2}\delta_{B,B} B^2 + \delta_{ATM} ATM + \\ \tfrac{1}{2}\delta_{ATM,\,ATM} ATM^2 + \delta_{B,\,ATM} B \cdot ATM + \Sigma\alpha_{iB} q_i B + \\ \Sigma\alpha_{iATM} q_i ATM] \cdot \exp[\Sigma\beta_k \ln r_k]\}^{(\phi)} + u, \end{aligned} \tag{1}$$

where the superscript (ϕ) refers to the Box-Cox transformation. In sum, there are three banking output services, two deposit delivery methods, and three input prices specified in (1). Further estimation and data details are noted in the appendix.[17]

[15] An alternative way to quantify the trade-off between branches and ATMs would be to determine the (Allen partial) elasticity of input substitution between these two deposit service delivery methods. Unfortunately, accurate data by individual banks on the total cost of supplying only transaction services and the per-transaction price of using an ATM or a branch office needed to compute such a measure from a "transaction cost function" are not generally available. Similarly, detailed transaction volume data for individual banks are also not generally available to derive this measure from a "transaction production function."

[16] The simplifications are that the price-output interaction and price-squared terms that are specified in the full composite model have been deleted in order to reduce collinearity problems and to focus on only those variables and relationships thought to be most important. The coefficient symmetry restriction and the restriction that the three price terms sum to 1.0 for input price linear homogeneity are imposed in estimation.

[17] A translog cost function could be obtained from (1) if $\phi = 0$ and the terms inside the brackets were multiplicative. While the translog form is log linear and thus easier to estimate than is the nonlinear composite form in (1), the translog form does not provide stable and robust estimates of banking scope economies while the composite does. This was shown in Pulley and Humphrey (1993).

Cost Scope Economies Between Branch and ATM Deposit Delivery Methods

Cost savings arise when the predicted total cost of delivering deposit services using offices (B), along with a minimal amount (ϵ) of ATMs, *plus* the predicted cost of using ATMs, along with a minimal amount of offices, *is larger than* the predicted cost of using the median amount of both delivery methods. Expressed formally, this condition is:

$$C[q_i, B(1-\epsilon), \epsilon ATM, r_k] + C[q_i, \epsilon B, ATM(1-\epsilon), r_k] > C(q_i, B, ATM, r_k).$$

This is the only way the costs associated with relying on either offices or ATMs to deliver deposit services can be properly compared while keeping the total use or scale of offices and ATMs constant at their median value.[18] In the inequality, the minimal amount of offices or ATMs (ϵB or ϵATM), added to their use when they are being primarily relied upon to deliver deposit services ($B[1-\epsilon]$ or $ATM[1-\epsilon]$), sum to their median values when used jointly (B and ATM).[19] The minimal amount of either delivery method used (ϵ) is set at 20 percent of their median values since it is not realistic to presume, in today's world, that deposit services will generally be delivered only through banking offices, and certainly not only through ATMs.[20]

The percent amount of cost savings is determined from:

[18] It is not appropriate to compare, say, $C[q_i, B(1-\epsilon), \epsilon ATM, r_k]$ with $C(q_i, B, ATM, r_k)$ since the total use of B and ATM would not be kept constant in the cost comparison. Thus we also need $C[q_i, \epsilon B, ATM(1-\epsilon), r_k]$ so the total use of B and ATM on both sides of the inequality are the same and the cost difference measured will be due to a different *mix* of deposit delivery methods.

[19] The distributions of B, ATM, and the other variables are skewed to the right, so median values are used rather than the means. Importantly, because the cost of producing all banking output is counted twice on the left-hand side of the inequality but only once on the right-hand side, an adjustment is required to the usual scope formula. Specifically, the predicted costs of producing all banking output $C(q_i, r_k)$ has to be subtracted from one of the cost estimates on the left-hand side of the inequality for a proper cost comparison to be made. $C(q_i, r_k)$ is computed using the estimated cost function (1) but with α_0, B, ATM, and their interactions with q_i all set to zero. Although not shown, the same adjustment is applied in (2) by subtracting $C(q_i, r_k)$ in the numerator.

[20] In addition, it has been demonstrated that some cost functions used to predict banking costs at zero or low levels of output provide a more accurate estimate of scope economies when the points of evaluation (here at ϵB and ϵATM) are within the range of the data and reasonably distant from zero. The minimal number of offices (ϵB) for the set of banks in 1992 is $.20(102) = 20$ which is contained within the range of the office data (where the sample minimum number of offices is 2 and the sample maximum is 1,827). Similarly, ϵATM is $.20(103) = 21$ and is contained within its range of 2 to 1,678.

$$\text{SCOPE} = \frac{C[q_i, B(1-\epsilon), \epsilon ATM, r_k] + C[q_i, \epsilon B, ATM(1-\epsilon), r_k] - C(q_i, B, ATM, r_k)}{C(q_i, B, ATM, r_k)} \quad (2)$$

but is perhaps clearer when expressed in words:

$$\frac{\text{cost of using offices} + \text{cost of using ATMs} - \text{cost of using both}}{\text{cost of using both}}.$$

Data from a special *American Banker* survey on ATM ownership for 161 large and small bank holding companies over 1991 and 1992,[21] augmented with Call Report information discussed in the appendix for the same two periods, are used to estimate the banking cost function (1) and compute the apparent cost savings from substituting ATMs for banking offices in (2).

Cost Savings from ATM Use

The estimated cost savings from joint use of ATMs and branch offices to deliver deposit services in 1991 and 1992 is shown in Table 3. Our preferred case—because it is the most realistic—is where the minimal amount of either ATMs or banking offices represents 20 percent of their median value and is in boldface in the table (where $\epsilon = .20$). Evaluated at this point, the estimated cost savings are -2.5 percent in 1991 and -1.4 percent in 1992. The negative value indicates that costs are *higher,* not lower, when ATMs and offices are jointly used to deliver deposit services. Because the ratio of total bank interest and operating expenses to total assets is 7.2 percent, a 2.5 to 1.4 percent increase in total cost due to ATM use would effectively translate into a possible decrease in ROA of 18 to 10 basis points.[22] Put differently, the substitution of ATMs for traditional banking offices represents a "technological change" in deposit service delivery methods that apparently has led to a permanent 2.5 to 1.4 percent upward shift in banks' average cost. However, only the 1991 point estimate of the cost effect of expanded ATM use is significantly different from zero. On balance, the scope measure indicates that ATMs have not lowered costs to banks. On the contrary, costs appear to have been marginally increased rather than reduced.[23]

The relative stability of the cost scope economy estimate is illustrated in Table 3 by changing the point of evaluation, letting ϵ vary between 0.0 and .50. At $\epsilon = 0.0$, which is a standard point for scope economy evaluation, the

[21] The ATM data were published in a special supplement to the *American Banker* for December 7, 1992.

[22] This assumes that (adjusted) revenues per dollar of assets (TR/TA) would be constant so that the basis point change in costs per dollar of assets (TC/TA) also is the change in ROA (since ROA = TR/TA − TC/TA).

[23] No conclusions are changed if, instead of all banks, only the set of low-cost banks on the efficient (thick) frontier were used in the analysis.

Table 3 Cost Scope Economies Between Branch Offices and ATMs: Composite Functional Form

Minimum Percent Use of Alternative Deposit Delivery Method (ϵ)		Scope Economy Estimates 1991 Cost Savings	1992 Cost Savings
(scope)	0.0	−.025 (.021)	−.027 (.036)
	.01	−.025 (.020)	−.026 (.035)
	.05	−.025 (.018)	−.023 (.032)
	.10	−.025 (.016)	−.020 (.028)
	.20	**−.025 (.013)***	**−.014 (.023)**
	.30	−.025 (.012)*	−.011 (.021)
	.40	−.024 (.012)*	−.008 (.019)
(scale)	.50	−.024 (.012)*	−.007 (.019)

* Significantly different from zero at the .05 level.

Notes: Cost scope economies are computed from equations (1) and (2). Profit scope economies are computed in a similar manner. Asymptotic standard errors are in parentheses. All values have been rounded off. None of the cost or profit scope measures are significantly different from zero in 1992, as all t ratios are less than 1.00. See Mester (1987), pp. 436–37, for the method used to compute the standard errors.

estimated cost increase is 2.5 percent in 1991 and 2.7 percent in 1992 but neither value is significantly different from zero.

At the other extreme, when $\epsilon = .50$, the scope calculation actually gives a measure of cost scale economies (see Pulley and Humphrey [1993]). When $\epsilon = .50$, the scope formula (2) compares the predicted costs of two banks each using 50 percent of the median number of offices and ATMs with the predicted costs of one bank using 100 percent of the median number of both delivery methods. Thus the mix of deposit delivery methods is unchanged but their scale of use is being doubled. This is in direct contrast to when $\epsilon = 0.0$ where the scale of use is held constant at the median but the mix of delivery methods is being varied (giving scope economies). The scale economies associated with using more of both branches and ATMs is estimated to raise costs by 2.4 percent in 1991 and 0.7 percent in 1992, but only the 1991 value is significantly different from zero.[24]

In sum, neither the scale nor the scope cost economy measures associated with the delivery of deposit services suggest lower costs. The point estimates are robust to different points of evaluation and, if anything, suggest that costs have risen, not fallen. The statistical significance of the increased cost results,

[24] Note that this is not the same thing as scale diseconomies for the production of deposit and loan services plus their delivery to bank customers. Overall, statistically significant output scale economies exist for smaller institutions but constant average cost—or not important scale economies—seems to be the rule for the largest banks (Berger and Humphrey 1991).

however, is weak as only a few points of scope and scale economy evaluation were significantly different from zero.[25]

5. PROFIT EFFECTS OF SUBSTITUTING ATMS FOR BRANCHES

A Nonstandard Profit Function Model

While costs do not fall as the mix of ATMs and offices used to deliver deposit services is varied, the same may not be true for bank profits. As noted earlier, fees are charged for ATM use. Just as important, the convenience provided by ATMs may enable a bank to retain a more profitable customer base: revenues from non-deposit services may be higher; a bank may be able to pay a lower interest rate on deposits; and a higher monthly minimum balance on deposit accounts may be required. All of these influences could lead to higher bank profits.

The approach to determine the effects of ATMs on bank profits closely follows the approach used to determine cost scope economies above. Profit scope economies are determined from a composite multi-output profit function where bank net income replaces total cost in equations (1) and (2).[26] This reflects a nonstandard profit function. With a standard (textbook) profit function, bank net income would be a function of exogenous output and input prices since the markets for banking outputs and inputs would be assumed to be perfectly competitive. With a nonstandard profit function, banks are assumed to have some market power to vary output prices with their assessment of the value of the product mix offered to consumers, *or* consumers value different mixes of services and bid up prices when these services are offered jointly in a competitive market.[27]

Profit Scope Economies Between Branch and ATM Deposit Delivery Methods

Profit scope economies are computed in an analogous manner to cost scope economies above. Profit scope economies arise when the predicted net income associated with delivering deposit services using offices (B), along with a minimal amount (ϵ) of ATMs, *plus* the predicted net income associated with using

[25] There is some indirect support for this result. Berger, Leusner, and Mingo (1993) found that one large bank provided far too many banking offices: the average office was only about one-half the efficient size, and if these smaller offices were consolidated, total costs could fall by 4 percent.

[26] Specifically, where NI is bank net income, $\ln NI$ replaces $\ln C$ in (1) and $NI(q_i, B, ATM, r_k)$ replaces $C(q_i, B, ATM, r_k)$ in (2).

[27] Some studies supporting price-setting behavior in markets for banking output are Hancock (1986), Hannan and Liang (1990), and English and Hayes (1991).

ATMs, along with a minimal amount of offices, *is smaller than* the predicted net income associated with using the median amount of both delivery methods. Thus profit scope economies exist—and profits are higher—if the scope measure is positive (just as cost savings would exist if the cost scope measure were positive).

Increased Profits from ATM Use

The estimated increase in net income from joint use of ATMs and branch offices to deliver deposit services is shown in Table 4. Our preferred case is still where the minimal amount of either ATMs or banking offices represents 20 percent of their median value and is in boldface in the table (where $\epsilon = .20$). Evaluated at this point, the estimated increase in bank net income is 3.6 percent in 1991 and 1.6 percent in 1992. Since ROA in 1992 was 92 basis points, the increased use of ATMs appears to have permanently contributed about 3.3 to 1.5 basis points to banks' ROAs. However, only the 1991 profit scope economy measure is significantly different from zero. At the usual point of scope economy evaluation of $\epsilon = 0.0$, neither profit scope measure is significant. Therefore, although the point estimates show a rise in bank net income, ATMs seem to have only marginally raised net income or profits to banks. The same conclusion applies to the profit scale measure (at $\epsilon = .50$) as this value is only significant in one year.

Table 4 Profit Scope Economies Between Branch Offices and ATMs: Composite Functional Form

Minimum Percent Use of Alternative Deposit Delivery Method (ϵ)		Scope Economy Estimates	
		1991 Profit Increase	1992 Profit Increase
(scope)	0.0	.049 (.030)	.031 (.037)
	.01	.048 (.029)	.031 (.036)
	.05	.045 (.026)	.027 (.033)
	.10	.042 (.022)	.023 (.030)
	.20	**.036 (.017)***	**.016 (.024)**
	.30	.032 (.015)*	.011 (.022)
	.40	.029 (.014)*	.009 (.020)
(scale)	.50	.029 (.014)*	.008 (.020)

* Significantly different from zero at the .05 level.

Notes: See Table 3.

6. SUMMARY AND CONCLUSIONS

The greatest change in the availability of deposit services over the last two decades has been the introduction of ATMs to augment, and replace, the traditional bank branch office in delivering these services. In 1973, there were 40,600 banking offices and less than 2,000 ATMs. On average, one banking office or ATM served 3,700 individuals (age 18 and older). ATMs were not intensively used as there were only five ATMs for each 100 banking offices. By 1992, there were 63,900 offices and 90,000 ATMs. Now there are three banking offices or ATMs for each set of 3,700 individuals—an expansion of convenience per person of over 200 percent. As a total, there are now 141 ATMs for each 100 banking offices. The increased availability of ATMs has benefited bank customers by both expanding the number of locations where deposit services can be obtained and by the fact that ATMs are typically "open" 24 hours a day.

Unfortunately, the expectation that ATMs would reduce bank costs has not been realized. Indeed, costs appear to be slightly higher, although the effect is weak. It is true that substantial scale economies exist for ATMs and that current transaction volumes are high enough to realize these economies. However, the potential benefits which should follow from the fact that an ATM transaction costs about half as much as a similar transaction in a traditional banking office has been largely offset by depositors who, because of the increased convenience of ATMs, use them up to twice as often as they previously used a banking office. Thus while ATMs were successful in reducing the cost of each depositor transaction, depositors increased the number of transactions, leaving total costs relatively unchanged or slightly higher. This suggests that the cost savings which could have been reaped by banks by substituting ATMs for branch offices has instead largely flowed to depositors who have shown their preference for the increased convenience provided by ATMs by substantially expanding the number of transactions they undertake.

The negative effect from higher costs can be offset if the revenues raised from bank provision of ATMs have been sufficient to raise bank profits. While profits are higher with ATM use, the effect is weak and is not consistently significant. Even so, profits appear to be marginally higher with ATM use, which likely represents a small net benefit to banks. Overall, however, it is probably the case that users of bank deposit services have benefited more from the change in the delivery of these services than have the banks.

APPENDIX

The ATM data are for 161 bank holding companies for 1991 and 1992 from the *American Banker* (special supplement, December 7, 1992) plus Call Report data on these same institutions for the same time periods. The medians of the data used are shown in Table A1 for 1992, while the parameter estimates for 1992 are in Table A2.

Table A1 Median Values of the Data: All Banks in 1992

Total cost (C)	\$.372 b	Number of offices (B)	102
Net income (NI)	\$.042 b	Number of ATMs (ATM)	103
Value of all deposits (q_D)	\$4.163 b	Price of labor (r_L)	\$33,200/yr.
Value of loans (q_L)	\$2.647 b	Price of capital (r_K)	.326
Value of securities (q_S)	\$1.512 b	(depreciation/book value)	
(b = billion)		Price of deposits (r_D)	4.50%

Note: Sample size was 161 for the cost function but 152 for the nonstandard profit function: nine observations with negative net income were deleted from the 1992 estimation.

Table A2 Parameter Estimates: Composite Cost and Profit Functions for 1992

Coefficient	Variable	Cost	Profit
ϕ	Box-Cox Parameter	.341*	.347*
α_0	Constant	−7.0E + 04	−6.9E + 04
α_D	Total Deposits	.305*	.321*
α_L	Loans	.438*	.426*
α_S	Securities	.407*	.415*
α_{DD}	(Deposits)^2	−.31E − 08	−.12E − 08
α_{LL}	(Loans)^2	−.35E − 07	−.25E − 07
α_{SS}	(Securities)^2	.26E − 07	.27E − 07
α_{DL}	Deposits · Loans	.18E − 07	.15E − 07
α_{DS}	Deposits · Securities	.16E − 07	.15E − 07
α_{LS}	Loans · Securities	−.59E − 07	−.59E − 07
δ_B	Offices	−1.1E + 03	−9.5E + 02
δ_{ATM}	ATMs	1.3E + 03	1.3E + 03
$\delta_{B,B}$	(Offices)^2	−2.39	−4.98
$\delta_{ATM,ATM}$	(ATMs)^2	−4.61	−4.56
$\delta_{B,ATM}$	Offices · ATM	7.74	9.21
$\delta_{D,B}$	Deposits · Offices	−.38E − 03	−.36E − 03
$\delta_{L,B}$	Loans · Offices	.57E − 03	.54E − 03
$\delta_{S,B}$	Securities · Offices	−.33E − 03	−.29E − 03
$\delta_{D,ATM}$	Deposits · ATMs	−.19E − 03	−.19E − 03
$\delta_{L,ATM}$	Loans · ATMs	.97E − 04	.45E − 04
$\delta_{S,ATM}$	Securities · ATMs	.80E − 03	.82E − 03
β_L	ln(r_L)	.127*	.119*
β_D	ln(r_D)	.783*	.779*
Log of the likelihood function		99.54	94.26

* Statistically significant at the .05 level.

Note: Although it is difficult to identify precisely the individual first- and second-order coefficients and interaction terms in a second-order (quadratic or log-quadratic) output specification, functions of those coefficients—such as the scope measure—can be identified with greater precision since correlations among coefficients are accounted for in the formulas for (approximate) asymptotic standard errors.

REFERENCES

Amel, D. "State Laws Affecting the Geographic Expansion of Commercial Banks," Working Paper. Washington: Board of Governors of the Federal Reserve System, September 1993.

Bank Administration Institute. *Checking Account Usage in the United States*. Park Ridge, Ill.: Bank Administration Institute, September 1979.

Barthel, M. "ATM Fees Rise for Usage at Customer's Bank," *American Banker,* September 17, 1993a, pp. 1 and 17.

_____. "No-Teller Retail Transactions Seen Hitting 40% by Yearend," *American Banker,* January 4, 1993b, pp. 1 and 3.

_____. "ATM Growth Put a Lid on Branch Costs," *American Banker,* December 7, 1992, pp. 1 and 3.

Baumol, W. "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quarterly Journal of Economics*, vol. 66 (November 1952), pp. 545–56.

Berger, A. "The Economics of Electronic Funds Transfers," Working Paper. Washington: Board of Governors of the Federal Reserve System, October 1985.

_____, and D. Humphrey. "The Dominance of Inefficiencies Over Scale and Product Mix Economies in Banking," *Journal of Monetary Economics*, vol. 28 (August 1991), pp. 117–48.

Berger, A., J. Leusner, and J. Mingo. "Efficiency of Bank Branches Estimated Using a Fourier-Flexible Frontier Approach," Working Paper. Washington: Board of Governors of the Federal Reserve System, December 1993.

Board of Governors of the Federal Reserve System. *Functional Cost Analysis.* Washington: Board of Governors, 1991.

Boeschoten, W. *Currency Use and Payment Patterns*. Financial and Monetary Policy Studies, Vol. 23. Dordrecht, the Netherlands: Kluwer Academic Publishers, 1992.

Carroll, R., and D. Ruppert. *Transformation and Weighting in Regression.* New York: Chapman & Hall, 1988.

_____. "Power Transformations When Fitting Theoretical Models to Data," *Journal of the American Statistical Association,* vol. 79 (June 1984), pp. 321–28.

Daniels, K., and N. Murphy. "The Impact of Technological Change on the Currency Behavior of Households: An Empirical Cross-Section Study," Working Paper. Richmond, Va.: Virginia Commonwealth University, 1993.

English, M., and K. Hayes. "A Simple Test of Market Power," Working Paper. Dallas: Southern Methodist University, 1991.

Federal Deposit Insurance Corporation. *Statistics in Banking*. Washington: Federal Deposit Insurance Corporation, various issues.

Hancock, D. "A Model of the Financial Firm with Imperfect Asset and Deposit Elasticities," *Journal of Banking and Finance*, vol. 10 (March 1986), pp. 37–54.

Hannan, T., and N. Liang. "Inferring Market Power from the Time-Series Data: The Case of the Banking Firm," Finance and Economics Discussion Series No. 147. Washington: Board of Governors of the Federal Reserve System, January 1990.

Humphrey, D., and A. Berger. "Market Failure and Resource Use: Economic Incentives to Use Different Payment Instruments," in D. Humphrey, ed., *The U.S. Payment System: Efficiency, Risk and the Role of the Federal Reserve*. Boston: Kluwer Academic Publishers, 1990, pp. 45–86.

Laderman, E. "The Public Policy Implications of State Laws Pertaining to Automated Teller Machines," Federal Reserve Bank of San Francisco *Economic Review,* Winter 1990, pp. 43–58.

McAndrews, J. "The Evolution of Shared ATM Networks," Federal Reserve Bank of Philadelphia *Business Review,* May/June 1991, pp. 3–16.

Mester, L. "A Multiproduct Cost Study of Savings and Loans," *Journal of Finance*, vol. 42 (June 1987), pp. 423–45.

Pulley, L., and Y. Braunstein. "A Composite Cost Function for Multiproduct Firms with an Application to Economies of Scope in Banking," *Review of Economics and Statistics*, vol. 74 (May 1992), pp. 221–30.

Pulley, L., and D. Humphrey. "The Role of Fixed Costs and Cost Complementarities in Determining Scope Economies and the Cost of Narrow Banking Proposals," *Journal of Business*, vol. 66 (July 1993), pp. 437–62.

Savage, D., and D. Humphrey. "Branch Laws and Banking Offices," *Journal of Money, Credit, and Banking*, vol. 11 (May 1979), pp. 227–30.

Tracey, B. "Study Sees a 20% Drop in Branches," *American Banker*, November 22, 1993, pp. 1 and 16.

van der Velde, M. *ATM Cost Model: A Survey of Fully Weighted and Incremental ATM Transaction Costs—1984*. Rolling Meadows, Ill.: Bank Administration Institute, 1985.

Walker, D. "Electronic Funds Transfer Cost Models and Pricing Strategies," *Journal of Economics and Business*, vol. 33 (Fall 1980), pp. 61–65.

_____. "Economies of Scale in Electronic Funds Transfer Systems," *Journal of Banking and Finance*, vol. 2 (June 1978), pp. 65–78.