Direct Investments in Securities: A Primer

RAMON P. DEGENNARO

The author is the SunTrust Professor of Finance at the University of Tennessee and a visiting scholar at the Atlanta Fed. *He gratefully acknowledges the support of a University of Tennessee Finance Department Summer Faculty Research Award, a CBA Scholarly Research Grant, and a University of Tennessee Professional Development Award. T. Shawn Strother and Samuel L. Tibbs supplied excellent research support and, along with Gerald Dwyer, Mark Fisher, Paula Tkac, and Larry Wall, provided helpful discussions. He also thanks the research librarians at the Federal Reserve Bank of Atlanta.*

Equity investors today face the same problems that previous generations of investors have faced: transactions costs, diversification, and the relatively large dollar amounts necessary to purchase certain assets. Investors want to minimize transactions costs, but doing so usually means buying round lots (100 shares), which implies large initial investments. Investors also want diversification, but that, too, requires large investments. For investors of constrained means, direct stock ownership has brought high fees and inadequate diversification or has simply been impossible.

Recently, dividend reinvestment plans and their more general cousins, direct investment plans, have virtually eliminated the problems of direct stock ownership by permitting investors to bypass traditional investment channels, such as securities brokers.[1] For the purposes of this article, a dividend reinvestment plan is defined as a mechanism that permits shareholders to reinvest the dividends paid on their shares in additional shares automatically, without the use of a broker. These plans may or may not restrict investors to being current shareholders. If the firm does not restrict its plan to current shareholders, instead permitting them to purchase their first share directly from the company without resorting to a broker, then the plan is also what is called a direct investment plan or an open-enrollment plan. For brevity, in this article both types of plans are labeled DRIPs and are differentiated only when the distinction is relevant. Also, following common usage, the redundant term "DRIP plan" is occasionally used.

DRIPs are not a different class of security, such as swaps or futures contracts. Rather, they are a new way of selling the traditional equity security. The privileges and obligations of equity ownership are unchanged. For example, DRIP investors retain all voting rights and receive all mailings, including annual reports and proxy statements. For taxable investors, dividends are still taxable income, and sales still generate capital gains or losses. DRIP investors are still subject to the rules of the stock transfer agent and to state and estate taxes. Many companies allow their DRIP to be used as a vehicle for an individual retirement account.

The key date in the proliferation of DRIPs was December 1, 1994. On that date, the Securities and Exchange Commission (SEC) granted an exemption from Rule 10b-6, essentially approving two model plans. This exemption eased restrictions on implementing and marketing these low-cost plans, cutting the time to set up a plan from as much as two years to under five weeks. Along with rapidly advancing technology, DRIPs have driven transaction costs to a bare minimum and brought diversified stock ownership to investors whose portfolios are well below modest. Companies may sell shares directly to investors without the services of brokers or investment bankers, in many cases absorbing all costs so

that the investor's transaction cost is measured in pennies. As of this writing, at least fifty companies impose absolutely no transactions costs at all, often with tiny minimum investments. Trust Company Bank of New York permits investors to purchase shares directly from the company with an initial investment of as little as $25. This is less than the cost of a single share, and there are no transactions costs. In all, well over 1,100 corporations—over 5 percent of the firms listed in the 1999 Compustat database—offer some type of direct investment plan.

This article serves as a primer on direct investment plans. The discussion describes how the financial services industry has evolved to address the problems facing the small investor, identifies the remaining limitations, and presents reasons why companies might offer such plans. The article then describes the data and identifies empirical differences between the types of companies that offer DRIPs and those that do not. Finally, the discussion speculates about the future of direct investments and provides conclusions.

DRIPs are not a different class of security, such as swaps or futures contracts. Rather, they are a new way of selling the traditional equity security. The privileges and obligations of equity ownership are unchanged.

History and Overview

Until the later portion of the twentieth century, equity investors of minor means may well have bemoaned the problems they faced. The financial services industry, however, recognizes that identifying and resolving consumers' financial problems is a profit opportunity. The industry attacks these problems from two directions. One approach involves pooling assets of many small investors, using professional managers to operate the fund. Mutual funds and closed-end investment companies are common examples. The other approach relies on different delivery systems to reduce transactions costs and the size of initial investments, thereby preserving the individual investor as the direct owner. Direct investment plans result from this approach. This section traces the development of mutual funds and DRIPs, clarifies similarities and differences between them, and describes the limitations of both of these investment options.

Some history. Mutual funds and closed-end investment companies were among the first innovations to use emerging technology to address small investors' needs. The Investment Company Institute credits the Scudder funds with opening the first no-load mutual fund in the 1920s (Carlson 1997). Such funds failed to generate much interest initially, and growth, at least in terms of dollars, was slow. By 1945, mutual funds' assets were about $1 billion, and by 1955 the figure was only $8 billion (Bogle 1982). By 1999, though, total equity investments under fund management had risen an average of over 15 percent annually, to $4.041 trillion.

The concept behind a mutual fund is simple: Fund managers pool money from small investors to permit large purchases of many stocks. Through this indirect ownership mechanism, each investor receives or bears a pro rata share of the fund's earnings or losses and pays a similar share of any costs the fund bears. This financial innovation solves, or nearly solves, many of the problems small investors encounter. First, large purchases have proportionally lower transactions costs. Thus, mutual fund investors pay lower transactions costs. Second, diversification becomes easy. Because fund managers invest a large pool of individuals' assets, managers are able to invest in many different companies in many different industries. In addition, investors may choose between index funds and actively managed funds. Index funds seek to match the returns on a popular index, such as the Standard and Poor's 500. By extension, managers of these funds do not attempt to uncover underpriced securities. Actively managed funds, by contrast, seek to purchase securities that fund managers believe are likely to outperform the market in general. Direct investments in individual securities remain extremely costly for small investors, though.

Another approach to solving the problems small investors face was the New York Stock Exchange's (NYSE) Monthly Investment Plan, which began in 1954 and ended in 1976.[2] This plan, which was operated by the NYSE itself, permitted individuals to invest in about 1,200 stocks, starting with as little as $40. Total fees were small for that time, about 6 percent for investments under $100. Moreover, there were no fees to open an account, no annual dues, and no obligation to invest. Participants could reinvest dividends for a small fee and could sell shares through the program. They could even purchase fractional shares if their investment did not purchase an integer number of shares. Participation in the NYSE's Monthly Investment Plan peaked in 1970, but despite its apparent appeal for small

investors, participation began to decline. The NYSE terminated the plan in 1976.

By 1976, and perhaps contributing to the demise of the Monthly Investment Plan, direct investment plans had emerged. DRIPs have obvious similarities to the NYSE's Monthly Investment Plan. Among these are low or even no commissions and often no explicit transactions costs at all. As with mutual funds, DRIP investors enjoy full investment of their funds because plans offer fractional shares. Initial investments are very low, and additional investments are convenient, even with tiny amounts. Coca-Cola, for example, accepts contributions of as little as $10 from plan participants. Yet wealthier investors can invest large amounts optionally. Unilever permits optional investments of up to $100,000 per year while for Southern Company the figure is $150,000 per year.

By the end of the Monthly Investment Plan, companies also had begun to incorporate various new features into their plans. In 1972, almost eighteen years after the NYSE introduced its Monthly Investment Plan, Long Island Lighting Company offered the first new-issue direct stock-purchase plan (Finnerty 1989). This plan offered two fundamental differences from the Monthly Investment Plan. First, investors no longer dealt with the NYSE. Instead, they dealt with Long Island Lighting. Second, unlike transactions through the Monthly Investment Plan, purchases through Long Island Lighting's direct purchase plan increased the number of outstanding shares and raised capital for the firm.

Other innovations followed. In 1972, AT&T was the first company to offer shareholders the opportunity to buy more shares at a discount from the market price. For example, a $95 investment might buy $100 worth of stock. Other companies offered safekeeping of shares, began accepting sales orders by telephone, or permitted dividends on one security to be reinvested in a different kind of the company's securities. For example, dividends on common shares could be used to buy preferred stock. At one time, ABT Building Products even offered a no-load direct-purchase plan despite not paying a dividend. Rather than reinvest dividends to increase their holdings of ABT, investors simply mailed a check to the plan administrator to purchase stock. Several foreign stocks also allow direct purchases despite not paying dividends.

Beginning in the early 1980s, some corporations no longer restricted plan participation to shareholders of record. Among the leaders were Citicorp, Control Data, and W.R. Grace. Even investors who were not currently shareholders could buy their initial and subsequent shares through the plan without a broker.

Dollar-cost averaging. In addition to the benefits of these innovations, both mutual funds and DRIPs are well suited for investors who believe that dollar-cost averaging makes sense. In brief, dollar-cost averaging involves investing (approximately) the same amount in the same security at periodic intervals. The result is that the investor purchases more shares when prices are lower so that his average purchase price is less than the arithmetic average of the shares' prices on the purchase dates. The apparent appeal of this procedure has led to widespread acceptance of its economic value despite evidence that it has no wealth implications.[3] Regardless of its value as an investing tool, though, many writers tout it as a wise strategy, and many investors use it. These investors see a benefit from participating in DRIPs because reinvesting regular dividend payments automatically results in dollar-cost averaging.

Mutual fund limitations. The success of mutual funds as an investment vehicle and the growing number of DRIPs available stand as strong evidence that these mechanisms serve investors' needs. Both approaches, though, have limitations. For example, the concept of purchasing a pro rata share of a portfolio has inherent drawbacks. In particular, three likely unavoidable features remain. First, no one expects the mutual fund's managers to work without pay. More generally, any mutual fund incurs expenses that must be recouped. Such costs fall into several categories. Management fees can range from under 0.5 percent to 7 percent or more. Updegrave (2001) reports that the typical U.S. stock mutual fund has operating expenses of 1.43 percent of assets annually. Load funds sold through brokers charge 12b-1 fees to cover marketing expenses. Administrative expenses, including mailing costs, tend to be smaller. Yet even relatively small fees can lead to large reductions in accumulated value over time. For example, an investor who invests $1,000 per year for forty years at 6 percent accumulates $154,762. If the total annual fund charges are only

1. Despite the similarity in names, direct investment plans discussed in this article have no relation to direct foreign investment.
2. Much of this section draws on Carlson (1997).
3. The apparent appeal probably traces to confusing the share-weighted average with the equal-weighted average of purchases. See also Constantinides (1979).

1 percent (about two-thirds of the average), so that the realized return is 5 percent, then the figure falls to $120,800—a net reduction of $33,962. The precise magnitude, of course, depends on management and its investment strategy. Investors seeking to minimize total fund charges can select an index fund, but even that involves some trading costs. The Vanguard Index Trust, perhaps the best-known index fund, reports a total expense ratio of only 18 basis points. For accounts under $10,000, though, Vanguard imposes a $10 annual maintenance fee. Thus, investors can select a fund based on its investment strategy and fees but still incur costs and surrender direct control of expense charges.

An investor in a mutual fund loses a portion of the value of tax-timing options. This loss occurs because a mutual fund essentially combines the options on each stock into just one option—the option pertaining to the entire portfolio.

A second disadvantage of investing in a mutual fund rather than holding stocks directly is that doing so makes it difficult for investors to diversify optimally. All investors in a single fund hold the same portfolio for the portions of their investment in the fund. A bank employee, though, might not want to hold the same portfolio as his identical twin, who is an auto worker. The bank employee probably wants to hold fewer bank stocks because his earnings at work are positively correlated with bank stocks. In other words, he could lose his job about the same time that the bank stocks in his investment portfolio decline. By the same logic, the auto worker might wish to own fewer automotive stocks. Such portfolio adjustments are difficult with mutual funds. Similarly, investors who wish to overweight individual securities that they believe are underpriced cannot do so with mutual funds alone.

Third, and perhaps most important, a mutual fund investor loses direct control of tax-timing options. U.S. tax law generally recognizes only realized capital gains. Thus, an investor who owns stock directly, unlike a mutual fund investor, can recognize losses for tax purposes by selling shares that have declined in price while deferring capital gains on shares that have increased in price.

How much are these tax-timing options worth? Constantinides (1984) shows that their value depends on several factors, including the investor's trading strategy, transactions costs, the tax rate on capital gains, and the stock's volatility. Clearly, though, their value can be substantial. For example, for a stock of average volatility and 4 percent round-trip transactions costs, the tax-timing option is worth about 3 percent of the stock's value. If transactions costs are negligible, then the option's value increases to 6 percent. For high-volatility stocks, the corresponding figures are 10 percent and over 14 percent, and other scenarios imply option values of more than 25 percent of the original investment.

An investor in a mutual fund loses a portion of the value of these tax-timing options. This loss occurs because a mutual fund essentially combines the options on each stock into just one option—the option pertaining to the entire portfolio. This option is worth less than the combined value of the individual options. This reduced value has been shown formally by Merton (1973), but the intuition is simple. Consider a portfolio of stocks with some winners and some losers. An investor holding an option on the entire portfolio cannot take tax losses without also taking gains. In contrast, an investor holding a portfolio of options (one on each security in the portfolio) can selectively realize losses for tax purposes while continuing to defer gains.

A second type of tax penalty on mutual funds can be enormous. By law, funds must distribute nearly all of their realized capital gains. CNNMoney (2001) reports that in 2000, while the average U.S. stock fund lost 10.1 percent, it still paid 9.19 percent in taxable distributions. The SEC reports that more than 2.5 percentage points of the average stock fund's total return is consumed by taxes each year.

In principle, a mutual fund manager can behave in exactly the same manner as an individual investor, recognizing losses on the underlying securities and deferring the gains. Indeed, the Vanguard Group began offering tax-managed funds in 1994 (Jacob 1996). However, fund investors cannot force the manager to distribute gains and losses in this way. A mutual fund investor can choose to invest in funds that are sensitive to the tax-timing issue, but even if she does, she has no explicit control of the timing of sales and must bear the consequences of the manager's decisions. Even the most tax-conscious mutual fund imaginable cannot consider other factors that affect an individual investor's tax position, such as changes in marginal tax brackets due to, say, changes in marital status or a spouse's decision to enter or leave the work force.

Investors in mutual funds can also find themselves with a tax liability if their fund closes. Mutual

funds can and do cease operations more often than people realize. The Vanguard Group reports that of the 356 general equity funds that existed in August 1976, fully 45 percent had ceased operations by the summer of 2001 (Vanguard 2001). If the fund liquidates, then investors bear a pro rata share of any capital gain—and of the resulting tax liability. Sometimes, a mutual fund merges with another fund. In this case there are no taxes due immediately, but the new shareholders inherit liability for capital gains earned before they acquired shares of the ongoing fund. Perhaps worst of all, investors have no control over either liquidation or merger.

Finally, mutual fund investors face the problem of accumulating the funds to meet minimum investment requirements. Though this problem is not inherent in the concept of an intermediary holding a diversified portfolio of stocks, most funds impose an investment minimum that exceeds those of direct investment plans. Of course, this analogy is not an apples-to-apples comparison. An investment in a single mutual fund might provide sufficient diversification for most people; this claim cannot be made for an investment in a single DRIP. Still, investors face the problem of accumulating the initial investment that funds require. The Vanguard Index Trust, for example, requires a minimum investment of $3,000 in most cases.

DRIP limitations. Investors who hold stocks directly through DRIPs face a different set of problems than those of mutual fund investors. Scholes and Wolfson (1989) say that DRIP investors bear a variety of implicit costs. For example, DRIP investors must become informed about the plans' details and must monitor the plans for changes in terms. Of course, mutual fund investors must also do this but for a much smaller number of investments. Though DRIP purchases often have no explicit cost, nearly all DRIPs provide for transactions costs when selling shares. Some even require plan participants to request stock certificates and to deliver them to a broker for sale in the traditional manner. Nor are transactions costs the only costs plan participants face: they also bear the costs of any tax implications of their direct equity holdings. Foremost among these are the usual taxes on dividends and capital gains. In addition, in some plans, the company pays commissions for the investor when the shares are purchased. If so, then the IRS treats such commissions as taxable income. Discounts on purchases are also taxable income.[4]

Tax rules also probably limit the value of the individual tax options that DRIP investors hold. Though the tax options in a DRIP are clearly more valuable than those in a mutual fund, they are unlikely to reach the levels that Constantinides (1984) calculates. Under the current U.S. tax code, gains and losses are calculated relative to the basis, which is usually the purchase price plus any transactions costs. DRIPs usually generate four purchases each year, so calculations of gains or losses tend to be tedious compared to an investment strategy built around larger purchases. One way around this problem is to receive dividends in cash. Nothing prevents an investor from doing so, and she can still make purchases through the plan if she wishes. In terms of the timing of purchases, this strategy is essentially the same one that an investor using a traditional broker would follow. Another solution is to sell all of the shares in a company as a block. This strategy lets the taxpayer use the average basis of the shares in the block as the basis for all shares in the block. A third strategy is to use one of the popular personal financial management software packages available today. Most record the basis and compute the gain or loss automatically when the shares are sold.

Investors who hold stocks directly through DRIPs face a different set of problems than those of mutual fund investors.

In defense of brokers. DRIPs and mutual funds do, of course, carry disadvantages for investors. Focusing solely on transactions costs ignores other advantages that traditional brokers can provide. For example, these investments offer less liquidity than investments through traditional brokers. DRIP and mutual funds investors cannot place limit orders or buy on margin, and execution of transactions is usually slower. In contrast, brokers are almost always faster in delivering the proceeds of sales. In addition, brokers offer a much wider variety of investment choices, such as bonds, options, money funds, collateralized mortgage obligations, unit trusts, and so on. Competent brokers offer useful advice regarding

4. These costs, however, can be reflected in the tax basis so that subsequent gains taxes are reduced and tax losses are increased. Thus, the tax liability is not only deferred but is also converted to a gain and usually taxed at a lower rate.

tailoring portfolios, such as matching investment opportunities with an individual investor's risk preferences, and can help monitor an investor's asset mix. For example, after a prolonged increase in stock prices, a busy investor might not realize that his portfolio contains a far greater proportion of risky assets than he prefers. A good broker can monitor and notify the investor of this situation—as well as the unhappy opposite case, when prices have declined and the risk profile is too conservative. Finally, some investors prefer having all of their transactions on one statement rather than receiving separate statements for each company or fund they own.

DRIP plans tend to attract a specific clientele. These plans are likely to provide a broad, relatively stable base of shareholders who, because they hold relatively small positions, are likely to be passive investors.

Many brokers provide some or all of these services at no explicit cost to their customers probably because brokerage firms usually keep their customers' securities in street name (that is, in the name of the brokerage on behalf of the customers), giving them the right to lend the securities for short sales and to collect any fees for doing so. In a competitive market for brokerage services, brokerage firms must provide some compensation for this right or else customers would move their accounts to firms that do. By contrast, DRIP and mutual fund investors must keep their securities in their own names. As a result, lending the securities is impractical, and the investors forfeit the fees they might gain. Two forces could tend to offset this disadvantage. First, the plan administrator may be able to lend the securities. If so, competition would tend to force him to compensate investors, just as it does for brokers. Second, if the plan administrator cannot arrange to lend for short sales, then company management may well view the resulting reduction in the number of shares available for shorting as a benefit. This situation would be especially true for DRIPs, and perhaps this circumstance explains why some plans offer such attractive terms to investors.

The best way of conceptualizing the role of brokers in relation to direct investment plans is to realize that brokers are no different from other middlemen. They can stay in business only if they add sufficient value to earn at least a normal profit. In general, these services are of relatively little value to investors with limited portfolios (that is, mostly stocks) using a buy-and-hold strategy. Thus, DRIP plans, in particular, tend to attract a specific clientele. These plans are likely to provide a broad, relatively stable base of shareholders who, because they hold relatively small positions, are likely to be passive investors.

Why Do Corporations Participate?

What type of company might prefer a clientele of buy-and-hold investors? More generally, why do firms offer direct investment plans? Street lore offers several possible explanations. Quite possibly, funds can be raised more cheaply through DRIPs. DRIPs do incur expenses such as telephone charges, added personnel, extra printing, mailings, and so forth. One estimate is that such costs are between $12 and $16 per account, and this figure is virtually independent of the number of shares held (Carlson 2000). DRIPs, however, substitute these direct costs for the investment banker and the related administrative, legal, and accounting fees when issuing new shares.

These cost savings can be large. Carlson (1996, 16) reports new-issue costs of between 5 percent and 15 percent of the equity issue. Eckbo and Masulis (1992) report that total issue costs as a percentage of gross proceeds average 6.09 percent for industrial firms and 5.53 percent for utilities. Underwriter costs alone account for about 90 percent of that amount. In addition, existing stockholders can be worse off as a result of an issue of additional shares. Asquith and Mullins (1986) use event-study methods to conclude that the two-day abnormal return for industrial firms that announce equity issues is –3.14 percent. For utilities, the figure is –0.75 percent. Eckbo and Masulis report returns of –3.34 percent and –0.8 percent for announcements of firm-underwritten offers for industrial and utility firms, respectively. Again, these costs affect the entire equity base, not just the new issue. Scholes and Wolfson (1989) report that the equity base largely avoids these costs with DRIPs.

In addition to avoiding some of these costs, new-issue DRIPs permit large sums to be raised. For example, the prospectus for OneOk, Inc., dated February 7, 2001, reports that, "This prospectus covers 4,424,502 shares. . . ." Given a share price on that date of about $22.25, almost $100 million could conceivably be raised—and raised quickly—under the terms of this single offering. South Jersey Industries raised $8 million with its DRIP in June 1990 alone. Scholes and Wolfson (1989) report such benefits in terms of the amount of dividends paid.

They report that firms with no discounts on reinvestments raise an average of 12 percent of the total common and preferred dividends they pay. If firms offer a 5 percent discount, then this amount rises to about 98 percent of the common and preferred dividends they pay. Clearly, many investors are reinvesting dividends or are making large optional payments. Why firms simultaneously pay dividends and encourage reinvestment in newly issued shares is a separate question, likely related to the question of why firms simultaneously pay dividends and raise funds through equity or debt issues.[5]

A second reason often given for the existence of DRIPs is that companies simply wish to provide a service to their owners. Goodwill is valuable, and owners who desire to increase their stake in their company want to do so in the lowest-cost manner. Certainly, high levels of telecommunications and computer technology are essential to administering such plans efficiently, and this has become easy and inexpensive in recent years. To the extent that a firm enjoys scale economies in transactions in its own stock, DRIPs are a logical option.

Third, having more shareholders could boost sales of a company's products. For example, an investor, even one who is not a current shareholder, can enroll in Bob Evans' DRIP with a minimum investment of only $50 and with no transactions costs. Once he is an owner, an investor may be more likely to eat at Bob Evans rather than at a competing restaurant. Owners are also more likely to refer new customers to the restaurant.

A fourth reason for the existence of DRIPs could be what economists call economies of scope. A company that provides several goods or services may have an advantage in satisfying consumers' needs if the consumers are shareholders. Because a shareholder is already on the company's mailing list and is familiar with the company, normal shareholder correspondence provides an easy, inexpensive way to approach these investors as customers for other services. For example, Regions Financial often includes pamphlets regarding refinancing home mortgages or second mortgages with its mailings to shareholders. ExxonMobil recently sent information promoting SmartPass, a transponder system designed to save customers time at gasoline stations and convenience stores. ExxonMobil also announced its participation in Upromise, a plan to assist families saving for college expenses.

A fifth reason is rarely mentioned. Most plan administrators usually retain the option to execute the plan's trades on more than one exchange or market. Thus, a company or its agent might collect fees for routing order flow.

A firm also might want to attract its own employees as shareholders. Employees who are not owners have greater incentives to shirk because a larger portion of the costs are borne by the company's owners. To the extent that employees own the firm, shirking becomes less attractive to them. This motivation explains in part the popularity of employee stock option plans (ESOPs) and 401(k) matching programs. Consistent with this idea, some companies permit their employees to purchase their first share of stock directly from the company while requiring others to use a broker. Such preferential treatment is impossible with a regular stock issue. The SEC would probably prohibit a public offering that was available only to employees of a specific company.

One commonly cited reason for offering DRIPs is that they generate price pressure by providing a steady stream of buyers, keeping share prices high. This argument is implausible. For this scenario to be true, DRIP investors must consistently be net buyers. Although this situation may occur around the time of dividend payments, there is no reason to expect it to occur during other periods. Even if DRIP investors were net buyers, that motivation would still be insufficient for a price-pressure argument to carry force. This argument must further assume that other investors make no adjustment in their purchases because of the higher prices around dividend dates. In fact, though, other traders would probably time their purchases to take advantage of such predictable price behavior. They would sell around dividend payment dates and buy at other times. In fact, considerable academic work has shown that price pressure tends to have little impact on share prices (for example, see Smith 1986).[6]

Clienteles. Clearly, offering a DRIP appeals to some companies, and buy-and-hold investors are more likely to use DRIPs. What type of firm might prefer a clientele of buy-and-hold investors? One obvious candidate is a company that offers many products and services so that it can benefit from cross-selling. Customers who purchase one product from a company are more likely to choose another from that same company rather than incur the costs of learning about competing products.

5. Both questions are fascinating and well beyond the scope of this paper.
6. Both Harris and Gurel (1986) and Ederington and Goh (2001) provide some evidence that price pressure is indeed large enough to measure. Both report that any price effects vanish within a few weeks.

A second candidate could be firms that are subject to regulation and are therefore more heavily exposed to the political process. Voters, unlike shareholders, are equally weighted. Other things being equal, having ten shareholders with fifty shares each is a better political base than having one owner with 500 shares. Having many shareholders (and thus many investors who are also voters) makes it less likely that government will impose onerous regulations on the company. The company can even claim that voters are small investors and set them against the allegedly helpless groups typically cited as the people protected by regulation. Especially in the case of utilities, owners are less likely to complain to regulators about rate increases or to demand tight environmental restrictions.

A reason for the existence of DRIPs could be what economists call economies of scope. A company that provides several goods or services may have an advantage in satisfying consumers' needs if the consumers are shareholders.

This advantage for regulated firms is magnified because management has routine access to its shareholders and can tell its side of any political story to more people at lower cost. For example, CH Energy Group, Inc., included copies of its chairman's remarks at its annual stockholders' meeting with its routine DRIP statements. About one-fifth of the remarks were dedicated to explaining the company's position regarding power shortages in California (Ganci 2001). Similarly, Duke Energy Corporation used two full pages of a letter to shareholders to explain and defend its position on the California crisis. This explanation included reports of investigations by the Federal Energy Regulatory Commission, the compliance unit of the California Power Exchange, and the Northwest Power Planning Council—none of which found any basis for the charges by government officials in California that electricity producers were artificially driving up power prices. The letter called for "the cooperation and support of the highest levels of State government" and added that "the regulatory process must be streamlined to encourage investments in new power plants and the market must be restructured to allow all participants the ability to manage and hedge their exposure to power and gas prices" (Priory 2001).

This line of reasoning can be carried still further: Not all voters are equally valuable to a company. In the case of utility firms, management would particularly want to have its customers and state residents also be owners. These companies should be expected to offer plans with features designed to entice these individuals to buy shares. In fact, some DRIPs do exactly that. For example, Carolina Power and Light requires investors to purchase their first share from a broker unless the prospective plan participant is a customer. In that case, the company will sell the first share directly to the customer. Until Central Fidelity Banks, Inc., was acquired by Wachovia Corporation, its plan required participants to be existing shareholders unless the prospective plan participant was a resident of the state in which it operated. State residents, of course, carry more weight with local politicians than residents of other jurisdictions.

Regulated industries such as public utilities and financial institutions are not the only ones that can benefit from improved public relations and political influence. Companies at risk of being regulated, or at risk of increased regulation, may concentrate their efforts on U.S. investors to provide a channel for disseminating the company's position on major issues. For example, Pfizer, Inc., is a major force in pharmaceutical products, another industry often targeted for government intervention. Prior to the presidential election of 2000, Pfizer's letter to shareholders stated: "In the heat of campaigning, rhetoric thrives. It would, however, be a sad day for American health care if anti-industry rhetoric were translated into policy. It would stifle pharmaceutical research, deprive millions of people of new treatments and herd every American senior into a vast drug-access scheme administered by government bureaucrats" (Clemente 2000).

One of the clearest attempts to rally shareholders to support a company appeared in a SCANA Corporation mailing. This letter explicitly asked stockholders to join the Association of SCANA Corporation Investors. This association was begun in 1978 "to help insure the Company received a fair rate of return on its shareholders' investment from the state utility regulatory body." The letter added that, "More recently, the Association represented the interests of its members and other shareholders in the debate over restructuring the electric utility industry in South Carolina. Association leaders testified before the South Carolina Public Service Commission and legislative committees while Association members from South Carolina explained our organization's position on this issue to their individual legislators" (Quattlebaum and Strock 2001).

Public relations is clearly an important component of shareholder mailings. To the extent that DRIPs increase the number of stockholders, such plans can play a part in maintaining a positive corporate image and lowering the costs of reaching them.

A company might also institute a direct investment plan to insulate and protect management. To the extent that DRIP investors hold small positions, they are less likely to be active in monitoring management. While it can make sense for an institutional investor holding millions of shares to take action against weak management, becoming informed about management practices and acting on this information is very unlikely to be worth the effort if one owns only a few hundred shares. Such investors are likely to vote with management, usually by proxy, or not to vote at all. Thus, an active investor faces an uphill battle to convince a majority of voting shares to support his position. Management benefits by becoming entrenched, and most research concludes that such entrenchment is detrimental to shareholders.

In summary, there are several reasons that corporations offer DRIPs. Not all of these reasons have equal appeal to all companies or industries. This reasoning suggests that there may be systematic differences between companies that offer DRIPs and those that do not. The next section explores this possibility empirically.

Comparisons between DRIP Companies and Their No-DRIP Counterparts

To explore direct investment plans empirically, this study examines the firms listed in the *Guide to Dividend Reinvestment Plans* (1999). According to the publisher, Temper of the Times Communications, Inc., this guide encompasses all firms that offered DRIPs on the publication date. Of these approximately 1,135 companies, 906 provided plan terms and are included in the 1999 Compustat annual database for 1999.

It might seem tempting to compare these 906 companies with the universe of Compustat firms without plans. The problem with this comparison is that DRIP firms are, on average, much larger than firms without DRIPs. For example, using total assets as the measure of size, the mean DRIP firm in 1999 has total assets of $13.87 billion compared to only $2.33 billion for firms without DRIPs. The mean DRIP firm in 1999 is more than five times larger. The likelihood of DRIP firms being a random sample of all companies in Compustat is less than 0.01 percent.

Clearly, large firms are more likely to offer DRIPs than small firms. This likelihood suggests that large firms have an advantage in sponsoring direct investment plans. This advantage is not too surprising since some administrative expenses are likely about the same for 50,000 shareholders as they are for 25,000 shareholders. Thus, the cost of providing DRIPs is lower per participant for larger companies. For some purposes, such as investing, this preponderance of large companies may not be a problem. Small companies would be underweighted in a portfolio comprising only DRIP companies, but many mutual funds also underweight small firms. For gaining an understanding of the economic forces driving the decision to offer a direct investment plan, though, this large size differential complicates the analysis because large firms differ from smaller ones in many ways. Not the least of these differences is access to capital markets; large firms have many more options to obtain financing. To circumvent this difference, this analysis constructs a size-matched sample based on total assets in 1999. Each of the 906 companies offering DRIPs and having data in 1999 is matched to a company without a plan, for a total of 1,812 companies. Paired differences are also computed for the variables. Some observations for certain variables are missing from some firms, however, so the number of matched pairs for many variables is less than 906.

A company might also institute a direct investment plan to insulate and protect management. To the extent that DRIP investors hold small positions, they are less likely to be active in monitoring management.

To obtain some evidence on how well the matching procedure worked, the mean total assets for the two groups of 906 companies are computed. Those without DRIPs have average total assets of $14.41 billion in 1999 while firms that offer DRIPs average $13.87 billion. The difference is less than 4 percent, and a t-test (0.25) is insignificant by any usual standard. Overall, the two groups are very similar in size. But the size-matching procedure can go only so far: For some ranges of total assets, there simply are not enough companies to provide a good match to each individual company. Thus, the difference in total assets of the 906 paired differences does differ statistically from zero.

The discussion in the previous section suggests that some industries might benefit more from

instituting DRIP plans than others. If that is true, then DRIPs would not be distributed randomly across industries. To test this assumption, a chi-square test using two-digit Dun & Bradstreet Standard Industrial Classification (SIC) codes is conducted.[7] This test rejects the hypothesis that DRIP firms are randomly distributed across industries. Some caution is in order here, as some industries have too few observations to merit too much faith in the results. Still, the results are illuminating. The likelihood that the departures from a random distribution are due to chance is less than 0.01 percent.

Table 1 shows the ten industries with the largest absolute deviations from expected outcomes if the distribution were random. The biggest departures from the expected distribution are, in descending order, the electric, gas, and sanitary services industries (DRIPs are over-represented), communications (under-represented), holding and other investment offices (over-represented), and business services (under-represented).

Clearly, large firms are more likely to offer DRIPs than small firms. This likelihood suggests that large firms have an advantage in sponsoring direct investment plans.

The higher concentration of electric and gas companies as DRIP providers makes sense because these industries tend to be regulated. Holding and other investment offices are over-represented because the category includes real estate investment trusts (REITs). REITs must distribute at least 95 percent of their earnings to shareholders to retain their preferred tax status. This limitation makes it nearly impossible for a REIT to grow using internal funds. Rather than continually going to the capital markets to raise funds, many REITs offer DRIPs to encourage reinvestment and essentially reduce the dividend yield.

Why might communications and business services be under-represented? The easiest answer is that, because some industries are over-represented, some must be under-represented, and communications and business industries happen to be among them. More insight can be gained, though, by realizing that the size-matching procedure, designed to eliminate the large difference in scale economies between the universe of DRIP companies and no-DRIP companies, has limitations in practice. First, because DRIP companies tend to be large, smaller companies tend to be eliminated, and the size-matched sample comprises larger companies than the universe of Compustat firms. Second, industries dominated by smaller companies tend to be under-represented in the size-matched sample. Just as there are not enough companies to permit a good match to each individual company, there simply are not enough companies to permit accurate matching within industries; size-matching the entire sample is the best available option.

The business services industry illustrates these effects. Business services companies, which include advertising agencies, pest control services, employment agencies, computer-related services, security systems, and so on, tend to have fewer total assets than most other companies do. In the universe of Compustat companies, the average total assets of business services companies is about $352 million compared to over $2 billion for other industries. The size-matching procedure reduces the discrepancy substantially: the average total assets of business services companies is about $5.08 billion compared to about $14.49 billion for other industries. The size differential declines from about seven-to-one to less than three-to-one, but business services companies still tend to be on the smaller side. This tendency might explain why the business services industry is under-represented among companies that offer DRIPs.

This explanation fails for the communications industry, however, because communications companies tend to be a little larger than average after size matching. A better explanation might be that this industry includes telephone communications and cable services, which grew rapidly during the late 1990s. Many of these companies paid low or no dividends at all.

These deviations from random distribution suggest that the concentration of DRIPs might be due to a dividend effect. That is, perhaps DRIPs appear more often in certain industries because those industries tend to pay dividends more often. The easiest way to check this supposition is to drop all companies that reported no dividends in 1999 and repeat the chi-square tests. The result shows that companies offering DRIPs are still concentrated in certain industries, but the specific industries differ. In terms of absolute deviations from the expected distribution, the biggest departure is again electric, gas, and sanitary services, which are over-represented. The insurance carriers industry is second, and it is under-represented relative to the expected distribution.

TABLE 1

The Ten Highest Absolute Deviations from the Expected Frequency of DRIP Plans, by Industry

	Actual DRIP frequency	Expected DRIP frequency	Difference
Electric, gas, and sanitary services	118	75.5	42.5
Communications	26	52.5	–26.5
Holding and other investment offices	79	54	25
Business services	11	32.5	–21.5
Insurance carriers	35	54.5	–19.5
Chemicals and allied products	63	45.5	17.5
Depository institutions	152	134.5	17.5
Nondepository credit institutions	8	23	–15
Amusement and recreation services	1	7.5	–6.5
Transportation by air	1	7	–6

Source: Author's calculations using data from Compustat and the *Guide to Dividend Reinvestment Plans*

Third are depository institutions, which are underrepresented. Holding and other investment offices slip to fourth and remain over-represented. One problem with this analysis, however, is that the sample sizes are much too small for the results to be reliable. Another is that almost all of the excluded companies (ones that paid no dividends) do not offer DRIPs. This fact points to a third problem: The decision to pay a dividend and the decision to offer a DRIP are not independent. Disentangling the effects of DRIPs from those of dividends themselves remains a difficult problem for future work.

Table 2 reports the 1999 sample means for Compustat data for the size-matched sample of 906 companies with DRIPs and 906 companies without DRIPs. It shows the means for companies with and without plans and reports *t*-statistics for a test of equality. It also contains similar results for the subset of paired differences.

These paired differences permit an analysis of variance—specifically, whether or not the paired differences between companies with DRIPs and companies without them are jointly nonzero. This distinction is important because the right-most column in Table 2 reports almost forty *t*-tests. Some of those tests are likely to appear statistically significant even if they are not. Economists call this a type I error, and the chance of committing it increases as the number of tests increases. An analysis of variance takes this possibility into account. The trade-off is that, if differences are found, the test provides no information about which variable or variables are the source of the difference. In such cases, further tests are necessary.

Here, the analysis of variance produces an *F*-statistic of 5.14. A value this large is very unlikely to be caused by chance, and the implication is that the magnitude of the paired differences between DRIP companies and no-DRIP companies is reliably different. The next task is to explore which of the variables are likely to be driving this result.

Previous discussion suggests that DRIPs are likely to provide a broad, relatively stable base of shareholders who, because they hold relatively small positions, are likely to be passive investors. The data support this. For example, in the size-matched sample of companies, firms with DRIP plans averaged 49,650 common shareholders in 1999 compared to only 23,460 for companies without such plans. The probability that this pattern is due to chance is less than 1 percent. In addition, companies with DRIPs had only 143.6 million common shares compared to 191.9 million traded by the shareholders of companies without plans. Put differently, the average number of shares traded per shareholder in a company offering a DRIP is about 2,890 shares annually compared to almost 8,200 for a company without a DRIP. Thus, the differences are economically and statistically significant. Using only the 521 paired companies with data on both firms, the results are similar. DRIP firms have almost twice the number of shareholders, but each of them trades only about one-third as much on average. DRIP companies generally have more stable shareholder bases.

7. SIC codes classify companies according to industry. For example, codes from 6000 through 6099 apply to the general category of depository institutions. Subcategories within this range represent specific different types of depository institutions.

TABLE 2

Means and *t*-tests, 906 Companies with DRIP Plans Compared to 906 Companies without, Size-Matched Sample, 1999

	Unmatched sample					Matched sample			
	No DRIP plan		DRIP plan			No. of paired	No DRIP plan	DRIP plan	*t*-statistic, paired
Variable	Observations	Mean	Observations	Mean	*t*-statistic	differences	Mean	Mean	differences
Total assets (MM$)	906	14,412	906	13,870	0.25	906	14,412	13,870	2.41*
PPE, gross (MM$)	711	4,710	657	5,779	–1.55	509	4,313	5,742	–2.82**
PPE, net (MM$)	851	2,428	831	2,644	–0.70	780	2,504	2,546	–0.18
PPE, capital expenditures (MM$)	751	535.05	712	484.81	0.54	587	501.89	475.36	0.35
Capital expenditures (MM$)	753	534.82	726	480.22	0.60	599	499.98	470.74	0.39
Research and development (MM$)	317	265.96	381	260.40	0.09	131	263.30	105.00	2.34*
Common equity (MM$)	897	2,394	905	2,645	–0.72	896	2,397	2,655	–0.91
Stockholders equity (MM$)	904	2,510	906	2,682	–0.49	904	2,510	2,685	–0.61
Net sales (MM$)	904	4,737	905	5,842	–1.79*	903	4,742	5,855	–2.54*
Interest expense (MM$)	758	266.72	741	268.99	–0.03	617	250.98	205.21	1.17
Dividends to common (MM$)	876	111.17	881	161.37	–2.45**	851	113.70	161.39	–2.98**
Dividends per share ($)	876	0.36	905	0.85	–8.37**	875	0.36	0.85	–8.20**
Payout ratio (%)	875	19.34	880	52.77	–4.23**	849	19.54	49.41	–3.85**
Dividend yield (%)	777	1.46	905	3.96	–5.07**	776	1.46	4.07	–4.91**
Number of common shares outstanding (MM)	862	181.43	903	207.73	–1.20	859	181.04	205.07	–1.29
Number of common shares traded (MM)	774	191.90	906	143.56	1.92*	774	191.90	128.20	2.50*
Treasury stock, number of shares (MM)	884	5.16	891	14.39	–4.12**	870	5.12	13.96	–4.09
Number of common shareholders (M)	675	23.46	713	49.65	–3.17**	521	25.25	48.85	–2.34*
Number of employees (M)	800	18.49	820	24.10	–2.08**	722	18.66	24.67	–2.22*
Interest income (MM$)	500	31.11	448	34.31	–0.35	261	31.09	44.40	–0.99
Sales, common & preferred (MM$)	757	155.93	712	51.07	4.32**	594	144.18	53.86	4.21**
Purchases, common & preferred (MM$)	706	75.49	711	170.21	–4.16**	551	72.90	157.38	–3.81**
Pretax income (MM$)	905	473.80	905	621.55	–2.09**	904	474.33	622.24	–2.80**
Net income (MM$)	906	313.83	906	403.33	–1.88*	906	313.83	403.33	–2.45*
Interest expense per share ($)	678	8.40	739	1.38	1.32	550	10.01	1.26	1.28
Net profit margin (%)	902	1.30	905	8.20	–3.06**	901	1.30	8.14	–3.09**
Return on stockholders' equity (%)	903	13.35	906	13.25	0.01	903	13.35	13.24	0.01
Pretax interest coverage (X)[1]	735	30.04	735	14.91	0.61	595	11.58	16.47	–0.26
Pretax profit margin (%)	902	5.45	905	12.28	–2.57**	901	5.47	12.22	–2.62**
Pretax return on assets (%)	905	3.30	905	6.31	–5.66**	904	3.31	6.31	–5.82**
Operating income before depreciation to total assets (%)	857	9.16	825	11.38	–4.56**	780	9.18	11.38	–4.45**
Aftertax interest coverage (X)[1]	735	22.87	735	9.84	0.53	595	3.09	10.79	–0.43
Aftertax ROE (common, %)	896	14.79	905	4.85	0.78	895	14.80	4.78	0.78
Aftertax return on total assets (%)	906	1.53	906	4.04	–5.53**	906	1.53	4.04	–5.67**
Debt ratio	905	0.69	905	0.69	0.44	904	0.69	0.69	0.46
Market-to-book (ratio)	766	3.07	903	2.87	0.46	763	3.06	2.72	0.76
P/E at fiscal year-end (ratio)	777	20.80	905	16.31	0.91	776	20.80	15.73	0.94
Market value of common stock at calendar year-end (MM$)	775	10,112	903	10,468	–0.21	772	9,995	9,409	0.36
Earnings per share	822	1.64	905	1.73	–0.21	821	1.64	1.69	–0.11

Note: The table is constructed so that positive *t*-statistics imply that the value for the companies without DRIPs is larger than for the companies with them. ** indicates significance at the 1 percent level; * indicates significance at the 5 percent level.

[1] Interest coverage is the ratio of income to interest expense. For example, $30 of income times $1 of interest expense yields an interest coverage of 30.

Source: Author's calculations using data from Compustat and the *Guide to Dividend Reinvestment Plans*

To the extent that employee ownership is advantageous, companies in labor-intensive industries would also be expected to offer more DRIPs for at least two reasons. First, if they have more employees, the advantage to be gained is presumably larger. Second, many employees are likely to be at most small investors, and DRIPs tend to attract such investors. In fact, DRIP firms are more labor-intensive than their no-DRIP counterparts. Computed from data on all 906 pairs, the mean number of employees for DRIP firms is 24,100 while corresponding no-DRIP firms average only 18,490. Statistically, this difference is much too large to be the result of chance.[8]

Previous discussion also suggests that companies in industries subject to relatively high levels of regulation are more likely to offer preferential access to their plans for customers, state residents, and employees. The data confirm that these effects are important. Of the twenty-three companies that offered customers, state residents, or employees preferential access to their plans in 1999, all but one are utilities. Moreover, the other is a financial services company. Given that only 16.7 percent of the DRIP companies in the sample are utilities, this difference is very unlikely to be due to chance, and tests confirm this.

One can make a case from Table 2 that DRIP companies tend to be more mature than those without DRIPs. Mature firms have more assets in place and fewer growth opportunities than younger firms. Older firms also tend to pay higher dividends and carry higher debt levels. Because such firms have fewer growth options, they tend to have higher current earnings but (relatively) lower expected future earnings; consequently, they usually have lower price-earnings ratios and market-to-book ratios.

All of these predictions for mature companies hold for DRIP firms except for the debt ratio, for which there is no significant difference. DRIP firms do, however, pay higher dividends per share and have higher payout ratios. They also tend to have more property, plant, and equipment (assets in place) but make smaller current capital expenditures, a pattern consistent with fewer growth opportunities. DRIP firms have higher net sales and higher profit margins. The evidence regarding price-earnings ratios and market-to-book ratios is mixed but generally supportive of the conjecture that DRIP firms tend to be more mature. In 1999 the mean ratios are higher for companies without DRIPs, but the difference is small enough that it may be due to chance. On balance, the evidence supports the conjecture that DRIP firms tend to be more mature.

The Future

Continuing technological advances, especially if unimpeded by regulatory constraints, are sure to foster the evolution of most financial services, including DRIPs. More DRIP plans are introduced every month, making it easier for investors to diversify as time passes. Another obvious tool for DRIP investors is the Internet. Ford, McDonald's, and Fannie Mae, among others, already let investors use the Internet to service their accounts. The Home Depot, Inc., takes this convenience a step further, permitting investors to buy their first share directly from the company via the Internet.

The Internet has fostered competition for many industries, and brokerage is no different. On-line brokers are now common; in a statement dated January 27, 1999, then-SEC Chairman Arthur Levitt reported that on-line brokers handle about 25 percent of all retail stock trades. Because on-line brokers offer fewer services than traditional brokers, on-line services tend to be cheaper. It seems unlikely, though, that on-line brokers can match the low costs of DRIPs. On-line brokerage accounts typically require a deposit balance, and these can be large. Brown & Company, for example, requires a $15,000 minimum. Such a large minimum balance is unlikely to appeal to new investors, who tend to have smaller accounts.

Broker-run DRIPs provide another evolutionary direction. Competitive pressures have led most major brokerage firms to offer in-house DRIPs. These plans are similar to true DRIPs only in that dividends can be reinvested automatically and only sometimes without brokerage fees. However, the brokerage house usually holds the securities in street name, usually does not credit fractional shares, and charges commissions on optional purchases. This is not to say that these accounts are necessarily inferior to true DRIPs. Rather, brokerage DRIPs provide a different menu of services and costs that may or may not appeal to a given investor.

Conclusion

No one expects direct investment plans to be the answer to all of the modern investor's needs. Mutual funds continue to offer convenience and unmatched diversification for small accounts. Investors seeking to hold individual stocks, whether to compensate for nondiversifiable human capital, to place bets on mispriced securities, or for some other reason, can choose from a rich menu of financial service providers. Traditional brokerage accounts cost more than transactions using DRIPs but offer a wide

8. The numbers are almost identical for the 722 pairs for which data are available on both firms.

range of services that many investors find valuable. On-line brokers offer lower costs but fewer services; such brokers target investors who place less value on the services that a traditional brokerage firm can provide. Direct investment plans, which are concentrated by industry, make diversification more difficult. To offset this disadvantage, they offer a transactions cost advantage; they appeal to the buy-and-hold clientele who seek the lowest possible transactions costs.

Viewed in the broadest sense, all of these methods of distributing securities compete in the same arena for customers' favor. When examined more closely, though, differences become clear. Each offers a different combination of services and costs that appeals to different investors. The key is no different than for any other menu of costs and services: Customers choose the product that offers services that they value and that charges less than the value of those services to them.

What sets direct investment plans apart from the other offerings of financial service providers is a clientele that is well suited for certain companies. A broad, stable ownership base provides benefits to companies that face political or regulatory scrutiny because the company has easy access to many voters. Such shareholders also tend to vote with management; hence, direct investment plans offer potential as a takeover defense. Finally, a broad ownership base provides opportunities for cross-selling, which is more valuable to companies with large-scope economies.

REFERENCES

Asquith, Paul, and David Mullins Jr. 1986. Equity issues and offering dilution. *Journal of Financial Economics* 15: 61–89.

Bogle, John C. 1982. Mutual funds. In *The complete guide to investment opportunities*, edited by Marshall E. Blume and Jack P. Friedman, 509–34. New York: The Free Press.

Carlson, Charles B. 1996. *Buying stocks without a broker*. 2d ed. New York: McGraw-Hill.

———. 1997. *No-load stocks*. Rev. ed. New York: McGraw-Hill.

———. 2000. *Let's talk DRIPs.* Hammond, Ind.: Horizon Publishing Company.

Clemente, C.L. 2000. Pfizer, Inc., letter to shareholders. September 7.

Constantinides, George M. 1979. A note on the suboptimality of dollar-cost averaging as an investment policy. *Journal of Financial and Quantitative Analysis* 14 (June): 443–50.

———. 1984. Optimal stock trading with personal taxes: Implications for prices and the abnormal January returns. *Journal of Financial Economics* 13 (March): 65–89.

CNNMoney. 2001. Fund taxes do matter. February 21. <cnnfn.cnn.com/2001/02/21/mutualfunds/q_funds_taxes_wg/> (February 21, 2001).

Eckbo, B. Espen, and Ronald W. Masulis. 1992. Adverse selection and the rights offer paradox. *Journal of Financial Economics* 32 (December): 293–332.

Ederington, Louis H., and Jeremy C. Goh. 2001. Is a convertible bond call really bad news? *Journal of Business* 74 (July): 459–76.

Finnerty, John D. 1989. New issue dividend reinvestment plans and the cost of equity capital. *Journal of Business Research* 18 (March): 127–39.

Ganci, Paul J. 2001. Remarks presented at the annual meeting of CH Energy Group, Inc., April 24.

Guide to Dividend Reinvestment Plans. 1999. Mamaroneck, N.Y.: Temper of the Times Communications, Inc.

Harris, Lawrence, and Eitan Gurel. 1986. Price and volume effects associated with changes in the S&P 500 list: New evidence for the existence of price pressure. *Journal of Finance* 41 (September): 815–29.

Jacob, Nancy. 1996. Tax-efficient investing: Reduce tax drag, improve asset growth. Dow Jones Publications Library, Trusts & Estates, June.

Levitt, Arthur. 1999. Securities and Exchange Commission statement concerning on-line trading. January 27. <www.sec.gov/news/press/pressarchive/1999/99-9.txt> (January 27, 1999).

Merton, Robert C. 1973. Theory of rational option pricing. *Bell Journal of Economics* 4, no. 1:141–83.

Priory, Richard B. 2001. Duke Energy Corporation letter to shareholders. February 12.

Quattlebaum, Paul, and Otto Strock. 2001. Association of SCANA Corporation Investors letter to shareholders. January.

Scholes, Myron S., and Mark A. Wolfson. 1989. Decentralized investment banking: The case of discount dividend-reinvestment and stock-purchase plans. *Journal of Financial Economics* 24 (September): 7–35.

Smith, Clifford W., Jr. 1986. Investment banking and the capital acquisition process. *Journal of Financial Economics* 15 (January/February): 3–29.

Updegrave, Walter. 2001. What's the best way to invest on a low income? August 14. <www.money.com/money/depts/planning/expert/archive/010814.html> (August 14, 2001).

The Vanguard Group. 2001. A quarter-century of success proves the power of indexing. *In the Vanguard* (Summer):1.

Pension Systems and Aggregate Shocks

KARSTEN JESKE

The author is an economist in the macropolicy section of the Atlanta Fed's research department. *He thanks Thomas Cunningham, Juan Rubio-Ramírez, and Ellis Tallman for helpful comments.*

To some analysts the prospective imbalances between pay-as-you-go pension system benefits and the tax base present a great economic challenge in the decades to come.[1] It is currently estimated that the U.S. Social Security Trust Funds will run out of money in 2041, after which either benefit cuts or major tax hikes would have to occur.[2] Other countries face similar or even tougher and shorter-term problems. For example, in Germany wages reported to the defined benefit system are already taxed at a rate of about 20 percent of gross income, compared to 12.4 percent in the United States, and are expected to rise to a staggering 28 percent by 2035. Sinn (1999) calculates that the current value of all implicitly promised future benefits amounts to a number roughly 250 percent of current German annual gross domestic product (GDP), an overwhelming figure compared to the current government-debt-to-GDP ratio of 60 percent.

In the United States and in most Western democracies, proposals to cut benefits to the extent necessary to save social security are politically infeasible. Raising contributions is—from an economic point of view—undesirable because proportional payroll taxes have distortionary effects on both labor supply and savings decisions. In light of these impending funding problems, politicians and academics alike are calling for a reform of social security. In the United States there is a near consensus that such reform is necessary, but there is also controversy about what the reform should look like. One suggestion, put forward by the President's Commission to Strengthen Social Security, is a partial privatization of social security (President 2001). In this proposal current social security surpluses could be used to fund private, individual retirement account (IRA)-style accounts, and the private savings could make up for future benefit cuts.

In the privatization debate, two opposing perceptions seem to hinder productive discussion. First, social security is considered a low-risk vehicle for the provision of old-age income. Second, the exact size of the social security funding problem is extremely difficult to forecast. Minute changes in the growth assumptions of the forecasting models lead to large changes in the long-term forecasts for social security feasibility. In fact, a few economists argue that with only slightly larger annual growth rates than the conservatively chosen rates used by the Social Security Administration, social security will not face any funding problem whatsoever.[3] For example, Robert Reich, a former trustee of the Social Security Trust Fund, argues that "The actuary's projections are based on the pessimistic assumption that the economy will grow only 1.8 percent annually over the next three decades. Crank the economy up just a bit, to a more realistic 2.2 percent a year, and the fund is nearly flush for the next seventy-five years" (Reich 1998).

Unfortunately, the argument also works in the opposite direction; just slightly smaller growth rates than the ones predicted lead to even more catastrophic scenarios than the ones already discussed. To see how sensitive the trust fund finances are with respect to aggregate variables, one need look only at the calculations of the Social Security Administration. Under the benchmark assumption, the value of the trust fund (in current dollars) will be $6.7 trillion in 2030, and the value decreases until 2041, when the fund is depleted. Increasing the wage growth rate from 1.1 percent to 1.6 percent and the labor force growth rate from 0.2 percent to 0.6 percent would not only almost double the trust fund value in 2030 to more than $12 trillion— almost $7 trillion in today's dollars—but would also ensure that the fund is never depleted over the horizon of eighty years. On the other hand, slightly lower growth rates of 0.6 percent for wages and –0.3 percent for the labor force would cause the trust fund to be depleted by 2029. The implicit confidence interval for the estimated trust fund value in 2030 then covers everything between a slight trust fund liability up to a staggering $7 trillion surplus in today's dollars, more than four times the federal budget for fiscal year 2002.

The results of the model simulations show that, in the long run, privatization makes every generation better off, even if a large aggregate shock occurs.

From an economist's perspective, the two perceptions of low-risk social security and the extreme sensitivity of social security finances with respect to unpredictable economic fundamentals contradict each other. Social security cannot be completely riskless, as the uncertainty about the predicted funding problem demonstrates. The viability of social security crucially depends on what such volatile variables as productivity growth, fertility, and immigration turn out to be over the next decades. From a macroeconomic perspective, a pay-as-you-go (PAYGO) system therefore implies a substantial amount of risk, contrary to the amount that proponents of social security would admit.

Moreover, the factors that tend to drive the performance of a PAYGO system are the same that drive returns on financial market assets, and they push in the same direction. For example, lower productivity growth indeed has a negative effect on the returns to physical capital and therefore reduces financial market returns. At the same time, however, lower productivity growth also jeopardizes a PAYGO system because the system's promised benefits become more difficult to finance if wage growth rates are lower than expected.

PAYGO returns also tend to be lower than financial market returns. If, in addition to this low return, PAYGO also has potentially high risk and a high correlation with financial market returns, then—from the perspective of a Sharpe (1964) and Lintner (1965) capital asset pricing model (CAPM)—a PAYGO system may be viewed as a very undesirable asset. To determine whether and to what degree a PAYGO system is undesirable, one must design a model economy in which aggregate shocks affect not only financial market returns but also a PAYGO system. For policy analysis, this model can help to evaluate proposals. For example, privatization may look unattractive because it exposes the retirement income of a representative generation to a substantial amount of risk. If, at the same time, however, social security faces a symmetric risk profile, then privatization appears more attractive.

This article provides such a model in order to address the sensitivity of different retirement schemes to large aggregate shocks, such as a major drop in productivity growth or demographic shocks like a baby-boomer generation. The workhorse model used in the analysis is a so-called life-cycle economy in which agents work when they are young and their old-age consumption is financed by a combination of a PAYGO pension system and private savings. The model is then used to determine how different retirement schemes perform under different kinds of shocks.

The results of the model simulations show that, in the long run, privatization makes every generation better off, even if a large aggregate shock occurs. The intuition for this result is that, under a privatized pension system, savings are higher, and this higher savings level increases the aggregate capital stock because private savings are more desirable and affordable if both benefits and contributions are lower. This higher capital stock increases welfare by an amount high enough to insulate all future generations even from large aggregate shocks. Aggregate risk is mainly a concern for the period immediately after a social security reform.

A Simple Life-Cycle Model

This section introduces a simple model that allows one to assess how sensitive different retirement systems are with regard to a variety of aggregate

shocks. The model is extremely stylized and abstracts from many real-world matters in order to be analytically and computationally tractable. Many simplifications, however, are performed in such a way as to give a PAYGO system the best possible chance to perform well relative to private retirement accounts. Consequently, the model consistently underestimates gains from privatizing social security; therefore the potential welfare improvements presented here serve as a lower bound.

The model is an overlapping generations (OLG) model; people live for a maximum number of periods, and, to lend the model a greater degree of realism, in each period they have a given probability of dying. That is, in every period there is a distribution of agents of different ages. At the beginning of each period a new generation of young agents enters the economy, and in each of the existing generations a specified fraction of agents dies and leaves the economy.

Just as in the real world, agents have a hump-shaped labor productivity profile; that is, people start with a relatively low labor productivity associated with lower wages and then accumulate human capital until their labor productivity peaks at about the age of fifty, after which productivity slowly decreases until age sixty-five. The model assumes that retirement is mandatory at age sixty-six; that is, productivity (and therefore labor income) falls to zero, and people must live on their private savings and a government-sponsored pension plan thereafter.

In the model simulations in the next section, it is assumed that one period is six years. People enter the labor force at the age of eighteen and may live for up to fifteen periods, that is, to an age of 108; the survival probabilities are matched to the probabilities computed from data from the National Center for Health Statistics. For the computations and the quantitative results presented later in the article, this more sophisticated and realistic multigenerational model is used, but most of the intuition works just fine with a simpler version of the model with only two generations alive at any given time.

In this simpler model, even though the economy has infinitely many periods, individuals live for only exactly two periods; that is, there is no probability that a person will die before reaching the second period. (This assumption will be relaxed in the more sophisticated model used in the numerical examples.) The precise timing is illustrated in Figure 1. In period 1 there is an initial old generation (generation 0) that lives only for exactly one more period and then dies and a young generation (generation 1) that lives for two periods. In period 2, generation 2 is born and serves as the young generation whereas generation 1, the previously young generation, is now the old generation. More generally, in period t there are two generations alive: Generation $t - 1$ is the old generation and generation t is the young. For example, in period 3 generation 2 is old and generation 3 is young.

Introducing social security or even expanding an existing social security system is beneficial only for the initial old and middle-aged generations.

While they are young, people work, receive labor income, and pay a payroll tax to finance the government-sponsored retirement scheme for the old generation. When they are old, people cannot work but must finance their consumption with the government pension and private savings. In the notation used throughout the article, subscripts denote time and superscripts denote the generation's birth period. For example, c_t^t, c_{t+1}^t is consumption of a generation t agent in time periods t and $t + 1$.[4] The budget constraint of this agent in period t takes the form

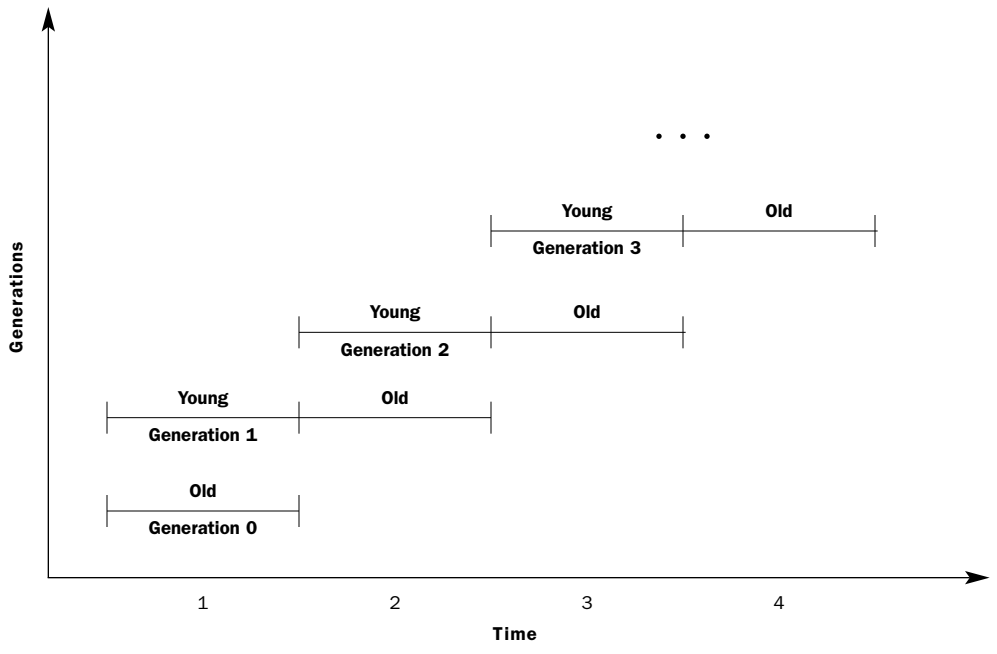
$$(1) \quad c_t^t + k_{t+1}^t = (1 - \tau_t)w_t.$$

The terms on the left-hand side are the expenses of an agent. He can dedicate his income to either consumption, c_t^t, or private savings, k_{t+1}^t, that pay off principal and interest next period. The right-hand side is net labor income, w_t, after paying the proportional

1. It is important to note that the U.S. social security system does not fit the definition of a pay-as-you-go plan in the narrow sense. The current surpluses due to the baby-boomer generation are saved in the Social Security Trust Fund rather than being used to lower contribution rates.
2. The Social Security Trust Fund is actually two funds—the Old Age and Survivors Insurance Trust Fund and the Disability Trust Fund. This article will refer to these two funds as a single fund.
3. Biggs (2000) has the opposite view. He addresses uncertainties about the estimates of a wide range of economic fundamentals underlying the Social Security Administration's computations, such as fertility, longevity, and productivity, and concludes that the estimates tend to be at most reasonable and sometimes on the optimistic side from a historical point of view.
4. This model assumes that agents in each cohort are identical, so there is no within-generation heterogeneity of wealth.

FIGURE 1

Timing in the Overlapping Generations (OLG) Model



payroll tax, τ_t. After working for one period, the agent retires in period $t + 1$ and has two sources of income:

$$(2) \quad c^t_{t+1} = \delta w_t + (1 + r_{t+1})k^t_{t+1}.$$

The first term on the right-hand side is the government-sponsored retirement scheme, promising to pay a fraction δ of an individual's wage as a social security benefit. This social security system is a defined-benefits system because it promises a fixed replacement ratio of δ of the previously earned wage. The second term is private savings plus the earned interest. Agents take wages and interest rates as given and maximize the objective lifetime utility function $U = u(c^t_t) + \beta u(c^t_{t+1})$ subject to the two budget constraints (1) and (2).

On the aggregate level there is both population and productivity growth. It is easy to see that both growth rates play a vital role for the feasibility of social security. The higher the population growth rate, the lower the retiree-to-worker ratio, and the higher the productivity growth rate, the higher the wage growth rate and thus the easier it is to finance promised benefits out of the payroll tax base. Let g_λ denote the population growth rate, which is assumed to be fixed for now, and λ_t, the size of the generation born in period t. Then, by definition, $\lambda_{t+1} = (1+g_\lambda)\lambda_t$. Output ($Y_t$) is produced by combining labor (L_t) and capital (K_t) according to a production function $Y_t = A_t K_t^\alpha L_t^{1-\alpha}$, where A_t is total factor productivity. Market clearing dictates that the amount of labor used in production is equal to the amount of labor from the young generation, and the amount of capital is equal to the amount of savings the old generation accumulated during its working years. That is, output is equal to $Y_t = A_t(\lambda_{t-1}k_t^{t-1})^\alpha \lambda_t^{1-\alpha}$, where productivity, A_t, grows at rate g_A; that is, $A_{t+1} = (1 + g_A)A_t$.

In the model it is assumed that the government balances its budget period by period, and in order to do so it sets the payroll tax so that the tax revenue is exactly equal to the payments to retirees. (Later in the article, when aggregate shocks are considered, this assumption implies that the tax rate adjusts in order to finance the predetermined benefits to the retirees.[5]) The government budget constraint then implies that the payroll tax revenue, $\lambda_t \tau_t w_t$, is equal to the promised benefits to the old generation, $\lambda_{t-1}\delta w_{t-1}$;[6] that is, the equilibrium payroll tax is equal to $\tau_t = \delta(\lambda_{t-1}/\lambda_t)(w_{t-1}/w_t)$. The intuition that supports this result is that with no population or wage growth, benefits are equal to contributions. If there is wage or population growth, then a fixed benefit level can be achieved with lower contributions because the tax base is increasing.

Alternatively, one could assume that contributions into the pension system are constant and that the government distributes the proceeds to the

current retirees. In this case, if there are fluctuations in wages, payroll taxes as a percentage of wages would remain constant and benefits would adjust—the classic example of a defined-contributions pension system. However, benefit cuts as a reaction to aggregate fluctuations are considered politically difficult to implement. Hence, it is more realistic to use the defined-benefits scheme to model social security in the United States because it is a closer approximation to reality.

There are two main simplifying assumptions in this model. First, labor supply is exogenous, that is, agents' provision of labor does not depend on the wage or the payroll tax. In a more sophisticated model where agents decide how much labor to supply, social security financed by payroll taxes has a distortionary effect because people would work less. Ruling out this effect makes a PAYGO system more attractive than in reality, and the gains from privatizing social security would be underestimated in the model relative to the real world.

The second assumption is that the private savings have to be invested in the domestic economy; that is, no international portfolio diversification is possible. Therefore, the effect on financial asset returns is likely to be overstated in the model, again making social security more attractive than in a more realistic, yet currently computationally intractable, model. This article, therefore, makes the best possible case in favor of social security, and the gains from privatization it depicts are likely a lower bound on the actual gains in a more realistic framework.

Balanced Growth Paths

When numerical simulations are performed, initial conditions in the model matter. In particular, the economy starts with an old generation that owns capital, and the researcher therefore has a choice about what the capital endowment of this initial old generation should be. The preferred choice in economics is to begin in a long-run equilibrium and see how a shock affects the economy, in particular what the path back to the long-run equilibrium looks like. Without population and productivity growth, this long-run equilibrium, called steady state, is an equilibrium in which all model variables stay constant over time. Since there is growth in both population and productivity, however, there is evidently no such steady state in this economy. Instead, there is a long-run equilibrium in which the growth rates rather than the levels stay constant for all variables (but may differ across variables). This long-run equilibrium is also called a balanced growth path. Here, this balanced growth path can be computed easily by noting that the tax, τ, must remain constant and the variables in equation (1)—namely, consumption when young, savings, and wages—all must grow exponentially at the same rate. This result, together with the fact that the wage is simply the marginal product of labor, implies that

Social security reform has beneficial effects in the long run, but in the short run a large portion of the population will be worse off.

$$(3)\quad 1+g_k = 1+g_w = (1+g_A)[(1+g_\lambda)(1+g_k)]^{\alpha}(1+g_\lambda)^{-\alpha},$$

and therefore growth rates of wages and per capita savings are given by

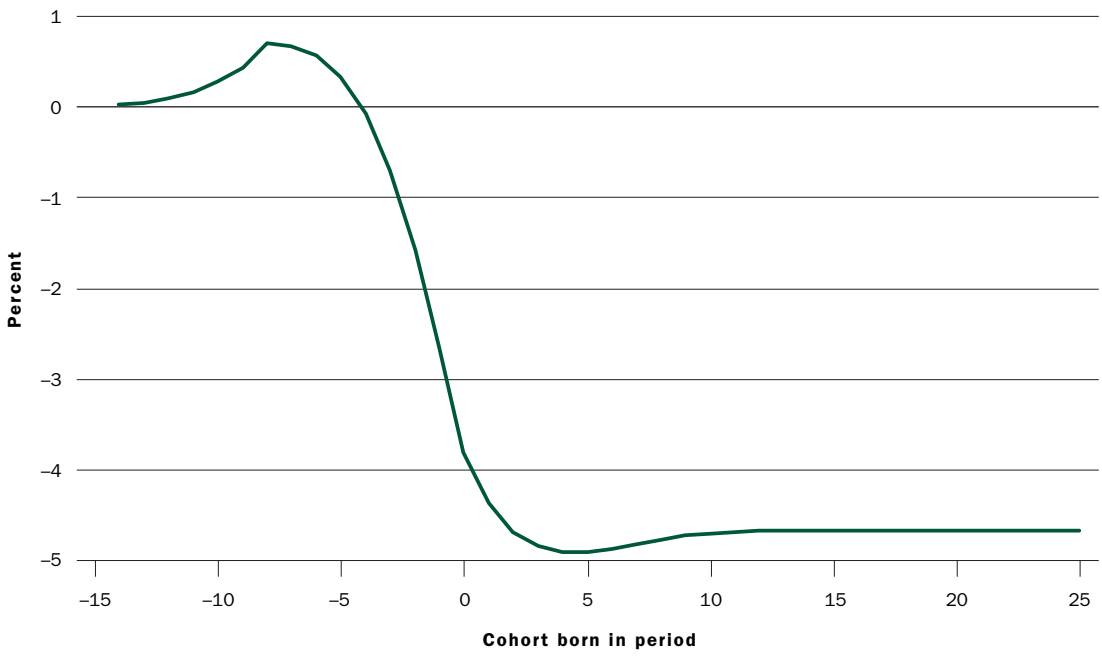
$$(4)\quad 1+g_k = 1+g_w = (1+g_A)^{1/(1-\alpha)}.$$

Equation (4) means that independent of the population growth rate, both wages and savings grow at identical rates equal to approximately $1/(1-\alpha)$ times the productivity growth rate.

5. There are three reasons why this article abstracts from the possibility of a trust fund and instead considers a PAYGO system in which the government adjusts payroll contributions period by period. First, a trust fund is only an option if temporary surpluses are saved to finance future deficits. In two of the three aggregate negative shocks considered here, there are no temporary surpluses and hence no room for a trust fund. Second, in the one case that does display temporary surpluses, namely, the baby-boomer shock, it turns out that the negative welfare effects become even more pronounced with a trust fund; thus, in order to give social security the best possible chance to do well compared to privatized social security, this article assumes there is no trust fund. Third, the bulk of the actual trust fund savings in the United States was accumulated after 1987, well after the first baby boomers entered the labor force, so that the trust fund can all but partially smooth out the future deficits. In other words, because the trust fund build-up occurred so late, the actual U.S. social security system is not far from being a pure PAYGO system.
6. Notice that the promised benefits depend on the lagged wage the current retirees earned when they were young.

FIGURE 2

Increasing Benefits: Average Lifetime Consumption Relative to No Policy Change



The Facts about Social Security

Equipped with the tools from the OLG model, the discussion now turns to some of the facts about pension systems that are well known in the academic literature but possibly less well understood in the popular media.

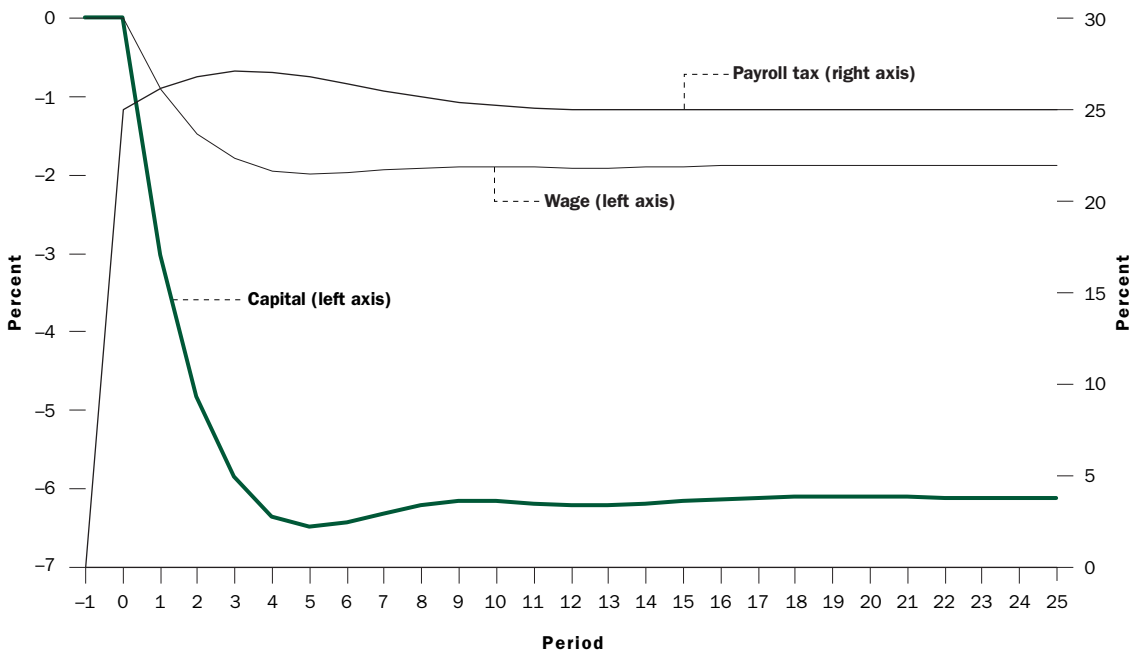
The internal rate of return. Without aggregate fluctuations, the internal rate of return of defined benefit systems for a representative generation is equal to the rate of population growth plus the growth rate of real wages. This result, from Samuelson (1958), can be easily derived with the two-period OLG model. According to the calculations from above, an agent's flow of payments into the social security system is $\delta w_t/[(1 + g_\lambda)(1 + g_w)]$ when young and $-\delta w_t$ when old. Consequently, independent of the value for the replacement ratio, δ, the internal rate of return on this flow is $(1 + g_\lambda)(1 + g_w) - 1$, which is approximately equal to the sum of the growth rates of population and wages. The intuition for this result is that the higher the growth rates for wages and population, and thus the higher the growth rate of the payroll tax base, the easier it is to finance social security. For every dollar of old-age benefits, only $1/[(1 + g_\lambda)(1 + g_w)]$ dollars of contributions have to be paid.

This result holds even if there is a positive probability of death before reaching old age. Suppose a fraction π of agents die after their first period of life so they naturally do not receive old-age benefits. In this case the expected benefits of a representative generation are reduced by π, and at the same time the contributions are cut by the same fraction because fewer old people receive old-age benefits, leaving the internal rate of return unchanged. The results from this stylized economy with only a two-period horizon for every generation carry over to more realistic settings in which agents live longer than two periods.

Financial market returns. In the real world, the potential average yield for social security is equal to population plus wage growth, an amount that is smaller than the average return on private savings. This observation spurred the entire privatization debate because an individual worker could invest her social security contributions in the financial market and, in expected terms, achieve a higher level of benefits. For example, for the period 1960 to 2001, real wage plus population growth averaged less than 3 percent per year, and it is expected to average less than 2 percent per year for the next three decades, according to the Social Security Administration's estimates. Individuals born in 2000 can even expect less than 1 percent return on their social security contributions. On the other hand, the long-term real return on financial assets is significantly higher—in the neighborhood of 5.5 percent annually, according to Feldstein and Ranguelova (2001), or 4.6 percent return from a diversified port-

FIGURE 3

Increasing Benefits in Period 0: Changes Relative to No Reform



folio considered by the President's Commission to Strengthen Social Security.

Introducing or enhancing a social security system. Introducing social security or even expanding an existing social security system is very beneficial for the initial old and middle-aged generations, but the young generation at the time of the program's introduction and all generations born in the future are worse off.[7] In particular, social security crowds out private savings and therefore leads to a reduction in the level of capital. To demonstrate this result, a more realistic version of the model above is used in which generations live for up to fifteen periods. In the simulation it is assumed that in period 0 pensions increase by 25 percent; that is, the benefit formula is adjusted unexpectedly to increase the ratio of pension benefits to lifetime earnings by one quarter.[8] Notice that this policy affects not only generations born in period 0 and thereafter but also all other current workers, namely, generations –7 to –1, as well as current retirees in generations –14 to –8.

The results of the simulation are demonstrated in Figure 2, which plots the average lifetime consumption of all cohorts affected by the policy change.[9] The ten cohorts that entered the labor force between periods –14 and –5 post a gain in average lifetime consumption. These are the seven cohorts currently retired who get higher benefits without ever paying higher contributions and the three cohorts of workers prior to retirement who profit from higher retirement benefits but must pay higher contributions only for a very limited amount of time. All other cohorts, whether currently alive or born in the future, suffer substantial losses of lifetime consumption on the order of about 4.5 percent.

The reason for this result lies in the response of macroeconomic variables to the policy change. Figure 3 plots aggregate capital, wages, and the

7. The scope of this article is social security reform, exactly the opposite of introducing or enhancing social security, but for completeness this well-known result is included.
8. In this version of the model, since people work for a maximum of eight periods, the calculation of benefits is more complicated than in the two-period OLG model, where the benefit formula consisted of only one replacement ratio, δ. Specifically, the more complex model must define how retirement benefits depend on a total of eight past wages and how they develop during the maximum of seven periods in retirement. It is assumed that retirement benefits stay constant (in real terms) during retirement and benefits are a share of the average lifetime earnings.
9. Average consumption in this context is defined as follows. Suppose an individual born in period t consumes a sequence $(c_t^t,\ldots,c_T^t)$; then average consumption $c_{average}^t$ is defined as the constant consumption profile that makes the individual indifferent between the actual and the constant profile. In the two-period example, $u(c_{average}^t) + \beta u(c_{average}^t) = (1 + \beta)u(c_{average}^t) = u(c_t^t) + \beta u(c_{t+1}^t)$.

payroll tax relative to an economy without policy change over time. A higher level of social security triggers a drop in aggregate capital relative to an economy with unchanged social security benefits. The reason for this drop is that the aggregate capital stock is equal to the sum of all savings, and with higher retirement benefits, private savings become both less desirable as people rely more on social security and less affordable as young and middle-aged workers receive smaller net wages because of higher payroll taxes. With the drop in capital comes a drop in wages (making private savings even less affordable), which in turn causes the payroll tax to initially overshoot to more than 25 percent above the initial level, because now the promised benefits to older cohorts must be financed out of a smaller payroll tax base. It is interesting to note that, in the long run, lifetime consumption drops by about 4.5 percent—much more than the 2 percent wage drop. More than half the drop in lifetime consumption is due to the low internal rate of return on the increased payroll tax contributions.

The longer ago privatization took place, the more likely it is that all cohorts alive will be better off under privatized social security. There may be arguments against privatization, but aggregate risk is not one of them.

Theoretically, there is one positive effect for all generations coming from the introduction or expansion of social security. If there are no markets for annuities, private savings have a disadvantage that, with a random lifespan, there is the risk of outliving one's savings. This risk can be eliminated by social security. Since the expansion of social security in this economy substantially reduces welfare in the long run, the risk-sharing effect from social security must be small, however. This effect is in line with the results from Storesletten, Telmer, and Yaron (1999), who show that the negative effect from crowding out private savings is indeed larger than the positive effect from risk sharing.

The effects of social security reform. Social security reform has beneficial effects in the long run, but in the short run a large portion of the population will be worse off. Naturally, a reform can take many forms, but ultimately all reforms involve lower benefits. This article assumes that a reform takes the form of an immediate one-time reduction of benefits. Alternatively, one could assume a delayed or a stepwise reduction, in which case the initial burden of the transition would be spread over more generations. The long-term effects, however, would be unchanged. The result of initial burden and long-term gains is the precise flip side of the previous result as shown in Figure 4. Cutting benefits hurts the currently retired cohorts (–15 through –9) and even some of the older cohorts currently working but close to retirement. In the long run, though, agents are far better off—by about 4.5 percent of average consumption—because lower retirement benefits in conjunction with a lower tax burden encourage more private savings and therefore more capital accumulation. With more productive capital in the economy, wages are higher, and the increase in lifetime income, coupled with the higher returns on private savings compared to the pension system, makes future generations substantially better off.[10] If retirement benefits are cut, however, all retirees and even workers who are near retirement have to suffer substantial losses. Even a delayed reform in which benefits will be lower in the future penalizes those generations that have to pay contributions into the social security system but receive only reduced benefits when they retire.

The fact that private accounts yield higher returns than social security is therefore often misunderstood as a miraculous way of saving the system by offering higher returns from private accounts. It is often ignored that, if such privatization took place, current retirees' benefits—being a burden on current and future tax payers—would have to be cut or would have to be financed through the general tax base or government debt. In other words, the problem of privatization is the unfunded liability to pay for current retirees if current workers start investing in their private accounts.

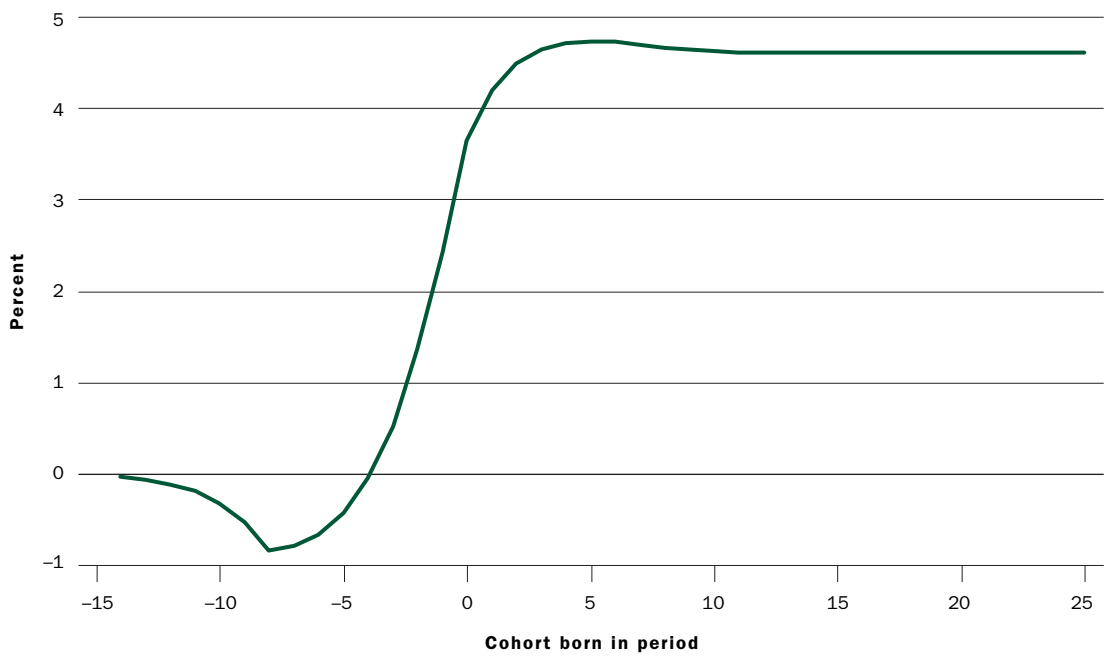
Related Literature

The model presented here is neither new nor the only one that tries to address the issue of aggregate fluctuations and social security. Auerbach and Kotlikoff (1987) set up a similar large-scale OLG model to address a wide array of interesting policy questions, including tax reform and social security reform. The model used here can be thought of as a simplified version of the Auerbach and Kotlikoff model that looks at a new set of experiments, namely, the performance of different pension systems if a large aggregate shock occurs.

De Nardi, Imrohoroglu, and Sargent (1999) do a numerical exercise to study the effects of the demo-

FIGURE 4

Partial Privatization through Decreasing Benefits: Average Lifetime Consumption Relative to No Reform



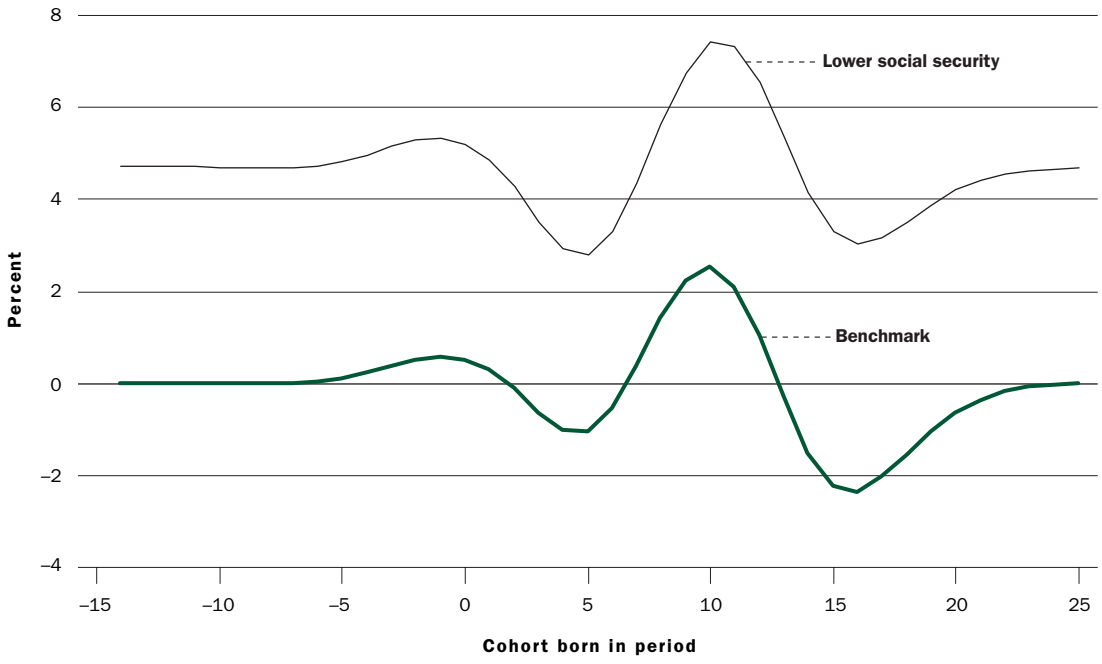
graphic shock on social security that will be caused by retirements among the baby-boomer generation during the coming decades. As one of the main points of the paper, the authors show that, without reform of the public pension system, contributions would have to increase substantially. Higher taxes, however, magnify the distortions in the economy to such a degree that the growth forecasts used by the Social Security Administration seem too optimistic from the perspective of a general equilibrium model because social security taxes discourage agents from both working and saving. In other words, De Nardi, Imrohoroglu, and Sargent quantify a potential feedback effect from higher payroll taxes into the growth forecasts that has been ignored by the Social Security Administration and show that this effect can be substantial.

As mentioned above and demonstrated by Samuelson (1958), introducing a social security system makes a few initial cohorts better off at the cost of making all other future generations worse off. Krueger and Kubler (2002) use an OLG model to test whether this result still holds if aggregate macroeconomic shocks occur—for example, shocks to productivity. Krueger and Kubler point out that it is theoretically possible that all generations may be better off with the introduction of social security, contradicting conventional wisdom. The main ingredients in their model are the assumptions that returns to labor and physical capital are imperfectly correlated and that the social security system is designed as a defined-contribution system. The latter assumption means that the contributions are fixed and benefits vary over time because they are a fixed proportion of the payroll tax base. This setup is quite different from the assumption used in this paper, in which taxes adjust to finance promised benefits, and from the current U.S. social security setup. The welfare-improving

10. It is important to note that, theoretically, it is possible for OLG models to display so-called dynamically inefficient equilibria, as first pointed out by Diamond (1965). The inefficiency involves overaccumulation of savings. In such a case, reducing social security may actually reduce welfare even in the long run because it involves even more overaccumulation of capital. For this reason, Imrohoroglu, Imrohoroglu, and Joines (1995) find in their economy, which displays this inefficiency, social security replacement ratios should optimally be above zero. In the real world this inefficiency, however, appears to be nonexistent because it would involve observing real interest rates on capital that lie below real GDP growth rates, which do not occur in the United States.

FIGURE 5

Baby Boomers in Periods 3–5: Deviation in Average Consumption Relative to No Shock



property of social security in Krueger and Kubler's environment stems from the fact that retirees put a value on the asset called labor because the returns of labor and capital are not perfectly correlated and, therefore, diversification gains are possible. Retirees, however, have no labor endowment left, so giving them a claim to labor income through the social security system improves their welfare.[11] Still, there is the well-known negative effect of social security crowding out private savings both through removing incentives on private savings and the tax distortion, but the former effect theoretically could be larger than the latter. Krueger and Kubler do, however, point out that in an economy modeled to match the U.S. economy the negative effect of crowding out savings dominates the risk diversification.

Computations

Using the more complex version of the model, with fifteen generations alive at any given period and having realistic survival probabilities, this section determines the impact of three different aggregate shocks to the welfare of all current and future generations. Since now one period is only six years, the groups entering the labor market each period will be labeled cohorts rather than generations.

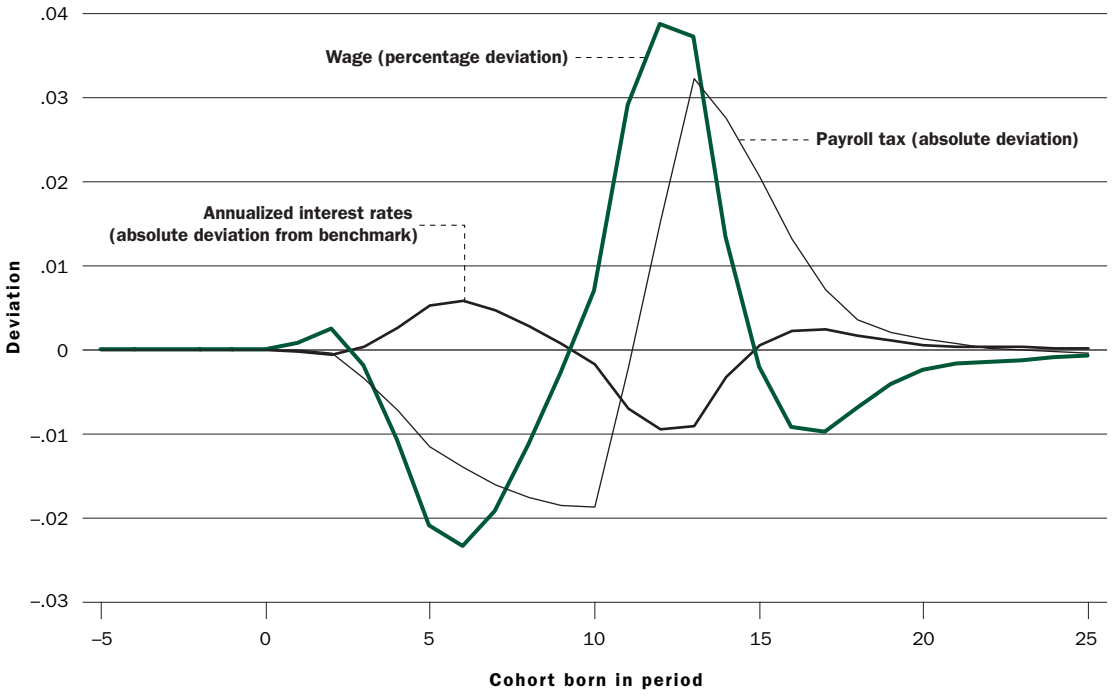
Two sets of experiments are conducted, each of which examines the following three aggregate shocks:

1. A baby-boomer generation—three cohorts that are larger relative to both their parents and children. In the computations using the fifteen-period OLG model, it is assumed that in period 0 it becomes apparent that the cohorts entering the labor force in periods 3–5 are 40 percent larger than in the benchmark economy. The three-period gap (equal to eighteen years) accounts for the years between the birth of a cohort and the time the cohort actually enters the labor market.
2. A permanent drop in the productivity growth rate by one-half starting in period 0, causing the long-run wage growth rate to drop from 1.1 percent—the benchmark growth rate used by the Social Security Administration—to only 0.55 percent.
3. A permanent drop in the labor force growth rate from 0.2 percent per year—the benchmark labor force growth rate used by the Social Security Administration—to –0.3 percent per year, which is the pessimistic demographic scenario used in the calculations of the Social Security Administration. Again, it is assumed, for the same reason as before, that the change takes effect in period 3.

The first set of experiments looks at the three alternative shocks' impact on cohorts' welfare under different payroll tax levels when the economies start down their respective balanced growth paths.

FIGURE 6

Effect of Baby Boomers on Factor Prices and Payroll Taxes



In particular, this analysis looks at one benchmark economy with a benefits structure generating a payroll tax of 12.5 percent, about the same as the payroll tax in effect in the United States, and another economy with both benefits and contributions reduced to three-quarters of the benchmark economy. This second economy is on its long-run balanced growth path and consequently has a higher capital stock than the benchmark economy. One could think of this economy as one that partially privatized social security many periods before and is now on its balanced growth path with lower social security benefits and higher private savings.

The results are plotted in Figures 5–9. In both economies the shocks have an impact on welfare, mostly negative, but people in the economy with the lower payroll tax are uniformly better off; that is, if asked in which economy it would rather live, every cohort would prefer the one with lower payroll taxes.

Specifically, Figure 5 plots the average consumption for both the benchmark economy and the lower-social-security economy relative to the benchmark economy without a shock. In the economy with lower social security coverage, every cohort is better off than its respective cohort in the benchmark economy. Cohort-by-cohort lifetime consumption could lie between 4 percent and 5 percent higher than in the benchmark economy.

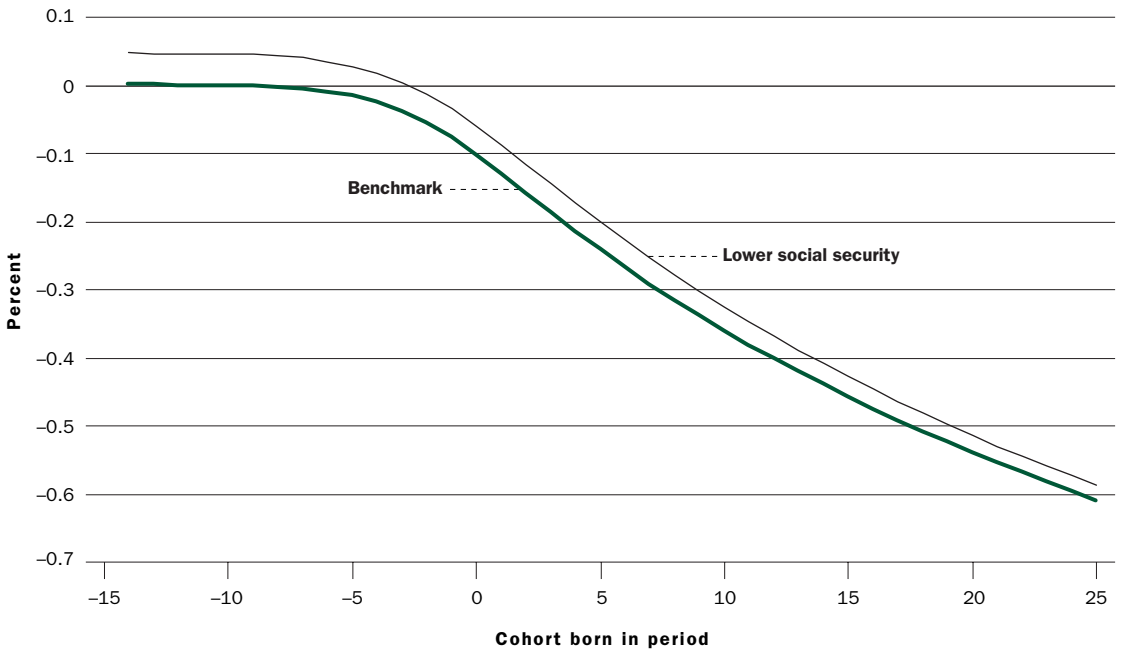
Quite interestingly, in the benchmark economy not all cohorts are worse off because of the arrival of the baby boomers. One could say that the parents of the baby boomers gain about 0.5 percent of average lifetime consumption; the baby boomers themselves lose 1 percent, the children of the boomers gain substantially, more than 2 percent, and the biggest losers are the grandchildren of the baby boomers, who lose more than 2 percent of average lifetime consumption. The intuition for this result comes from the general equilibrium structure of the model, in which factor prices (namely, wages and interest rates) and payroll taxes must adjust to macroeconomic shocks.

Figure 6 plots the response of factor prices and payroll taxes to the arrival of the baby-boomer generation. The baby boomers increase the amount of labor input available in the economy starting in period 3. Therefore, the parents of the boomers benefit because their savings yield higher interest as the amount of labor increases. According to the

11. This outcome, of course, raises the question of why a government has to get into the business of providing risk diversification. If there really is a diversification gain, a private market could do the same job without causing the distortions of private savings.

FIGURE 7

Lower Productivity Growth: Deviation in Average Consumption Relative to No Shock



simulations, interest rates are higher in this economy relative to the no-shock economy until period 9, during which the parents of the baby boomers accumulate the bulk of their savings, and the first couple of periods of their retirement. The baby boomers themselves suffer because the larger supply of labor drives down wages until period 9—for almost as long as cohorts 3–5 work. At the same time, interest rates are lower from period 10 on, the start of the baby boomers' withdrawal phase.

The children of the baby boomers gain substantially because they benefit from two developments. First, when they enter the labor force in periods 8–10, the boomers themselves are still working, driving down the payroll tax until period 11 and thus helping to alleviate the future payroll tax hikes. The baby boomers also drive up the aggregate capital stock through their savings, so when they retire they substantially drive up the wages of their children during periods 10–15, covering most of the working years of generations 8–10 and therefore causing large gains for these cohorts. This result is consistent with Bohn's (1999) for the same reason. The relatively small generation of the baby boomers' children posts a net gain because the factor price effect is larger than the fiscal effect coming from higher payroll taxes. By the time the grandchildren enter the labor force, this positive wage effect has reversed, and payroll taxes are still high, driving down their average consumption.

This result is intriguing because conventional wisdom suggests that the post-baby-boomer generations are worse off because they have to pay high payroll taxes to finance the boomers' retirement. The children of the baby boomers, however, post a net gain because their higher wages make up for the payroll tax hike. The large losers are the baby boomers and their grandchildren if general equilibrium effects are taken into account. Notice also that these results are true in a pure PAYGO system, in which contribution rates adjust period by period. If the government were to start a trust fund at the arrival of the baby-boomer generation, the path of the payroll tax would be smoothed out. This fund would even magnify the welfare effects, in particular for the baby boomers and their children. With a PAYGO system the lower payroll tax during the working years of the baby boomers alleviated the factor price effect. A trust fund would eliminate this alleviating effect, causing the baby boomers to be hit even harder. The children of the baby boomers, on the other hand, who already benefit from the movement in factor prices, would get an additional advantage from the trust fund because part of the burden of financing the baby boomers' retirement would now be financed by the baby boomers themselves.

FIGURE 8

Lower Population Growth: Deviation in Average Consumption Relative to No Shock

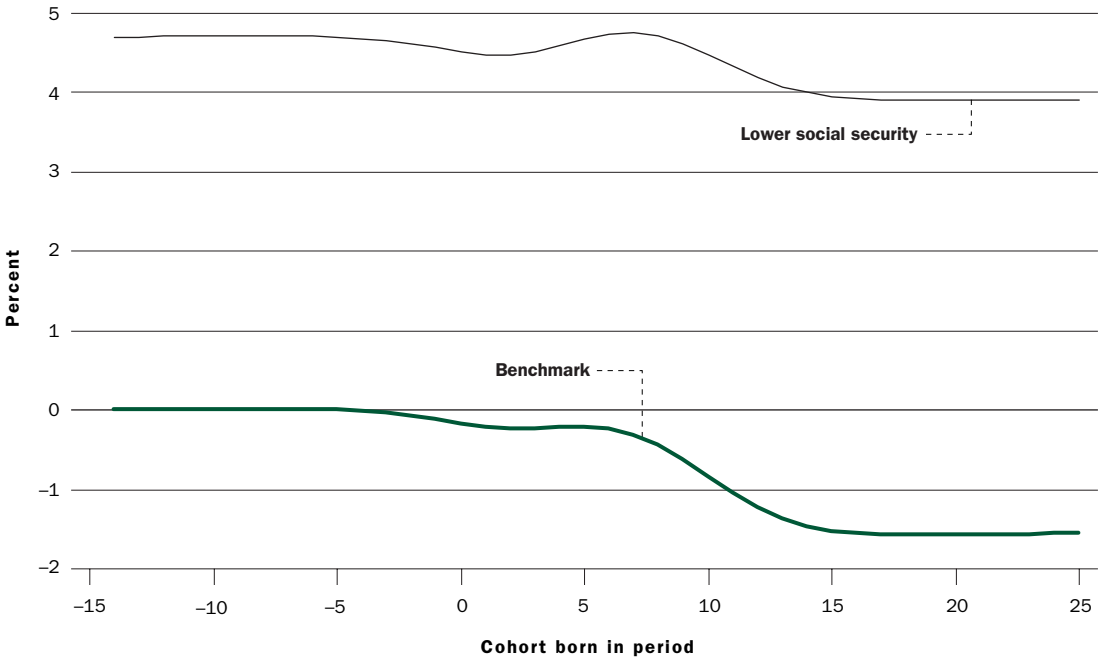


Figure 7 plots the effect of lower productivity growth on cohorts' welfare. All cohorts are uniformly better off in the economy with lower social security. All cohorts are negatively affected by the productivity slowdown, but agents in the economy with lower payroll taxes and higher private savings are shielded better from the adverse effects of the shock.

Finally, in the economy with lower population growth plotted in Figure 8, again every cohort is better off living in an economy with less social security. In addition, the long-run effect in the benchmark economy is more pronounced than in the economy with less social security; cohorts 15–25 lose about 1.6 percent of average consumption in the benchmark economy and only 0.8 percent in the lower-social-security economy. This result comes as no surprise because the long-run rate of return of social security is reduced by exactly the drop in the population growth rate (as demonstrated earlier) while the rate of return on private savings is not affected as much in the long run.[12] Hence, people living in an economy with higher social security levels get penalized more by the adverse demographic shock because their old-age income depends more on social security, an asset whose return is more negatively affected by the demographic shock than private savings are.

Another interesting dimension is the internal rate of return on both private savings and social security in the event of a shock. Figure 9 plots those internal rates of return in the benchmark economy when productivity growth drops in period 0. As expected, the drop in internal rates of returns for the first couple of cohorts is higher for private savings than it is for the social security system. It seems that indeed the social security system provides a better safety net from aggregate shocks than private savings do. However, this result is deceiving because in the long run the drop in the social security yield is higher, about 50 basis points, whereas the yield for private savings recovers for later cohorts and is only about 30 basis points below its original level. In other words, the relative safety of social security for the first couple of cohorts comes at the price of future generations losing a much larger share of their social security yield.

Policy Issues

What is the policy relevance of these results? Evidently, the U.S. economy could not jump

12. The reason is that lower population growth indeed lowers the capital returns because fewer workers are alive, but a large part of the decline is offset by the equilibrium effect of less capital accumulation coming from smaller cohort sizes.

immediately from the regime with high payroll taxes to the one with low payroll taxes and high private savings because a large investment would be necessary to build up the higher capital stock of the latter economy. The analysis shows that, after a partial privatization, in the long run people are better shielded against aggregate shocks. Less social security indeed makes old-age income more sensitive to an aggregate shock, but, since the aggregate capital level is so much higher in the low-payroll-tax economy, this effect is sufficiently cushioned that people are better off with more private savings. Put differently, a concern about a large aggregate shock many years from today would be a reason in favor of privatization, not an objection against it.

Consequently, aggregate risk can be an issue in the privatization debate only over the short horizon right after privatization occurred, that is, before the economy reached its new balanced growth path with higher capital levels. This is the only point at which social security has any chance to beat private accounts on welfare grounds—right after the privatization, when the aggregate capital stock is still low relative to its long-term path and retirees are exposed to the aggregate shock to a higher degree if their benefits are lower.

In the second set of experiments the economy is shocked early on during the privatization, before the new balanced growth path is reached. The three shocks outlined earlier are introduced in each of three new economies, in which the social security reform in the form of a 25 percent reduction of benefits is introduced in periods 0, –5, and –10, respectively. The aggregate shock thus occurs at various stages of the reform, namely, at the same time as the reform, five periods after the reform, and ten periods after the reform, respectively. The welfare effects of the reform combined with the shock are presented in Figures 10–12. The format of these figures is different from that in Figures 5, 7, and 8, which plot the percentage deviation from a no-shock, no-reform economy for the benchmark economy and an economy with lower social security, each of which experienced a shock. Figures 10–12 plot the percentage deviation in the three reform economies with a shock relative to a benchmark economy that also experienced a shock. One could view Figures 10–12 as Figure 4, where an aggregate shock occurred at various stages of the privatization process; that is, a negative number indicates that a cohort is worse off compared to no reform given that a shock occurred, and, vice versa, a positive

FIGURE 9

Internal Rates of Return after a Drop in Productivity Growth in Period 0

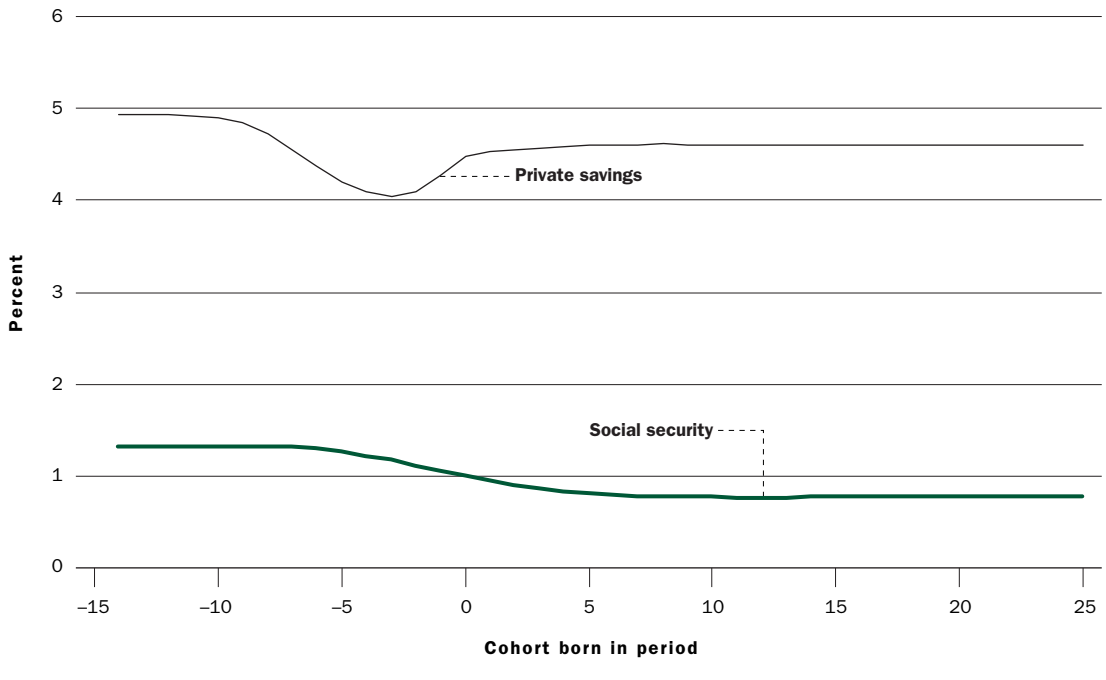


FIGURE 10

Baby Boomers in Periods 3–5: Gain of Partial Privatization

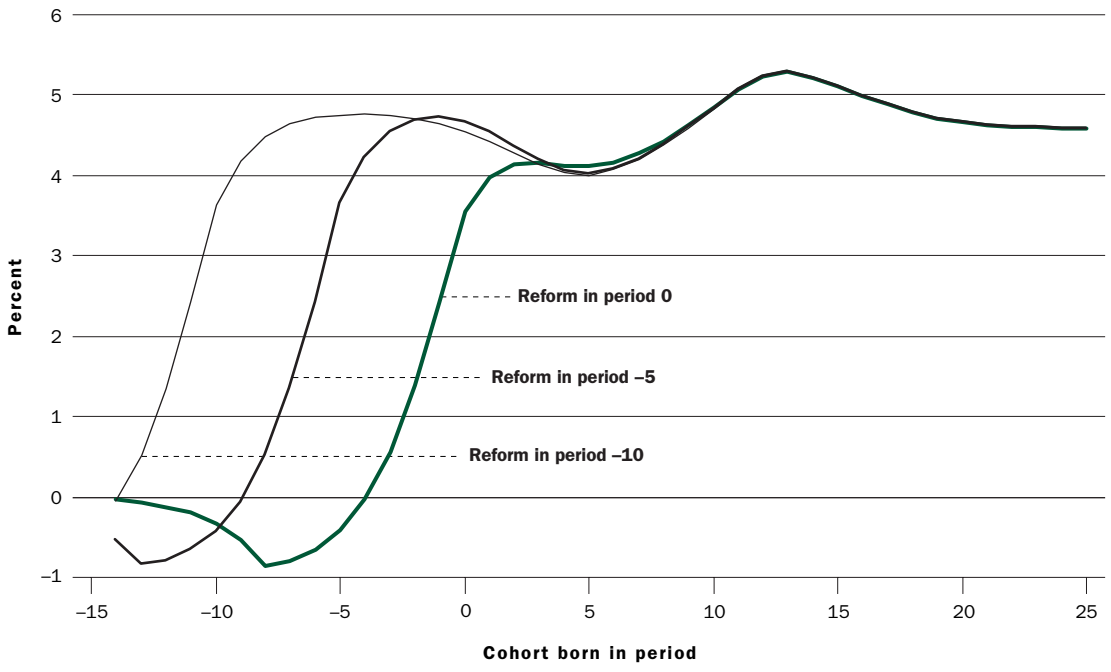


FIGURE 11

Lower Productivity Growth Starting in Period 0: Gain of Partial Privatization

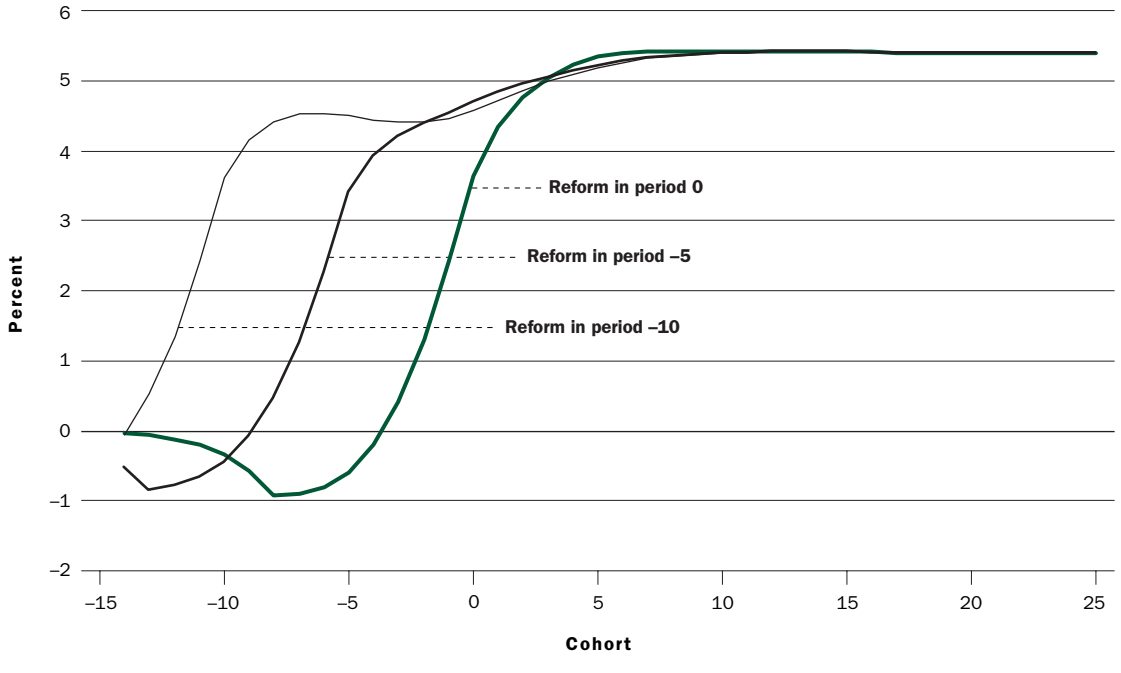
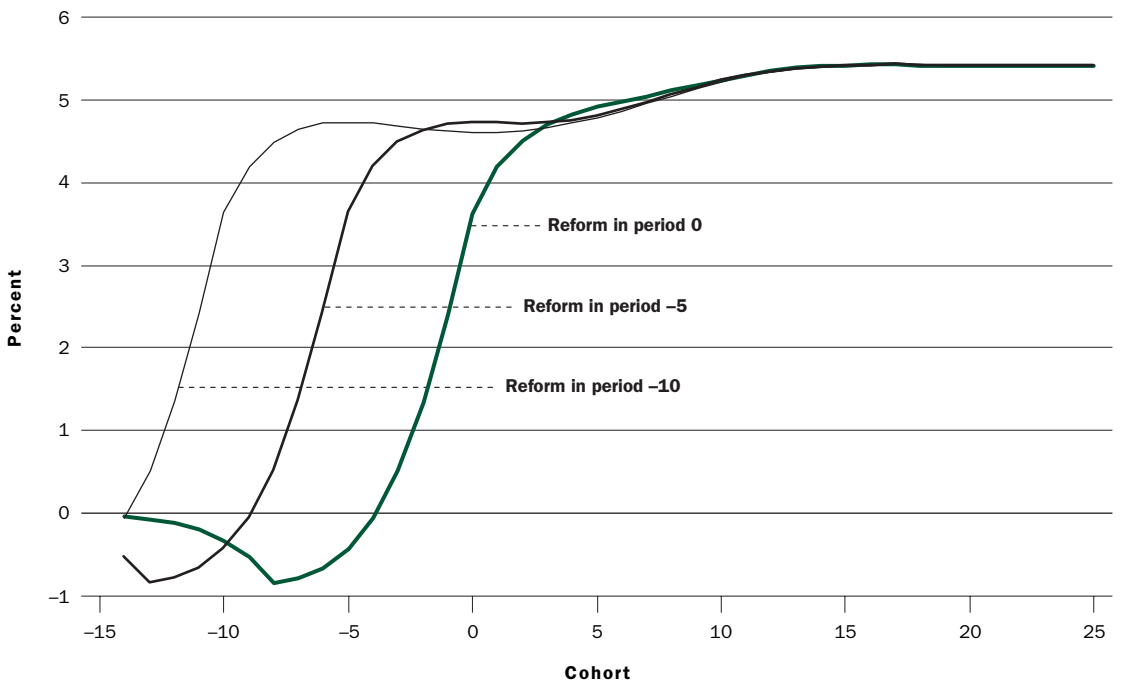


FIGURE 12

Lower Population Growth Starting in Period 3: Gain of Partial Privatization



number indicates a cohort is better off with privatization relative to no reform.

The welfare effects in this set of experiments differ from those in the previous sets. Before, every single cohort was better off with lower social security independent of the shock. Now, there are in fact cohorts that are worse off—precisely the older cohorts in the scenarios in which privatization took place in the current period and five periods ago.

The interesting question is whether the older cohorts are worse off because of the reform or the shock. The effects on consumption in Figures 10–12 look strikingly similar to those in Figure 4; if the reform takes place in period 0, in each case the welfare effects on cohorts –14 to 0 are very similar to the ones in Figure 4. If the reform occured five or ten periods ago, then the curves in Figures 10–12 are close to being shifted by five and ten periods, respectively. Consequently, the welfare effects seem to be due to the reform, not the shock. In other words, once privatization is agreed upon, with its consequences on welfare to the initial old cohorts, adding a shock to the economy will reduce their welfare even further, but, quantitatively, this effect is not much larger than it would have been without the privatization. The benefit of the reform is that future generations, who are normally hardest hit by the aggregate shock, benefit a great deal from the privatization.

Conclusion

Implementing a social security system seems like increasing social welfare out of nothing: One generation of initial old agents never has to pay contributions, but they receive benefits. What looks like the modern version of an alchemist's dream of turning lead into gold is in fact a rather costly endeavor for all future generations because the internal rate of return on their contributions and benefits is almost certainly lower than that of financial assets. There is no free lunch after all! Nevertheless, low returns do not necessarily imply that an asset is unattractive. In a standard Sharpe (1964) and Lintner (1965) CAPM model, the attractiveness of an asset is determined by both the expected return and the variance-covariance structure. People perceive a defined-benefits plan such as social security as a relatively riskless asset.

This article establishes a more cautious view on this matter. Social security is risky as well, and, even worse, it is risky for exactly the same reasons that financial assets are. The determinants of implicit social security internal returns are productivity and labor force growth, precisely the same as those for long-term capital market returns. It is certainly true that in the event of a negative shock, say, to productivity growth, the internal rates of return on the private savings of living cohorts drop by more than the

return on social security contributions. In the long term, however, social security returns take a greater hit than returns on physical capital. In other words, the security in pension system returns is deceiving because it pertains only to current retirees, not to future ones. This seems to be a general feature of PAYGO systems: They look attractive to current cohorts, but future cohorts may suffer substantially.

If an economy with a low payroll tax reaches its balanced growth path, then all people in that economy are better off than the people in an economy with higher payroll taxes if an adverse shock hits the economy. Low payroll taxes encourage more private savings in the form of a larger aggregate capital stock, and this cushion of savings ensures that people are better off than they would otherwise be with a higher payroll tax. This result is true for all three scenarios computed here. If an aggregate shock occurs during the transition to lower benefits, certain generations of current retirees and people about to retire suffer a stronger loss than they otherwise would under an unchanged PAYGO system. However, the effect is rather small compared to the welfare loss coming from the reform.

All future generations, on the other hand, in particular those that normally get hit the hardest by aggregate shocks, benefit greatly from the reform. Their average consumption can be far higher than it would be in an economy without privatization if an adverse shock occurs. Moreover, the longer ago privatization took place, the more likely it is that all cohorts alive will be better off under privatized social security. There may be arguments against privatization, but aggregate risk is not one of them.

REFERENCES

Auerbach, Alan, and Lawrence Kotlikoff. 1987. *Dynamic fiscal policy*. Cambridge, New York, and Melbourne: Cambridge University Press.

Biggs, Andrew. 2000. Social security. Is it a crisis that doesn't exist? The Cato Project on Social Security Privatization. SSP No. 21, October.

Bohn, Henning. 1999. Social security and demographic uncertainty: The risk-sharing properties of alternative policies. National Bureau of Economic Research Working Paper 7030, March.

De Nardi, Mariacristina, Selahattin Imrohoroglu, and Thomas J. Sargent. 1999. Projected U.S. demographics and social security. *Review of Economic Dynamics* 2 (July): 575–615.

Diamond, P.A. 1965. National debt in a neoclassical growth model. *American Economic Review* 55 (December): 1126–50.

Feldstein, Martin, and Elena Ranguelova. 2001. Individual risk in an investment-based social security system. National Bureau of Economic Research Working Paper 8074, January.

Imrohoroglu, Ayse, Selahattin Imrohoroglu, and Douglas H. Joines. 1995. A life-cycle analysis of social security. *Economic Theory* 6 (June): 83–114.

Krueger, Dirk, and Felix Kubler. 2002. Intergenerational risk sharing via social security when financial markets are incomplete. Stanford University manuscript, December.

Lintner, John. 1965. The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47 (February): 13–37.

President's Commission to Strengthen Social Security. 2001. Strengthening social security and creating personal wealth for all Americans. Final report, December. <www.commtostrengthensocsec.gov/reports/final_report.pdf> (December 2002).

Reich, Robert B. 1998. The sham of saving social security. *The American Prospect Online*. <www.prospect.org/columns/reich/rr980600.html> (December 2002).

Samuelson, Paul A. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66 (December): 467–82.

Sharpe, William. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19 (September): 425–42.

Sinn, Hans-Werner. 1999. The crisis in Germany's pension insurance system and how it can be resolved. National Bureau of Economic Research Working Paper 7304, August.

Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron. 1999. The risk-sharing implications of alternative social security arrangements. *Carnegie-Rochester Conference Series on Public Policy* 50 (June): 213–59.

How Much Do We Really Know about Growth and Finance?

PAUL WACHTEL

The author is a professor of economics and the Jules Backman Faculty Fellow at the Stern School of Business at New York University. *This paper was presented at a Federal Reserve Bank of Atlanta conference on finance and growth, November 15, 2002, and at the XI Tor Vergata Conference, University of Rome, December 5, 2002. An earlier version of some of the ideas in this article appears in Wachtel (2001). The author thanks Peter Rousseau for invaluable comments and advice.*

Development economics has changed profoundly in the course of one generation. Twenty-five years ago the emphasis among development economists was on planning and allocation mechanisms, which separated the development community from the core of mainstream market-oriented economics. Academicians who followed development issues were often peripheral to the cutting edge in the economics literature. However, that situation has changed in recent years, and development issues are now at the forefront. As part of this transformation, the term "development" (which connotes a directed process) has been largely replaced by the term "emerging markets." The very term emphasizes the private sector and the market-oriented paradigm of contemporary economics. In no other area is the change in thinking more striking than in the analysis of the role of the financial sector—banks and capital markets—in the development process.

The modern literature on economic growth starts with Robert Solow's work in the mid-1950s, for which he was awarded the Nobel Memorial Prize in Economics. The early theoretical and empirical literature focused on the role of capital and labor resources and the use of technology as the sources of growth. For the most part, any possible role of the financial sector in the growth process was ignored. In fact, development economists up until the 1970s often advocated explicit manipulation of the financial sector in order to achieve development goals. Credit subsidies to favored activities were the rule rather than the exception. Inflation was attractive since a tax on financial assets gave governments with an otherwise weak tax base resources that could be devoted to development projects.

Nevertheless, a few influential economists began to draw attention to the contribution of the financial structure to growth and the benefits of liberalization (in particular, Goldsmith 1969 and McKinnon 1973). Economists slowly acknowledged that credit allocation, interest rates ceilings, and high reserve requirements were undesirable. Generally, high inflation, negative real rates, and inflation taxes create distortions that lead to extensive resource misallocations and discourage saving and the use of intermediaries. The pejorative term "financial repression" was introduced to refer to restrictive policies that inhibited the operation of the financial sector. In 1993 McKinnon could write with confidence that "Now, however, there is widespread agreement that flows of saving and investment should be voluntary and significantly decentralized in an open capital market at close to equilibrium interest rates" (12). However, he characterizes the path toward liberalization as a minefield where one misstep might be the last.

There has been a major shift toward a market-oriented approach to the financial sector over the past twenty-five years. Although capital controls

prevailed around the world in both developed and less developed economies, there have been significant liberalizations in recent years.[1] Today, countries that maintain capital controls are almost self-conscious pariahs in the international community. Liberalization of domestic financial markets has occurred at a somewhat slower pace. Nevertheless, support for directed credit, interest rate ceilings, and government ownership of financial institutions has also disappeared. The prevailing paradigm is that competitive private sector capital markets should be able to gather savings at market rates of interest and allocate capital to the most efficient private sector projects.

There are severe limitations to what we know. The empirical literature has not yet adequately explained what happens when the financial sector deepens and how that deepening affects behavior and economic growth.

The contemporary paradigm hardly needs restatement. Economists now take it for granted that a well-developed, market-oriented financial sector contributes to economic growth. However, it is curious how little solid evidence there is that relates the financial sector to economic growth and stability. The paradigm of financial liberalization was widely accepted before there was evidence to relate it to economic growth.

Only recently, since the early 1990s, has a large body of empirical knowledge accumulated that relates financial sector development—the depth and activity of financial intermediaries—to growth. An impressive array of econometric techniques has been used to show the robustness of the finance-growth relationship. However, it is now time to pause and take stock and ask what this literature has taught us.

This article will briefly describe the approach to assessing the finance-growth relationship that has become virtually standard. The literature provides some important results that relate different dimensions of financial sector development to economic growth. The observed relationships appear convincingly to be causal, from finance to growth, and not an artifact of simultaneity or reverse causality.

However, with all that said, there are severe limitations to what we know. The empirical literature has not yet adequately explained what happens when the financial sector deepens and how that deepening affects behavior and economic growth. There is convincing evidence that countries with money-to-GDP (gross domestic product) ratios of over 100 percent grow more rapidly than those with ratios of 20 percent. However, no good explanation exists of what happens when financial deepening occurs that causes growth. Thus, it is not easy to provide advice to a country with a weakly developed financial sector. The specific mechanisms that relate financial sector deepening to changes in the behavior of economic agents are still a mystery.

Although the finance-growth link is part of the liberal consensus in modern economics, there are still some detractors. Not everyone shares the same degree of confidence in the consensus conclusions. Economists as disparate as Joan Robinson and Robert Lucas have expressed doubts about the link.[2] More importantly, a number of authors have been less enthusiastic about the strength of the empirical consensus. There seem to be differences in temperament on either side of the Atlantic. The Americans (Levine, Barro, myself, and others) exhibit unbounded enthusiasm about the strength of the relationship. The Europeans (Temple and Arestis, among others) are much more cautious and give more emphasis to the variability of the effects and the lack of robustness in some studies. It might well be time to temper some of the enthusiasm with an examination of the skeptics.

There is an interesting analogy to this problem in the short-run macroeconomics literature. Monetarist empirical research in the 1960s and 1970s provided an impressive and convincing body of evidence for the influence of money on inflation and output. The econometric evidence about the direction of causality was convincing, and the description of lags in the effects is widely accepted. However, by the 1990s it was clear that our understanding was limited. We knew that money affected inflation but not *how* money did so; there was a mysterious and unknown "black box" that related money and inflation. Research began to investigate the "transmission mechanism" or the channels of influence that relate money to the economy. Empirical investigations of money and price aggregates are no longer in vogue and have been replaced by efforts to use micro data to illustrate particular channels of transmission.

The finance-growth literature is at the same crossroads. Aggregate investigations will soon be going out of style. In fact, empirical efforts to describe specific channels of interest have already begun to appear.

The discussion will first consider the consensus paradigm and selectively summarize the evidence on the aggregate relationships. The focus then turns to concerns about the strength of the econometric evidence. Finally, the newer developments in the literature—efforts to investigate the finance-growth transmission mechanism with disaggregated data—will be discussed.

Why Are Finance and Growth Related?

The financial sector is important because the financial intermediaries are responsible for resource allocation. Well-working financial intermediaries improve the efficiency of capital allocation, encourage savings, and lead to more capital formation. King and Levine (1993b) were among the first to emphasize that the efficiency-enhancing aspect of financial sector development is more important than the impact on the amount of investment. The financial sector's impact on the allocation of resources cannot be overemphasized. Think of countries with high rates of investment and savings and poor growth experience. The Soviet Union always had high savings rates; there was always an abundance of machinery and equipment, which simply was not allocated to effective uses. Generally speaking, countries with higher investment-to-GDP ratios experience higher growth rates, but the evidence is not overwhelming. The simple correlation of investment ratios and subsequent growth rates was 0.43 in the 1980s and 0.24 in the 1990s.[3] There is substantial variation in growth rates among countries with similar investment ratios. Countries with similar levels of capital investment can have widely diverse growth experiences. The ability to allocate investments efficiently—the role of the financial services industry—might be responsible for the differences.

In the process of providing payments and intermediary services, the financial industry promotes the efficient allocation of resources. There are at least four ways in which the financial sector contributes to growth. They are described in the surveys by Pagano (1993) and Levine (1997) and presented as a rationale for the endogenous growth model in King and Levine (1993b). First, the financial sector improves the screening of fund seekers and the monitoring of the recipients of funds, and these activities improve the allocation of resources. Second, the industry encourages the mobilization of savings by providing attractive instruments and savings vehicles. Such encouragement may also increase the savings rate. Third, economies of scale in financial institutions lower costs of project evaluation and origination and facilitate the monitoring of projects through corporate governance. Finally, financial intermediaries provide opportunities for risk management and liquidity. They promote the development of markets and instruments with attractive characteristics that enable risk sharing.

Broadly speaking, the role of the financial sector in all economies is to channel resources from savers to investment projects. In planned economies, the process is conducted by administrative arrangements with few, if any, market-oriented elements of the financial sector. Emerging market economies will often rely on a single institution—the banking sector—to provide intermediary functions. In contrast, modern economies have a wide range of market-oriented institutions for facilitating intermediation. A successful financial sector will have a broad continuum of financing techniques that channel resources to investment opportunities. The effect of entrepreneurial finance—self financing, informal funding, etc.—on growth is not well explored because there is little data. Nevertheless, the role of venture capital financing is an area of considerable research interest in the United States. More is known about bank financing, and many countries have bank-dominated financial sectors. Capital markets are rudimentary in many countries, including some highly developed ones. There continues to be considerable debate concerning the relative merits of bank-dominated financial sectors and those that give equal weight to capital markets.[4]

1. The International Monetary Fund (IMF) reports large numbers of countries taking measures to liberalize capital flows while the number of tightening measures has declined (IMF 1999, chap. 3).
2. Lucas (1988) suggests that the role of finance is overemphasized, and Robinson (1962, 80) argues that "enterprise leads, finance follows."
3. Average investment-to-GDP ratios for 1979–83 and 1988–92 are compared to growth in 1980–88 and 1989–98, respectively. GDP growth is real per capita; GDP is converted to dollars using purchasing power parity exchange rates and corrected for U.S. inflation. Investment is gross domestic investment. There are eighty-seven countries with available data and a population of at least 2 million. Data are from the World Bank (2000).
4. The differences between Anglo-Saxon bank–dominated and European capital–dominated systems have been diminishing in recent years as a result of globalization and technological and regulatory changes. One of the consequences of European unification is the increased importance of capital markets on the continent. In the United States, regulatory changes virtually allow continental-style universal banking in which banks are involved in the entire spectrum of financing.

The Evidence on Financial Sector Development and Growth

Empirical investigations of the relationship between financial sector development and economic growth began to appear in the 1990s with King and Levine's (1993a, b) cross-country studies for the postwar period and Wachtel and Rousseau's (1995) evidence from long-time series for several countries. These studies showed that the depth of financial sector development and greater provision of financial intermediary services are associated with economic growth. In the decade since those studies appeared, there has been a veritable explosion of empirical interest in the finance-growth relationship.

Broadly speaking, the role of the financial sector in all economies is to channel resources from savers to investment projects.

Furthermore, the research has been extensively surveyed elsewhere starting with Levine (1997) and more recently in Theil (2001).

The first cross-county study of growth and financial development was Goldsmith (1969), which introduced the idea of using a broad measure of the size of financial intermediaries (his specific choice was the value of intermediary assets to GDP) as an indicator of the provision of intermediary services. Looking at decade averages for thirty-five countries for about one hundred years, he found broad indications of a relationship between finance and growth. Goldsmith's work was econometrically unsophisticated and did not seem to spur much research interest at that time. More extensive econometric work was needed to (1) hold constant other determinants of growth and (2) identify the direction of causality.

Barro (1991) and King and Levine (1993a, b) introduced growth studies with cross-country data sets for the postwar period that have become the benchmark for other studies. Their empirical specifications are widely used. King and Levine included measures of intermediary activity developed from IMF and World Bank data sources that are available for a large number of countries. Table 1, reproduced from Levine (1997, 705), shows values for the indicators in 1985 for 116 countries divided into quartiles by real GDP per capita. The four measures of financial sector development are the ratios of

- liquid liabilities of the financial system to GDP,
- bank credit to bank and central bank credit,
- claims on the nonfinancial private sector to total domestic credit, and
- gross claims on the private sector to GDP.

The relationships are clear: Richer countries have more developed intermediaries, and market-based private sector institutions are more important than in poorer countries. Financial intermediary liabilities are over two-thirds of GDP in very rich countries and about half as much in below-median-income countries. Central banks allocate as much credit as commercial banks in countries with below-median income while they are only about one-tenth as large in the very rich countries. Almost three-quarters of credit is extended to the private sector in the richest countries, almost twice the percentage in the poorest countries.

The Standard Empirical Framework

This section presents the regression framework for panel data that has become the standard form.[5] Results from Rousseau and Wachtel (2000, 2001, 2002) are used to illustrate the empirical consensus concerning the relationship between growth and financial depth and to illustrate some of the drawbacks. Econometric investigations with panel data use a regression specification given by

$$X_{it} = \alpha F_{it} + \beta Z_{it} + u_{it}.$$

X_{it} is the growth of per capita real GDP or of the real capital stock or a measure of total factor productivity growth in the ith country for some time period, t. Z_{it} is a standard set of conditioning variables that usually includes the log of initial real GDP per capita (a convergence effect) and the log of the initial secondary school enrollment rate (human capital investment). Additional conditioning variables may include the ratio of government consumption to GDP (measure of private sector activity), the inflation rate, a black market exchange rate premium, or the ratio of exports plus imports to GDP (a measure of openness of the economy), among others. Finally, F_{it} is one of the measures of financial sector development.

There are two econometric problems with regressions of this type. First, there may be simultaneity or reverse causality between the finance variable, F, and economic growth, X. Simply speaking, growing countries might have well-developed

TABLE 1

Aggregate Measures of Financial Development for 116 Countries, 1985

	Very rich	Rich	Poor	Very poor	Correlation with real per capita GDP
Depth	0.67	0.51	0.39	0.26	.51
Bank	0.91	0.73	0.57	0.52	.58
Private	0.71	0.58	0.47	0.37	.51
Privy	0.53	0.31	0.20	0.13	.70
Real GDP per capita (1987 $)	13,053	2,376	754	241	

Note: "Depth" is the ratio of liquid liabilities of the financial system (currency plus demand and interest-bearing accounts of banks and nonbank intermediaries) to GDP. "Bank" is the ratio of bank credit (domestic deposit money banks) to bank credit plus central bank credit. "Private" is claims on the nonfinancial private sector to total domestic credit. "Privy" is gross claims on private sector to GDP.

Source: Derived from Levine (1997)

financial sectors because the income elasticity of the demand for financial services is large. That is, wealthy people demand banking services. Second, the regression specification assumes that any unobserved country-specific effects are part of the error term. Thus, correlation between the error term and included variables in *F* or *X* is likely, which leads to biased estimation of the regression coefficients. Modern econometrics offers a number of approaches to solving these problems.

To deal with simultaneity, researchers have used predetermined (initial) values for the independent variables or instrumental variable estimation. Since the underlying relationship is a long-run one, the time period for observations is often set as a five- or ten-year period. To avoid simultaneity, the independent variables are then measured as the initial (first-year) values of the observation period. For example, if *X* is the average growth rate for 1960–65, then *F* and *Z* are the 1960 values for the respective variables. More recent studies by Levine, Loayza, and Beck (2000) and Rousseau and Wachtel (2000) have introduced the use of instrumental variables to ameliorate the effects of simultaneity between *F* and *X*. Typically, the instruments are initial values of the regressors and perhaps some contemporaneous indicators not included as regressors such as the inflation rate and relative size of the government sector and the degree of openness.

Rousseau and Wachtel (2000) argue that neither of these approaches does an adequate job of solving the simultaneity problem. In that study, the predetermined components of the *F* measures remain correlated with the contemporaneous measures. In addition, the *X* measures tend to be serially correlated. Thus, the techniques described do not remove all doubt of causality from *X* to *F*.

Techniques for examining dynamic interactions among variables have long been available for time series where extensive data series are available. Vector autoregression (VAR) is a widely used technique for looking at causality from lagged *F* to current *X* and vice versa. Wachtel and Rousseau (1995) and Rousseau and Wachtel (1998), among others, have applied VAR to the handful of countries with adequate data for very long periods of time. The results are consistent with the cross-country data analyses for the postwar period.[6]

Panel VARs with a large number of cross-country observations and relatively few time series observations can be estimated with recently developed econometric techniques (see Holtz-Eakin, Newey, and Rosen 1988; Arellano and Bond 1991). Rousseau and Wachtel (2000) implement the technique to estimate panel VARs with annual data and develop Granger causality tests. Beck, Levine, and Loayza (2000) and Levine, Loayza, and Beck (2000) also find that measures of financial sector development have a significant causal effect on growth in panel VAR estimates.

The second econometric problem noted above was the estimation bias introduced in any panel estimation from unobserved country-specific influences. One way of dealing with this is to include

5. There is some literature that utilizes somewhat different frameworks to address some of the same issues, such as the work done for the Organisation for Economic Co-operation and Development (OECD) growth project (see Leahy et al. 2001) and Graff and Karmann (2001).
6. Both Arestis and Demetriades (1997) and Rousseau (2002) compare time series and cross-section approaches. Arestis is skeptical of cross-country results because of the differences among countries in time series results. Rousseau finds the different approaches to be consistent.

country-fixed effects (dummy variables) in all estimated equations. However, the colinearity between the fixed effects and the phenomenon under investigation leads to very imprecise and unstable coefficient estimates. A measure of the financial structure such as the ratio of credit to GDP varies considerably among countries but changes slowly over time in any given country. Thus, the country-fixed effects explain much of the panel variation in the financial structure variable. The sensitivity of the standard specification to the inclusion of country-fixed effects will be demonstrated below. Although many econometricians would argue in favor of such country-fixed effects, most analysts reject this approach or the simple solution of differencing the data on practical grounds. However, the Arellano-Bond estimator ameliorates the country-specific effects by differencing a VAR specification in levels of the data and leads to better estimates.

Countries with better creditor rights, rigorous enforcement, and better accounting information tend to have more highly developed financial intermediaries.

A Summary of the Evidence on Financial Depth and Growth

Despite the formidable econometric problems, a wide body of literature has firmly established a consensus in support of a relationship between financial sector development and economic growth. Several studies by Rousseau and Wachtel will illustrate both the approaches taken and the results established.

Rousseau and Wachtel (2000) examine the ratio of the broad money supply to GDP with panel data that include two eight-year average observations for forty-seven countries. Similarly, Rousseau and Wachtel (2001) use seven five-year averages (1960–95) for eighty-four countries. These studies present results with panel data sets using instrumental variables. The first paper also presents panel VAR models with forty-seven countries and sixteen annual observations, estimated with an application of the Arellano and Bond procedures.

The ratio of broad money to GDP averages about 40 percent; it is larger in countries where the depository institutions are more actively intermediating between savers and investors, and it is smaller where the banks do little more than provide transactions services. The Rousseau and Wachtel results indicate that an exogenous increase in the ratio of 10 percentage points (increasing the activity and depth of the depository institutions) will, particularly in countries without high inflation, increase the rate of growth by between 0.6 and 1 percentage point a year. Over a five-year period, real output would be between 3 and 5 percent higher.

To address the issue of causality more directly, we estimate VAR systems with the same data using the Arellano and Bond approach. We find evidence of significant causality from financial measures to real GDP and no evidence of feedback from GDP to the financial variables. These estimates indicate that an increase in M3 that raises its average share in output by 10 percentage points would raise output per capita over five years by 4.1 percent, or 0.8 percent per year. Interestingly, the results from the two approaches—panel regressions and panel VAR—are remarkably alike.

A change in the ratio of M3/GDP of 10 percentage points is quite large. For any given country, the ratio is serially correlated and trends occur slowly. However, there is a great deal of variation among countries at different stages of financial development, and at any given time the distribution of the ratio across countries is quite diffuse. In 1987, the ratio is less than 40 percent in 38 percent of the countries, between 40 and 60 percent in 34 percent of the countries, and over 60 percent in 38 percent.[7] Thus, an increase of 10 percentage points is not unreasonable for a country experiencing financial sector deepening. Both the VAR and panel results indicate that such a change would have profound effects on growth.

The results in Beck, Levine, and Loayza (2000), which extend Levine's earlier work and also introduce panel estimation, are very similar to those in Rousseau and Wachtel (2000). This paper introduces an improved measure of financial sector development—the ratio to GDP of credits from financial intermediaries to the private sector from a World Bank data set. This measure excludes credits from the central bank and government and credits among financial intermediaries. The researchers estimate a variant of the now-standard specification with data for seventy-seven countries for 1960–95 in two ways. First, they estimate a cross-section regression with instrumental variables (using thirty-five-year average data). Second, they estimate a panel of five-year averages using the Blundell and Bond (1998) modification of the Arellano and Bond

TABLE 2

Equity Markets, Financial Depth, and Growth: Summary of Panel Regression and VAR Estimates

	Country mean		Effect on growth rate of a 10 percentage point increase (five-year horizon)	
Ratio to GDP of	1987	1995	Panel regression	VAR model
Liquid liabilities (M3)	58.73		0.15	0.8
Market capitalization	29.12	65.11	0.08	0.4
Total value traded	10.75	24.22	0.52	1.0

Source: Calculated from Rousseau and Wachtel (2000)

technique called the systems estimator, which allows information in the levels of the variables to be retained in the procedure rather than be swept away through differencing.

When initial income and average years of schooling are the only conditioning variables, both estimation procedures give very similar results. An increase of the private credit-to-GDP ratio of 10 percentage points from its mean of 27.5 percent results in an increase in the annual growth rate of 0.69 percent with the cross-section and 0.74 percent with the panel. When a broader set of conditioning variables is used, the estimates vary between 0.5 and 1 percent.

The Role of Equity Markets

Equity markets are always of interest because data on equity market activity around the world are available and because the stock market—Wall Street—always attracts attention as the paramount symbol of capitalism. Studies of the finance-growth relationship with aggregate credit measures were quickly followed by studies of the influence of the equity market on growth.

Banks dominate financing in many places and even in the most advanced industrialized countries; equity markets are only a small part of the overall financial markets. Most new investment is funded either internally by firms, through banks and other intermediaries, or directly through bond markets. New issuance of stock is never a large fraction of total sources of funds. Nevertheless, the existence of a stock market is important even when equity issuance is a relatively minor source of funds.

Why is the existence of a stock market so important? First, an equity market provides investors and entrepreneurs with a potential exit mechanism. Second, capital inflows—both foreign direct investment and portfolio investments—are potentially important sources of investment funds for emerging market and transition economies. Third, the provision of liquidity through organized exchanges encourages both international and domestic investors to transfer their surpluses from short-term assets to the long-term capital market, where the funds can provide access to permanent capital for firms to finance large, indivisible projects that enjoy substantive scale economies. Fourth, the existence of a stock market provides important information that improves the efficiency of financial intermediation generally. Finally, the valuation of company assets by the stock market provides benchmarks for the value of business assets, which can be helpful to other businesses and investors, thereby improving the depth and efficiency of company assets generally.

Atje and Jovanovic (1993) construct a cross-country panel for the 1980s and show that trading volume has a strong influence on growth after controlling for lagged investment while bank credit does not. Demirguc-Kunt and Levine (1996) provide a descriptive investigation. Levine and Zervos (1996, 1998) introduce equity market measures to the standard growth-finance cross-section specifications discussed earlier. Finally, a more comprehensive effort to examine the dynamic relationships is found in Rousseau and Wachtel (2000).

The Rousseau and Wachtel paper uses two measures of stock market development as financial sector indicators in the panel regressions: the ratio of market capitalization to GDP and the ratio of total value traded to GDP. Both have a positive coefficient, but only the latter is significant at the 1 percent level. The study also uses a VAR model to examine causality and dynamic interactions among growth, a measure of financial intermediation, and a stock market indicator. Table 2 summarizes the results of panel equations with alternative measures of financial sector development.

7. This result is based on the sample of forty-six countries with active equity markets.

The results indicate that the development of a liquid and highly capitalized equity market increases growth. The mean ratio of value traded to GDP was just 10 percent in 1987; the panel regression results indicate that an increase in the ratio of 10 percentage points would add 0.5 percent to the growth rate. Similarly, a 10 percentage point increase in the ratio of M3 to GDP (with a 1987 mean of 59 percent) would increase the growth rate by 0.15 percent. The equity market effects are similar in magnitude to the effect of more developed financial intermediaries.

Other Financial Sector Characteristics

Research efforts so far have not examined the impact of other financial markets or instruments on economic growth in a similar cross-country framework. A major reason for this dearth of research is that data on other types of financial intermediaries (for example, private placements, venture capital, bond issuance, commercial paper, etc.) are not part of any standardized data collection efforts and are often simply not available. Furthermore, the number of countries with these other instruments and markets is not large. Although banks and related intermediaries are found everywhere and equity markets are found in most places, bond markets, commercial paper, organized venture capital industry, and so on are quite rare.

There are systematic differences in the finance-growth relationship among countries with different characteristics. For example, the evidence of finance effects is not as strong among developed countries as it is among less developed countries.

There is a body of work that focuses on the relationship between economic growth and the quality of the financial sector environment. For example, important elements of this environment that might effect growth include clear and universally applied accounting standards and auditing practices and a legal framework for debtor-creditor relationships. The effect of accounting, bankruptcy, and governance standards and procedures on growth and on financial sector development has been recently examined with the standard cross-country framework by Levine, Loayza, and Beck (2000). Among other things, they find that countries with better creditor rights, rigorous enforcement, and better accounting information tend to have more highly developed financial intermediaries. Thus, growth prospects are enhanced because a sound legal environment encourages the development of financial intermediation.

The analysis has already shown that "more banking"—a larger ratio of bank liabilities to GDP—is an important correlate of economic growth. Further investigation examines the type of banking activity, the environment in which it is conducted, and by whom it is conducted. Results indicate that the following banking industry characteristics may be related to growth and stability:

- more competitive and less concentrated banking industry,
- more private as opposed to government ownership or control, and
- more foreign participation in banking.

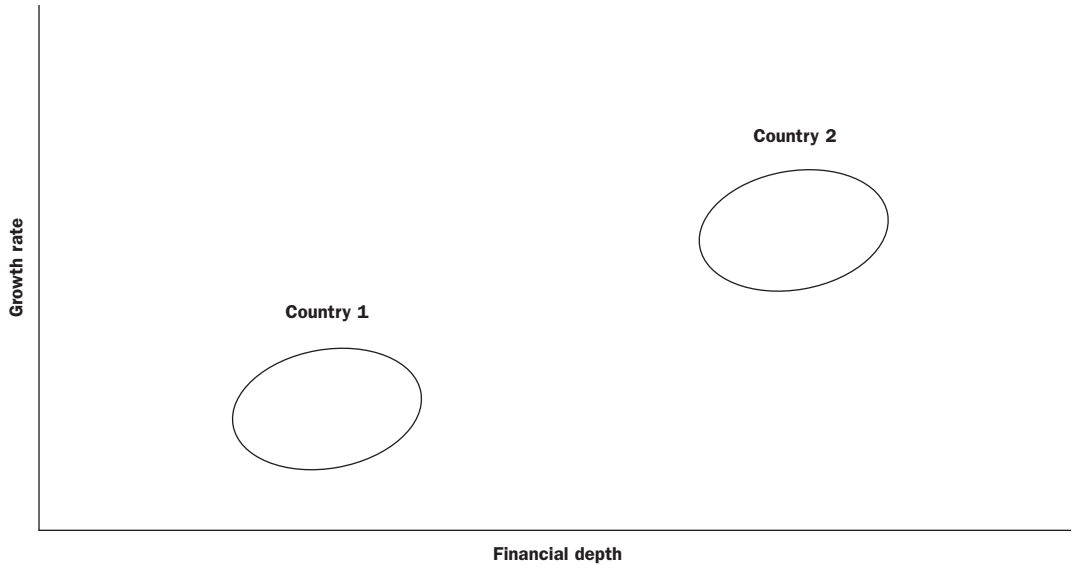
For example, La Porta, Lopez-de-Silanes, and Shleifer (2002) examine the effect of bank ownership on economic growth with the standard panel framework introduced earlier. They consistently find that higher initial government bank ownership has a negative impact on real per capita growth rates. A 10 percentage point increase in the proportions of assets of the largest banks owned by the government is associated with a decline in the annual growth rate of about 0.2 percent. These preliminary regressions do not address all of the econometric problems, but the overall thrust of these results will probably withstand a more careful empirical investigation.

Several recent papers relate the legal environment for the financial sector to economic growth. Part of the motivation for these inquiries is econometric. The origins of the legal system (for example, English common law or French civil law) are a completely exogenous variable determined by accidents of history (and colonialism). However, the legal systems have different approaches to creditor-debtor relationships that could be relevant to the performance of the financial system and, thus, economic growth (La Porta et al. 1998; Levine 1999). The exogenous characteristics (legal origins) can be used as instruments to improve econometric estimates of the basic finance-growth relationships.

A related issue addressed by Levine (2002) is whether bank-dominated (the German model) or market-dominated (the Anglo-Saxon model) financial systems generate better growth performances. He finds that the quantity of financial services is more important than the structure of the industry that provides them. Convergence of financial systems around the world will probably make this specific question moot over time.

FIGURE 1

Finance and Growth: Hypothetical Data



Drawbacks of the Standard Approach

The standard results seem to be very robust. The papers by Rousseau and Wachtel are consistent across techniques and data sets and are also consistent with the large body of work by Levine and various coauthors. Moreover, the results that relate growth to equity markets, banking sector structure, and the characteristics of the financial system strengthen the conclusions. Nevertheless, not everyone is convinced by these results. Although I think that the research results are convincing, there are still issues to look at and concerns to note. We should hesitate to declare victory.

Specifically, there are two questions I would like to pose. The first is whether the standard approach does an adequate job in controlling for country-specific effects. The second is whether the estimates of finance effects are robust or vary with other observed phenomena. These questions have come up before in regard to the growth literature in general (Temple 1999; Durlauf 2001; Kenny and Williams 2001). These authors argue that since the relationship between growth theory and empirical specifications is often tenuous, it is not surprising that many empirical results are sensitive to changes in specification.

My concern about the adequacy of efforts to hold country-specific effects constant is illustrated in Figure 1. If observations for growth and financial sector development are clustered by country, as shown in the figure, panel regressions could indicate a spurious aggregate relationship. The observed finance-growth relationship is due to between-country differences rather than within-country differences over time. In this case, regression results would not provide any reason to make inferences about the effects of financial deepening on growth.

This issue is further investigated with the regressions shown in Table 3. A standard panel specification is shown (with the panel data set from Rousseau and Wachtel 2001). The first equation is estimated by ordinary least squares (OLS), and the independent variables are all initial values (value for the first year of each five-year period). Estimates are indistinguishable from the second equation that uses contemporaneous values for the government and liquid liabilities variables and estimates the equation with instrumental variables. The choice of technique to correct for simultaneity is immaterial. Simultaneity bias does not seem to be an issue.

However, both of these equations include fixed effects for time periods but not for countries. The equation in the last column adds country-fixed effects to the equation. The introduction of country-fixed effects has a profound effect on the results. The fixed effects dominate the equation; the proportion of variance explained almost doubles, and some of the coefficients have the wrong sign. The finance effect is still positive, but the coefficient is very small and barely one-tenth of a standard error from zero. Figure 2 shows the strong relationship between the fixed effect coefficients and the average ratio of liquid liabilities to GDP. The between-country

TABLE 3

Panel Estimates for Five-Year Average Real per Capita GDP Growth

	OLS with initial values	Instrumental variables	OLS with initial values and country-fixed effects
Constant	–0.726 (1.0)	–0.743 (1.0)	
Log of initial real GDP	–0.203 (1.5)	–0.199 (1.5)	–3.447 (5.4)
Log initial secondary school enrollment	0.841 (3.7)	0.819 (3.7)	–1.715 (3.7)
Government expenditure to GDP	–0.060 (2.6)	–0.063 (2.5)	–0.081 (2.3)
Liquid liabilities to GDP	0.027 (4.7)	0.028 (5.0)	0.001 (0.1)
Fixed effects	Time periods	Time periods	Time periods and countries
Corrected R^2	.233	.247	.440

Note: Absolute values of *t*-statistics are shown in parentheses.

Source: Panel with 426 observations from Rousseau and Wachtel (2001) for 80 countries, 1960–95.

differences in the finance ratios are more important than the differences over time, and thus the fixed country effects and the finance ratios convey largely the same information. Although financial depth measures exhibit much short-run or cyclical volatility, development of financial systems evolves slowly. Data that span less than forty years may not reflect much long-run change in the financial system.

The devastating impact of fixed (country) effects on the estimates of a growth equation has been shown with a different panel specification by Benhabib and Spiegel (2000). They also show that adding fixed effects leads to coefficient instability and a loss of significance on the financial depth measures. Although they recognize this result, they seem reluctant to question the popular consensus that finance matters.

Proponents of the standard growth rate equation would argue that the specification does not call for country-fixed effects. The equation is derived from a production function relationship, so the country-specific unobserved effects disappear with the differencing. But the fact that they enter the equation significantly suggests that the country effects persist. It appears that the standard set of regressors does not provide an adequate framework for making inferences about the change in financial depth on growth from cross-country comparisons.

As noted earlier, there are some skeptics in the growth literature, mostly Europeans who are worried about a possible lack of robustness among empirical results. Kenny and Williams (2001) provide a scathing critique of the empirical growth literature (without any reference to the role of finance). In their view there is little consensus or robustness and most models are overly simple. A formal econometric investigation of robustness issues is found in Florax, de Groot, and Heijungs (2002).

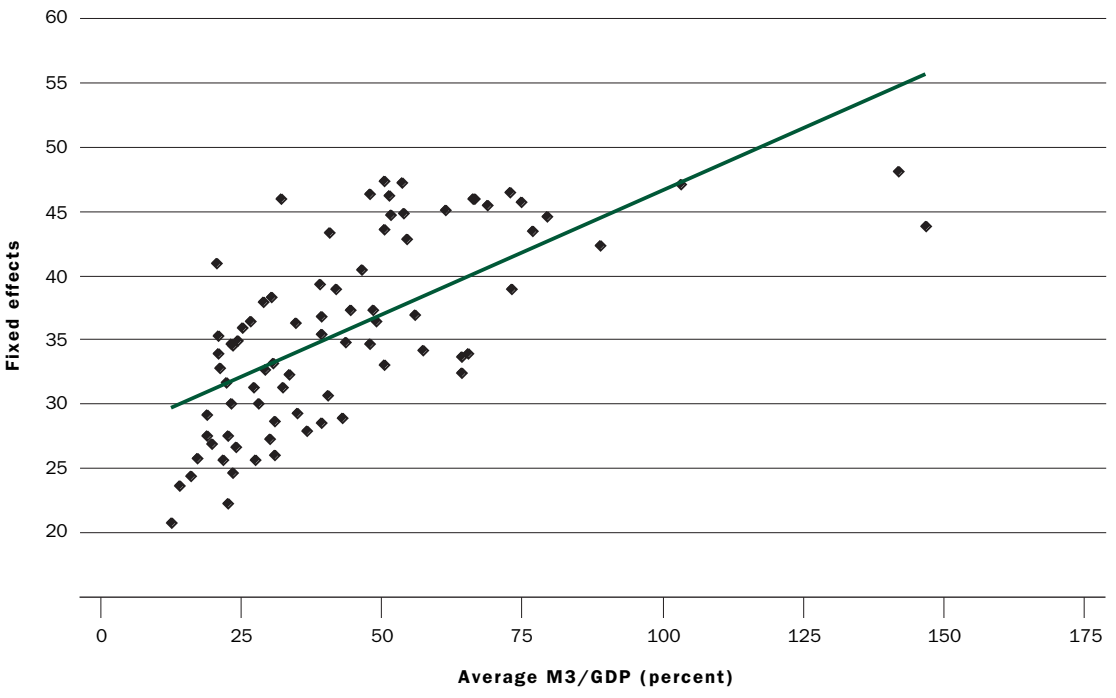
However, this issue highlights the importance of the recent papers with panel VAR estimates that remove the country-fixed effects by differencing and exploit the time series variation more fully. Nevertheless, there are several papers that are concerned with the robustness of VAR results. For example, Luintel and Khan (1999) find some evidence of bidirectional causality between financial sector development and growth in a VAR analysis of developing countries. Similar problems are noted by Shan, Morris, and Sun (2001) in VAR analyses of the OECD countries.

There are systematic differences in the finance-growth relationship among countries with different characteristics. For example, the evidence of finance effects is not as strong among developed countries as it is among less developed countries. In addition, the finance effect varies systematically with a country's inflation experiences (Rousseau and Wachtel 2001, 2002). These two studies find that the impact of financial deepening on growth disappears when inflation is high. This result would not be surprising with hyperinflation that erodes the value of financial intermediation. However, the results indicate that above a threshold inflation rate between 13 and 25 percent, financial deepening ceases to increase economic growth.

Estimation issues aside, there are at least two reasons why the consensus model is only the first stage of an important research agenda. First, even the refined measure of financial depth introduced by Levine, Loayza, and Beck provides a highly aggregated picture. There is wide variation in these financial sector ratios that is hard to understand. For example, the 1987 ratio of M3 to GDP is 73 percent in Spain and 51 percent in Sweden. Does this difference reflect more advanced financial sector development in Spain or greater reliance on bank-based financing? Second,

FIGURE 2

Fixed Effects and Average Ratio of M3 to GDP for Eighty Countries



Source: Calculated from the regression in the last column of Table 3.

a thrust of the earlier discussion was the variety of financial sector institutions and activities that contribute to efficient intermediation. The aggregate measures mask a rich and diverse set of activities and reveal little about how intermediation affects growth.

The Next Stage

Return for a moment to the analogy with the macro literature on monetary policy effects. The St. Louis model developed in the late 1960s was a standard reduced form that related money growth to output growth and inflation. Later research debated the stability and robustness of the relationship. Today hardly anyone pays attention to the St. Louis model specification. However, it played an important role in the development of monetary economics. Its reliability and usefulness aside, it established the consensus view of the impact of monetary shocks on the economy and set the scene for the next generation of research, which looks inside the black box and tries to explain the transmission mechanism for monetary policy.

The finance-growth empirical literature is in the midst of a similar development. The standard reduced-form equations might not be as robust as originally thought, and their predictive value for explaining the effects of financial deepening is limited. However, the research agenda of the 1990s firmly established the consensus view that finance matters and set the scene for the nest stage of research. Now it is time to look into the black box and develop empirical studies that shed light on the way in which financial sector development improves intermediation and generates economic growth.

The next stage has already begun with a few studies that exploit industry data to better understand how financial sector development works. Rajan and Zingales (1998) were among the first to exploit industry data to gain an understanding about the finance-growth relationship. A well-developed financial system removes or reduces the barriers to external financing for firms. Moreover, some industries tend to depend on external financing more than others because of differences in cash flow patterns, capital intensity, profit margins, and so forth. As a consequence, industries that are more dependent on external financing should do better in countries with better financial systems. Industry data for a number of countries gives Rajan and Zingales the opportunity to test this hypothesis. They examine data for forty-one countries during the 1980s. Their results support the hypothesis.

The innovative use of industry data opens the door toward more specific analysis of finance effects

on growth. The Rajan and Zingales paper is important for this reason although it makes a number of rigid assumptions. In particular, it uses U.S. experience to determine which industries are heavy users of external finance and assumes that these patterns hold elsewhere. Fisman and Love (2002) take issue with this assumption and provide a different interpretation. They are concerned that the Rajan and Zingales results imply that countries with poorly developed financial markets should concentrate on industries that rely on internal financing. Instead Fisman and Love provide support for the hypothesis that finance allows firms to respond to growth opportunities. Industry growth rates across countries are more highly correlated when the countries both have well-developed financial sectors. Thus, financial sector development enables industries to take advantage of global growth opportunities.

Although deeper financial intermediation may be a significant causal factor in economic growth, one cannot infer that every expansion of intermediary activity will be beneficial.

Cetorelli and Gambera (2001) extend this analysis by examining the effect of bank concentration on industries that rely on external finance. They find, paradoxically, that higher concentration in the banking industry is associated with more growth in industries that require more external finance. However, they also find an across-the-board depressing effect of concentration on growth. All in all, these studies provide specific illustrations of how financial sector development improves allocative efficiency by channeling financial resources.

Wurgler (2000) makes another important step in this literature with an effort to measure the relationship between allocative efficiency and financial sector development. He estimates the efficiency of capital allocation by the elasticity of industry investment to value added across industries in a given country. A higher elasticity indicates the extent to which a country is increasing investment in its growing industries. Using panel data for as many as twenty-eight industries (and up to thirty-two years of data), he obtains elasticity estimates for sixty-five countries. The highest elasticities are in Germany, Hong Kong, and New Zealand, and the lowest in Bolivia and Swaziland; the United States is thirteenth. Wurgler shows that the elasticities are related to characteristics of financial sector development. A specific mechanism of the finance growth relationship is that deeper financial sectors (measured by the ratio of either stock market capitalization or credit to GDP) help countries add to capital in growing industries. State ownership of industry inhibits this mechanism, and minority investor protections strengthen it.

Wurgler's paper takes some important steps toward identifying the channels of financial sector effects on allocative efficiency and growth. For example, he examines stock market synchronicity, a measure introduced by Morck, Yeung, and Yu (2000). We observed that equity market capitalization affects growth even though new equity issuance is always small. The markets are important because they assist the flow of information, which improves the efficiency of allocation. There will be more firm-specific information in markets where prices are not synchronized and seem to respond to firm-specific information.

Thus, the next stage of research has begun. Whether or not we are satisfied with the empirical literature of the 1990s, the finance growth nexus has become an established part of the economists' canon. The next generation of research is starting to delve into the black box and will show how financial deepening effects are transmitted to the real sector.

Conclusions

There is ample empirical evidence to make a convincing case that financial sector development promotes economic growth. However, this study has outlined some methodological reservations about the evidence used to establish this consensus. Nevertheless, the first decade of research on finance and growth identified relationships between growth and aggregate measures of financial sector development. The next stage, already under way, will identify specific institutional characteristics and financial sector channels that contribute to growth.

Research so far provides little in the way of rigorous guidance about how best to develop the financial sector. Although deeper financial intermediation may be a significant causal factor in economic growth, one cannot infer that every expansion of intermediary activity will be beneficial. Financial sector expansion that results from inflationary liquidity creation or deterioration in lending standards will not enhance long-run growth prospects. The observed association between financial sector deepening and growth does not, therefore, translate into

a simple prescription to encourage the unrestricted growth of financial intermediaries.

Similarly, the research on growth and finance provides policymakers with little guidance about the sequencing of financial sector developments. For example, we know that the expansion of bank credit is growth enhancing, but we do not how to promote credit expansion without compromising credit standards. Private sector credit evaluation capabilities, public sector regulatory oversight, and a sound legal and accounting infrastructure must all be in place as credit deepening occurs. The sequencing of financial sector developments is enormously important from a policy perspective. The recipe is not simple because the developments are likely to take place concurrently and mistakes are easy to make. Developing institutional capabilities and a legal tradition with enforcement standards is likely to be a slow process. It is easy to see how rapid credit expansion in a booming economy could wreak economic and political havoc even when a government is following a generally prudent prescription for financial sector development.

Recent history is full of examples of poor sequencing or a failure to have a robust institutional framework in place as financial deepening occurs. Bonin and Wachtel (2003) describe the problems that emerged in transition economies that opened equity markets before effective securities regulation was in place. Although securities laws were on the books, regulators were inexperienced and unable to apply them effectively. Thus, abuses were common, and the ensuing problems set back the development of equity markets.

The IMF has only recently introduced a program for financial sector stability assessments intended to evaluate financial sector developments in member countries and develop financial soundness indicators.[8] Previously, the IMF monitored macroeconomic developments and paid little attention to the financial sector. Perhaps as a result of some of the empirical research cited here, the IMF now understands that regulatory capabilities and the quality of institutions are as important as the growth of the money and credit aggregates. This change would be welcome since recent empirical work suggests that the quality of institutions is as important as their size.

Fundamental research on the finance-growth relationship has mushroomed in just the last few years. The strong evidence that financial development causes growth has contributed to the increased interest of the economics profession in financial institutions. However, much more needs to be done. Policymakers need to learn how to encourage the expansion of intermediation without creating inflation or excessive leverage. Researchers need to continue to develop the next stage of work on the channels of financial sector effects.

8. The program is described and reports can be found at www.imf.org/external/np/fsap/fsap.asp.

REFERENCES

Arellano, Manuel, and Stephen Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment fluctuations. *Review of Economic Studies* 58 (April): 277–97.

Arestis, Philip, and Panicos Demetriades. 1997. Financial development and economic growth: Assessing the evidence. *Economic Journal* 107 (May): 783–99.

Atje, Raymond, and Boyan Jovanovic. 1993. Stock markets and development. *European Economic Review* 37 (April): 632–40.

Barro, Robert J. 1991. Economic growth in a cross section of countries. *Quarterly Journal of Economics* 106 (May): 407–43.

Beck, Thorsten, Ross Levine, and Norman Loayza. 2000. Finance and the sources of growth. *Journal of Financial Economics* 58, no. 1–2:261–300.

Benhabib, Jess, and Mark M. Spiegel. 2000. The role of financial development in growth and investment. *Journal of Economic Growth* 5 (December): 341–60.

Blundell, Richard, and Stephen Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87 (August): 115–43.

Bonin, John P., and Paul Wachtel. 2003. Financial sector development in transition economies: Lessons from the first decade. *Financial Markets, Institutions, and Instruments* 12, no. 1:1–66.

Cetorelli, Nicola, and Michele Gambera. 2001. Banking structure, financial dependence and growth: International evidence from industry data. *Journal of Finance* 56 (April): 617–48.

Demirguc-Kunt, Asli, and Ross Levine. 1996. Stock market development and financial intermediaries: Stylized facts. *World Bank Economic Review* 10, no. 2:291–322.

Durlauf, Steven. 2001. Manifesto for a growth econometrics. *Journal of Econometrics* 100 (January): 65–69.

Fisman, Raymond, and Inessa Love. 2002. Patterns of industrial development revisited: The role of finance. World Bank Policy Research Working Paper 2877, August.

Florax, Raymond, Henri de Groot, and Reinout Heijungs. 2002. The empirical economic growth literature: Robustness, significance and size. Tinbergen Institute Discussion Paper TI 2002-040/3.

Goldsmith, Raymond. 1969. *Financial structure and development*. New Haven, Conn.: Yale University Press.

Graff, Michael A., and Alexander Karmann. 2001. Does financial activity cause economic growth? Dresden Working Papers in Economics 2/2001, February.

Holtz-Eakin, Douglas, Whitney Newey, and Harvey S. Rosen. 1988. Estimating vector autoregressions with panel data. *Econometrica* 56 (November):1371–95.

International Monetary Fund. 1999. *Exchange rate arrangement and currency convertibility: Development and issues.* World Economic and Financial Surveys. Washington, D.C.

Kenny, Charles, and David Williams. 2001. What do we know about economic growth? Or, why don't we know very much? *World Development* 29, no. 1:1–22.

King, Robert G., and Ross Levine. 1993a. Finance and growth: Schumpeter might be right. *Quarterly Journal of Economics* 108 (August): 717–37.

———. 1993b. Finance, entrepreneurship, and growth: Theory and evidence. *Journal of Monetary Economics* 32 (December): 513–42.

La Porta, Rafael, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2002. Government ownership of banks. *Journal of Finance* 57 (February): 265-301.

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. 1998. Law and finance. *Journal of Political Economy* 106 (December): 1133–55.

Leahy, Michael, Sebastian Schich, Gert Wehinger, Florian Pelgrin, and Thorsteinn Thorgeirsson. 2001. Contributions of financial systems to growth in OECD countries. OECD Economics Department Working Paper No. 280, March.

Levine, Ross. 1997. Financial development and economic growth: Views and agenda. *Journal of Economic Literature* 35, no. 2:688–726.

———. 1999. Law, finance and economic growth. *Journal of Financial Intermediation* 8 (January): 8–35.

———. 2002. Bank-based or market-based financial systems: Which is better? *Journal of Financial Intermediation* 11 (October): 398–428.

Levine, Ross, Norman Loayza, and Thorsten Beck. 2000. Financial intermediation and growth: Causality and causes. *Journal of Monetary Economics* 46 (August): 31–77.

Levine, Ross, and Sara Zervos. 1996. Stock market development and long-run growth. *World Bank Economic Review* 10 (May): 323–40.

———. 1998. Stock markets, banks, and economic growth. *American Economic Review* 88 (June): 537–58.

Lucas, Robert. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22 (July): 3–42.

Luintel, Kul, and Mosahid Khan. 1999. A quantitative reassessment of the finance-growth nexus: Evidence from a multivariate VAR. *Journal of Development Economics* 60 (December): 381–405.

McKinnon, Ronald I. 1973. *Money and capital in economic development.* Washington, D.C.: The Brookings Institution.

———. 1993. *The order of economic liberalization*. 2d ed. Baltimore and London: The John Hopkins University Press.

Morck, Randall, Bernard Yeung, and Wayne Yu. 2000. The information content of stock markets: Why do emerging markets have synchronous stock price movements? *Journal of Financial Economics* 58, no. 1–2:215–60.

Pagano, Marco. 1993. Financial markets and growth: An overview. *European Economic Review* 37 (April): 613–22.

Rajan, Raghuram, and Luigi Zingales. 1998. Financial dependence and growth. *American Economic Review* 88 (June): 559–86.

Robinson, Joan. 1962. *Essays in the theory of economic growth.* London: Macmillan.

Rousseau, Peter L. 2002. Historical perspectives on financial development and economic growth. Paper presented at the Federal Reserve Bank of St. Louis 26th Annual Economic Policy Conference, October.

Rousseau, Peter L., and Paul Wachtel. 1998. Financial intermediation and economic performance: Historical evidence from five industrialized countries. *Journal of Money, Credit, and Banking* 30 (November): 657–78.

———. 2000. Equity markets and growth: Cross country evidence on timing and outcomes, 1980–95. *Journal of Banking and Finance* 24 (December): 1933–57.

———. 2001. Inflation, financial development and growth. In *Economic theory, dynamics and markets: Essays in honor of Ryuzo Sato*, edited by T. Negishi, R. Ramachandran and K. Mino. Boston: Kluwer.

———. 2002. Inflation thresholds and the finance-growth nexus. *Journal of International Money and Finance* 21 (November): 777–93.

Shan, Jordan Z., Alan G. Morris, and Fiona Sun. 2001. Financial development and economic growth: An egg and chicken problem? *Review of International Economics* 9 (August): 443–54.

Temple, Jonathan. 1999. The new growth evidence. *Journal of Economic Literature* 37 (March): 112–56.

Theil, Michael. 2001. Finance and economic growth: A review of theory and the available evidence. European Commission Economic Paper No. 158, July.

Wachtel, Paul. 2001. Growth and finance: What do we know and how do we know it? *International Finance* 4 (Winter): 335–62.

Wachtel, Paul, and Peter L. Rousseau. 1995. Financial intermediation and economic growth: A historical comparison of the U.S., U.K. and Canada. In *Anglo-American Finance*, edited by M. Bordo and R. Sylla. Burr Ridge, Ill.: Irwin.

World Bank. 2000. *World development report 2000/2001: Attacking poverty.* Washington, D.C.: Oxford University Press for the World Bank.

Wurgler, Jeffrey. 2000. Financial markets and the allocation of capital. *Journal of Financial Economics* 58, no. 1–2:187–214.

Pricing Firms on the Basis of Fundamentals

MARK KAMSTRA

The author is a financial economist in the Atlanta Fed's research department. He is grateful for useful conversations with Lisa Kramer, Cesare Robotti, Tom Cunningham, and Paula Tkac and for extensive comments from Jerry Dwyer, Mark Fisher, and Larry Wall.

People often speculate that a particular stock is overpriced, or underpriced, and analysts sometimes issue stock price targets followed abruptly by price "corrections." A natural question is, What is the right price for a stock? Mergers and acquisitions of firms rely heavily on determining the right or fair price of a stock. One set of strategies to find the right price is to forecast cash flows from a stock market investment and calculate what that income is worth. Roughly speaking, this strategy is what fundamental valuation is all about, and it is the focus of this article.

Beyond an overview and illustration of commonly used fundamental valuation techniques, the article will discuss a new valuation approach developed in Kamstra (2001). The discussion will also explore severe market turndowns, such as the tech "bubble" of the late 1990s, to see if market prices reflected gross overvaluation of various stocks compared to the estimated fundamental values. Application of Kamstra (2001) to both blue chip and dot-com firms improves the ability to track market price movements, as will be demonstrated below with applications to BellSouth, Starbucks, Sun Microsystems, Microsoft, Yahoo, and the S&P 500 index.

The article first describes fundamental valuation approaches and establishes links between these methods. This review of techniques will draw on practitioner and academic financial literatures as well as the academic accounting literature.

The Literature

A large literature deals with the issue of stock valuation as a function of future cash flows and discount rates. Valuation methods based on fundamental analysis—forecasting future cash flows and discounting them to estimate the value of this income stream—all face the common criticism that these forecasts can be unreliable. Together with assumptions about the firm's ability to borrow funds and about market efficiency, such forecasts depend on a company's maintaining its investment and business strategies. Pricing by discounting future cash flows is intuitive, however.

The literature on fundamental valuation includes studies from accounting that explore restatements of the discounted dividend model in terms of accounting information (see Feltham and Ohlson 1995; Penman 1996; Burgstahler and Dichev 1997) and finance papers that often start with or derive the discounted dividend model (see Gordon 1962; Rubinstein 1976; Barksy and DeLong 1993; Campbell and Kyle 1993; Donaldson and Kamstra 1996; Chiang, Davidson, and Okuney 1997; Bakshi and Chen 1998). Finally, a literature written largely by practitioners for practitioners typically starts with the discounted dividend model of Gordon (1962) and augments it to allow for more flexibility.

A related literature has focused on the question of market efficiency, documenting abnormal return predictability based on earnings, size, and financial statement ratios.[1] There is considerable ongoing

controversy over the issue of market efficiency. The focus of the present work is not on market efficiency questions but rather on fundamental valuation *in the context of efficient markets*, though this study will comment on the efficient markets implications of the deviations observed between market and fundamental prices.

Contributions from the practitioner literature. The practitioner literature spans decades and provides a number of equity valuation approaches. There are somewhat indirect methods that are intended to rank stocks using price-earnings ratios (sometimes termed price relatives) or return-on-equity ratios combined with book-to-price ratios (see Beaver and Morse 1978; Wilcox 1984; Estep 1985; Peters 1991; Bauman and Miller 1997; Leibowitz 1999). There are equity valuation methods that use sales to calculate present value of future cash flows (see Leibowitz 1997). There are also methods based on the dividend growth model of Gordon (1962), or classic fundamental analysis.

Valuation methods based on fundamental analysis—forecasting future cash flows and discounting them to estimate the value of this income stream—all face the common criticism that these forecasts can be unreliable.

The papers based on Gordon's method start with a model equating market price to the sum of discounted future dividends. To produce a tractable formula, a structure is imposed, such as constant growth rates of dividends and constant discount rates. Many articles extend the simplest Gordon growth model to allow dividend growth rates to have several stages—for instance, permitting growth firms to start with high dividend growth rates and then decelerate to a stable long-run rate. Some studies also propose random but independent dividend growth rates.[2] The variations of the discounted dividend growth model used in this literature are rarely more than ad hoc attempts to capture real-world phenomena such as time-varying dividend growth rates. Pricing firms that do not pay out dividends is not considered explicitly, or else dividends are proxied as a constant fraction of observed earnings or sales. A good example of valuation based on sales in the absence of dividends is Damodaran (1994, 244–48), and standard investments texts outline how dividends may be replaced by earnings and payout ratios (see, for instance, Sharpe and Alexander 1990, 474–76).

An often-mentioned financial measure of fundamental value in this literature is the price-to-earnings (P/E) ratio. A high P/E ratio is often taken to imply that investors expect a high dividend growth rate, a low risk in holding the stock, or a high payout of earnings together with an average growth rate. The valuation of stocks using P/E ratios, most often termed relative value pricing, is studied by both academics and practitioners. P/E ratios are typically compared across similar firms to formulate buy/sell recommendations and to forecast price by multiplying a forecast of earnings by the current P/E ratio. Shares of firms that are not actively traded are often priced by finding an actively traded firm with similar risk, profitability, and investment opportunity characteristics and multiplying the actively traded firm's P/E ratio by the inactively traded firm's earnings.[3]

Contributions from the accounting literature. Studies in the accounting literature begin with the assumption of the discounted dividend model, imposing constant discount rates. The focus is on relating accounting information, such as earnings and book value, to stock valuation. The most popular techniques are the residual income valuation method and the free cash flow valuation method (see, for instance, Feltham and Ohlson 1995; Penman and Sougiannis 1998; Lee, Meyers, and Swaminathan 1999). Residual income is typically defined as earnings generated by a firm in excess of a normal rate of return on the company's book value (also termed abnormal earnings in the literature on residual income models).[4] Free cash flows are cash flows that could be withdrawn from a firm without lowering the current rate of growth.[5] The residual income method requires positive earnings and book value, and the free cash flow method requires positive free cash flows. Many firms have negative free cash flows, negative book value, and negative earnings. Among firms that have been included in the S&P 500 index at some point over the last twenty years, 6 percent have recorded at least one year with nonpositive book value; 12 percent have recorded at least one year with nonpositive earnings before interest, taxes, depreciation, and amortization (EBITDA); 89 percent have recorded at least one year with nonpositive free cash flow; and 32 percent have never had positive free cash flow. Of the more than 19,000 firms tracked by Compustat over the last twenty years, over 20 percent have recorded simultaneous nonpositive book value, nonpositive free cash flows, and nonpositive

EBITDA for at least one year on record; 6 percent have never had a positive book value; 52 percent have never had a positive free cash flow; and 22 percent have never had a positive EBITDA.

Contributions from the finance literature. In the finance literature, one approach taken to the fundamental valuation problem has been to implement some variant of the Gordon (1962) model of discounted dividends, which uses essentially the same starting point as the accounting literature. Although more formal, this literature also has much in common with the practitioner literature on fundamental valuation. The models that have been proposed vary from the simplest Gordon model with constant dividend growth rates and constant discount rates to multistage models with the growth rates varying in a step-wise manner—constant for a period of time (a step) and then shifting to a new level for a period of time (see, for instance, Brooks and Helms 1990, Barsky and DeLong 1993). The literature following directly out of Gordon (1962) motivates restrictions on dividend growth and discount rates either in an ad hoc fashion or by arguments based on analytic tractability. Another approach makes use of option-pricing methods but also imposes ad hoc assumptions to make the methods more straightforward to apply.[6]

Both streams of this literature—that following the Gordon (1962) growth model and that exploiting option-pricing tools—are closely related to each other. Both seek to impose sufficient structure on the dividend growth and discount rate processes to permit an explicit computable expression for the present value of future dividends.[7]

Donaldson and Kamstra (1996) generalize the Gordon (1962) model to allow arbitrary dividend growth and discount rate processes. The point of Donaldson and Kamstra's procedure is to avoid imposing structure on the dividend growth and discount rate processes and to let the data speak for themselves.[8]

Most investment professionals view any algorithmic valuation model as only a starting point to pricing equity, whether the model is based on price relatives like the P/E ratio or on classic fundamental analysis. For instance, in the context of zero-income stocks,

An often-mentioned financial measure of fundamental value in this literature is the price-to-earnings ratio.

1. An efficient market is one in which the price of assets reflects their fair value; that is, prices are unbiased. For work that presents evidence consistent with market inefficiency, see, for instance, Basu (1977), Jaffe, Keim, and Westerfeld (1989), Ball (1992), and Fama and French (1995). In contrast, Kirby (1997) demonstrates that the statistical significance of the evidence of market inefficiency from long-horizon returns is overstated.
2. A few examples include Hawkins (1977), Farrell (1985), Sorensen and Williamson (1985), Rappaport (1986), Hurley and Johnson (1994, 1998), and Yao (1997).
3. References to these sorts of rules can be found in textbooks like Brealey et al. (1992) and journal articles such as Peters (1991). See also Wilcox (1984), Estep (1985), Bauman and Miller (1997), and Campbell and Shiller (1998).
4. Preinreich (1938) derived that the stock price of a firm should equal the book value of the firm plus discounted abnormal earnings. Ohlson (1995) extends Preinreich and goes on to show the time period t stock price is a linear sum of time period t book value and abnormal earnings. This result assumes the discounted dividend model, constant discount rates, the clean-surplus relation, and linear autoregressive stochastic abnormal earnings. Ohlson also generalizes this relationship to admit information other than abnormal earnings. Feltham and Ohlson (1995) and Penman (1996), among others, extend Ohlson (1995). Feltham and Ohlson do so by focusing on the implications of conservative versus unbiased accounting for the Ohlson model while Penman focuses on the differential information contained in price-to-book versus price-to-earnings ratios in the context of the Ohlson model.
5. For a discussion of free cash flows and equity valuation, see Hackel and Livnat (1996) or Penman and Sougiannis (1998). Free cash flows are substantially different from accounting earnings and even accounting measures of the cash flow of a firm.
6. See Campbell and Kyle (1993), Chiang, Davidson, and Okuney (1997), Bakshi and Chen (1998), and Schwartz and Moon (2000, 2001) for examples of this approach. This literature starts with the representative consumer-complete market economic paradigm. Models are derived from primitive assumptions on markets and preferences, and the solution to the fundamental valuation problem is derived with the same tools used to price financial derivatives.
7. Even the solutions are often similar—the Gordon (1962) model is explicitly considered as a special case in the Bakshi and Chen (1998) option-pricing model.
8. The Donaldson-Kamstra methodology is similar to pricing path-dependent options because it involves a Monte Carlo simulation and numerical integration of the possible paths followed by the joint processes of dividend growth and discount rates, explicitly allowing path-dependence of the evolutions. See Donaldson and Kamstra (1996) for details.

Wilson (2000) argues that a practitioner should use discounted cash flow analysis together with scenario analysis, considering the fair value of a company under various possible scenarios and then judging which scenario is most likely to occur. If the market price is below the most likely fair value, he observes that it is appropriate to consider buying the stock. Wilson also points out the many difficulties involved in the simple application of discounted cash flow analysis, including the difficulty of determining the appropriate discount rate.

Fundamental Valuation

The fundamental value of a dividend-paying stock is merely the present value of the flow of dividends that are expected into the future.[9] That is, fundamental valuation involves solving equations A1 and A2 (in the appendix) to yield the market price equal to the expected discounted value of future dividends. This result holds if the stock market price contains no bubble—no "irrational exuberance."[10] Although this approach suggests that one must look into the distant future in order to price firms, there are a number of ingenious solutions that do not require complex forecasting methods. Among these are methods that simplify the basic formula to solve for future dividends and discount rates directly, such as Gordon growth models, and methods that make use of known market prices of other firms, such as the relative valuation model.

Practitioners use a variety of relative value models exploiting the notion that similar companies should have similar multiples of price to fundamental measures of value.

The relative value model. Practitioners use a variety of relative value models exploiting the notion that similar companies (in the same industry, at the same point in their growth cycle, of similar size, and so on) should have similar multiples of price to fundamental measures of value. That is, if company A is similar to company B, and company A has a price that is ten times its earnings (reflecting a 10 percent return on investment, roughly speaking), then company B's price is expected to be roughly ten times its earnings as well. The price for company A reflects a risk-return trade-off for that company, with market participants satisfied with a 10 percent return for a company with the characteristics of either company A or B. If market participants suddenly reassessed these companies as less risky, then a 10 percent return would be considered rich, and the price of both companies would be bid up, lowering the return until market participants would no longer consider the return to be unusually good.

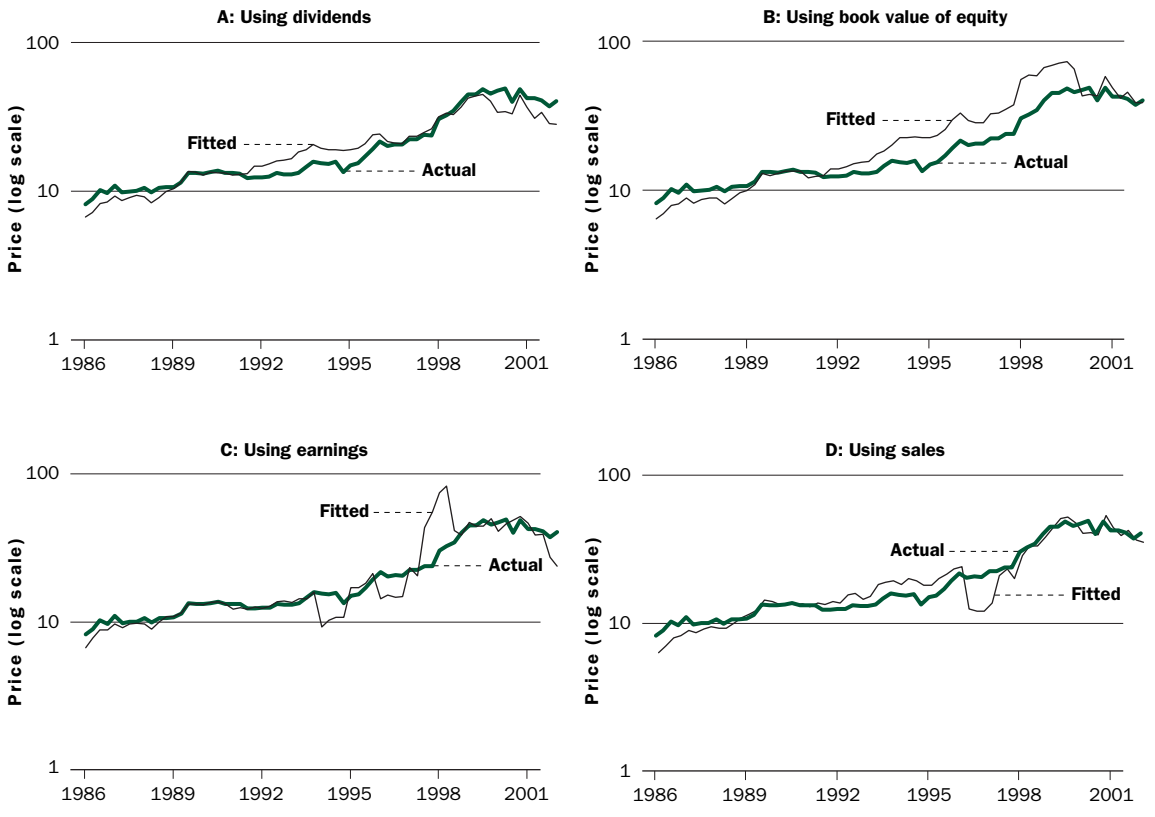
Relative value models do not require that firms pay out dividends. In the past, price-earnings multiples were most closely watched, but the advent of the technology boom in the 1990s led many to rely more on sales to price multiples because many companies did not have positive earnings (see, for instance, Leibowitz 1997). Other price relatives that are closely watched include the price-to-cash flow, the price-to-EBITDA ratio and the book-to-price (B/P) ratio. A B/P ratio of 1 is expected for relatively mature firms while growth firms are expected to produce lower ratios.

To illustrate relative value pricing, Figure 1 shows the price of BellSouth shares (NYSE:BLS), plotted quarterly, over the past sixteen years, using dividends, book value of equity, earnings, and sales of BellSouth and a similar firm, SBC Communications, another Baby Bell. In each panel the price scale is logarithmic, and the price is the closing price on the last day of trading in the first month of the quarter.[11] The respective relative value price is also plotted. The relative value price based on dividends for, say, the first quarter of 1985 is calculated by multiplying BellSouth's 1984 dividend by SBC Communications' 1985 price-dividend ratio.[12] The relative value prices based on earnings and on sales were calculated similarly. For the relative value price based on book value of equity, the book value for BellSouth reported for the fiscal year preceding a given quarter is multiplied by the price-to-book value reported for the same quarter for SBC; this calculation uses the closing price of SBC on the last trading day of the first month of the quarter and the book value reported for the fiscal year preceding the quarter.

A relative value model based on sales performs very well over most of the last sixteen years in this example; the relative sales price tracks the actual price very closely on average and in particular tracks actual market price remarkably well through the turmoil of the last three years. Relative valuation based on dividends also performs very well while relative valuation based on book value of equity or on earnings is much less reliable for the last sixteen years for BellSouth. More generally, relative

FIGURE 1

Relative Value Estimates of BellSouth Share Price



Note: Panels A through D present logarithms of the quarterly BellSouth share price level and the forecast price level from the relative value model. Panel A is based on dividends issued by BellSouth and the dividend yield of SBC Communications (SBC); panel B is based on the BellSouth book value of equity and the SBC book-to-market ratio; panel C is based on BellSouth earnings and the SBC earnings yield; panel D is based on BellSouth sales and the SBC sales yield.

valuation based on sales is attractive because most companies report sales while a great many companies issue dividends only rarely or have negative earnings or negative book value.

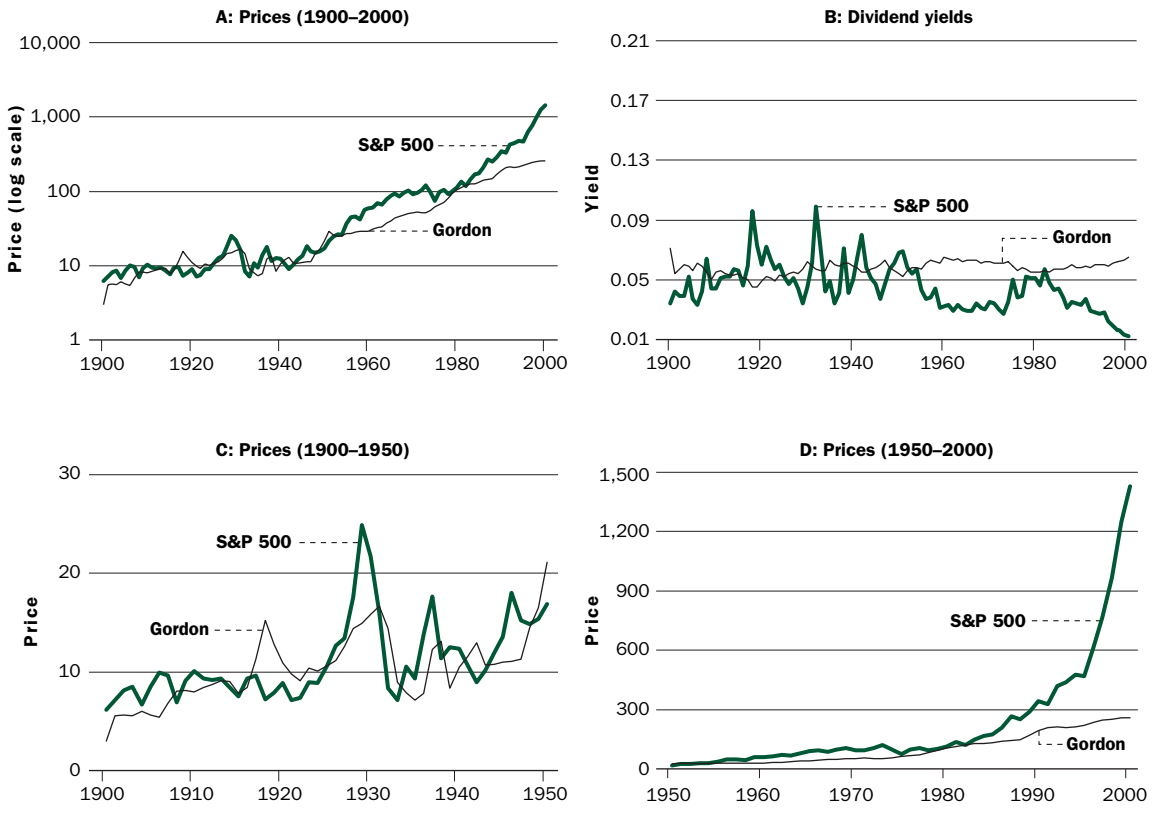
A word of caution—exploiting price relatives to value firms requires great care. Truly comparable firms must be found or the exercise is of little merit. Firms with advantages like a monopoly will be able to generate much higher profit margins and yields and will be grossly undervalued if benchmarked against otherwise similar firms. The best application of relative valuation is in valuing individual firms, provided a comparable firm can be found to the firm being valued. Pricing an index like the S&P 500 can be accomplished with relative valuation but only by using past values of the index. The classic dividend discounting valuation methods are easily applied to indices, however, as shown below.

The Gordon growth model. Perhaps the most widely used fundamental valuation method after relative valuation is the Gordon growth model. The Gordon fundamental price estimate does not, unlike relative valuation, require a comparable firm to the firm being valued and is derived with two simple assumptions: a constant discount rate and a constant growth rate of dividends. With these two assumptions, the valuation formula simplifies to a ratio involving the average dividend growth rate and the

9. The appendix provides technical descriptions of the valuation methods and models discussed in this article.
10. See Garber (1990), Kindleberger (1978), Shiller (1989), and White (1990) for a discussion of bubbles. Bubbles in asset prices are commonly defined as deviations of market prices from fundamental values.
11. The logarithm of price is presented to compress the scale of prices, making it possible to see detail throughout the period.
12. The SBC closing price used is that recorded on the last day of trading in January 1985, and the dividend used is the 1984 dividend.

FIGURE 2

The Gordon Growth Model for the S&P 500 Index



average discount rate multiplied by the most recent dividend (see equations A3 and A4 in the appendix).

To illustrate this pricing method, one can apply the Gordon growth model to the S&P 500 index over the past 130 years. During this time the S&P 500 index has enjoyed an average annual dividend growth rate of approximately 4 percent, and most measures of r, the average annual discount rate, would be close to 11 percent. The Gordon model price for, say, 1980 was calculated by estimating g as the average annual growth rate in dividends and r as the average annual return to holding the S&P 500 index for the 1871–1979 period and using dividends paid during 1979. This calculated Gordon price is then compared to the January 1980 price. Hence, the Gordon prices are all out-of-sample forecasts.

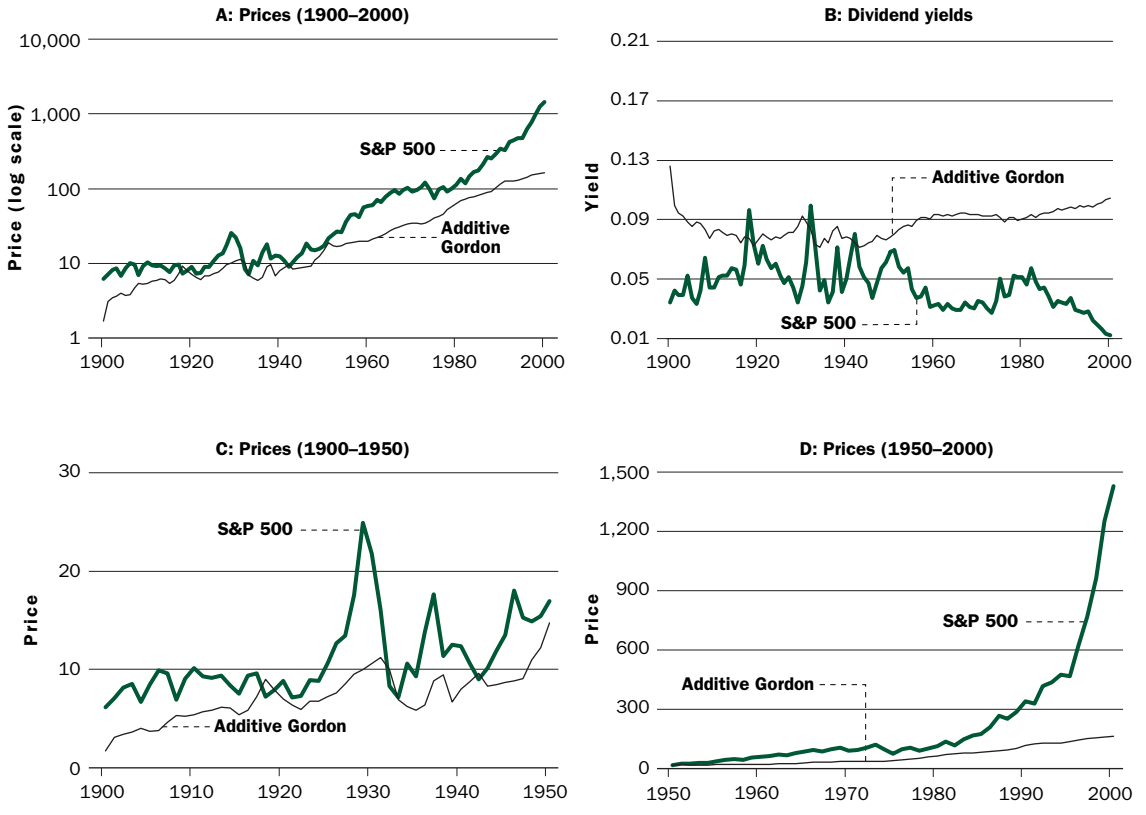
Figure 2 compares S&P 500 data with Gordon model estimates. Panel A shows prices and the Gordon model price for the period 1900 to 2000; a logarithmic scale makes it possible to see detail throughout the 100-year period. Panel B presents the market dividend yield (the S&P 500 index dividend divided by the market price) and the dividend yield using the Gordon price in place of the market price. Panels C and D present the actual market Gordon model prices (instead of the logarithmic values shown in Panel A) for the 1900–1950 and 1950–2000 periods, respectively. This display of fifty-year periods makes it easier to interpret deviations of the forecast and actual S&P 500 prices. The dividend yield shown in panel B highlights deviations of market and forecast prices. Evidence of large and persistent deviations between the market and forecast yields reveals whether the market is making systematic valuation errors or the forecasting model is performing very poorly.

Applying the Gordon model to the S&P 500 index annual data produces evidence of excessive market volatility (the forecast dividend yield is much less variable than the realized market yield) and of periods of inflated market prices—bubbles—in particular, during the 1920s, the 1960s, and the last half of the 1980s and 1990s. If the Gordon model is too simple, however—ignoring as it does changes in discount and dividend growth rates over time—this evidence may be misleading.

The additive and geometric Markov Gordon growth models. Hurley and Johnson (1994, 1998)

FIGURE 3

The Additive Markov Gordon Growth Model for the S&P 500 Index



and Yao (1997) develop Markov models—models that presume a fixed probability of, say, maintaining the dividend payment at current levels and a probability of raising it—to estimate dividends more realistically. These extensions of the Gordon growth model go back to the fundamental valuation equation, imposing less stringent assumptions. The simple Gordon growth model imposes a constant growth rate on dividends—dividends are expected to grow at the same rate every period—while these Markov models allow the probability of zero dividend growth. Two examples of these models found in Yao (1997) are the additive Markov Gordon model (equation 1 in Yao) and the geometric Markov Gordon model (equation 2 in Yao) (see equations A5 and A6 in the appendix).

For the S&P 500, over the last 130 years annual dividends have decreased 28.9 percent of the time and increased 71.1 percent of the time, the average absolute value of the change in annual dividends has been 0.161, and the average absolute value of the annual percentage change in dividends has been 9.2 percent.

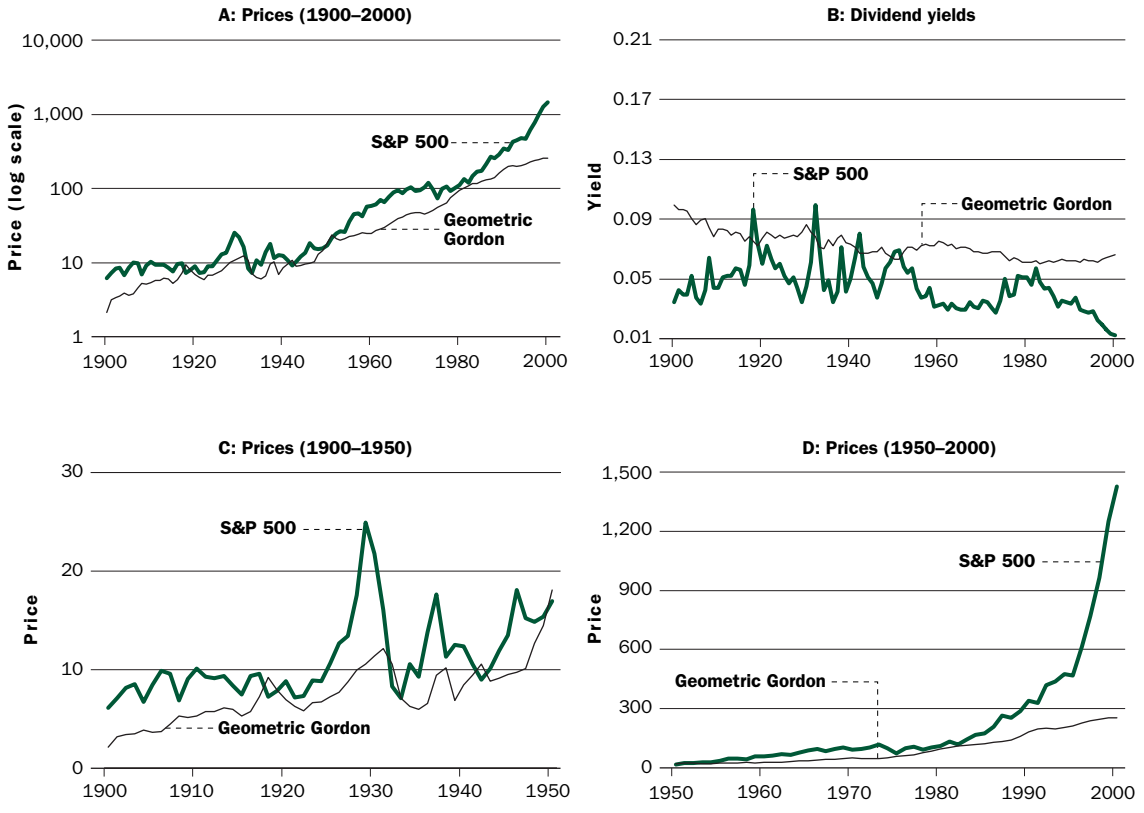
Figures 3 and 4 show prices and dividend yields from the additive and geometric Markov Gordon models versus the market price and dividend yield for the period 1900 to 2000. These models were implemented to produce out-of-sample price estimates just as the Gordon growth model was. The price for a given year was estimated using data up to but not including that year. Applying these two extensions of the Gordon model to the S&P 500 index annual data also produces evidence of excessive volatility and periods of inflated market prices—the 1920s, the 1960s, the 1980s, and the 1990s. Overall, the simplest Gordon model performs as well as the Markov model extensions, but none perform well.

Again, this poor performance could be the result of overly simple models that are not able to capture changes in value of the index or of a mispriced (irrationally priced) market.[13] The fact that the market price typically exceeds the forecast price from these models has led many to believe that the market has been overvalued at times, especially during

13. "Irrational pricing" can be defined as pricing based on expected price appreciation in the absence of expected cash flows.

FIGURE 4

The Geometric Markov Gordon Growth Model for the S&P 500 Index



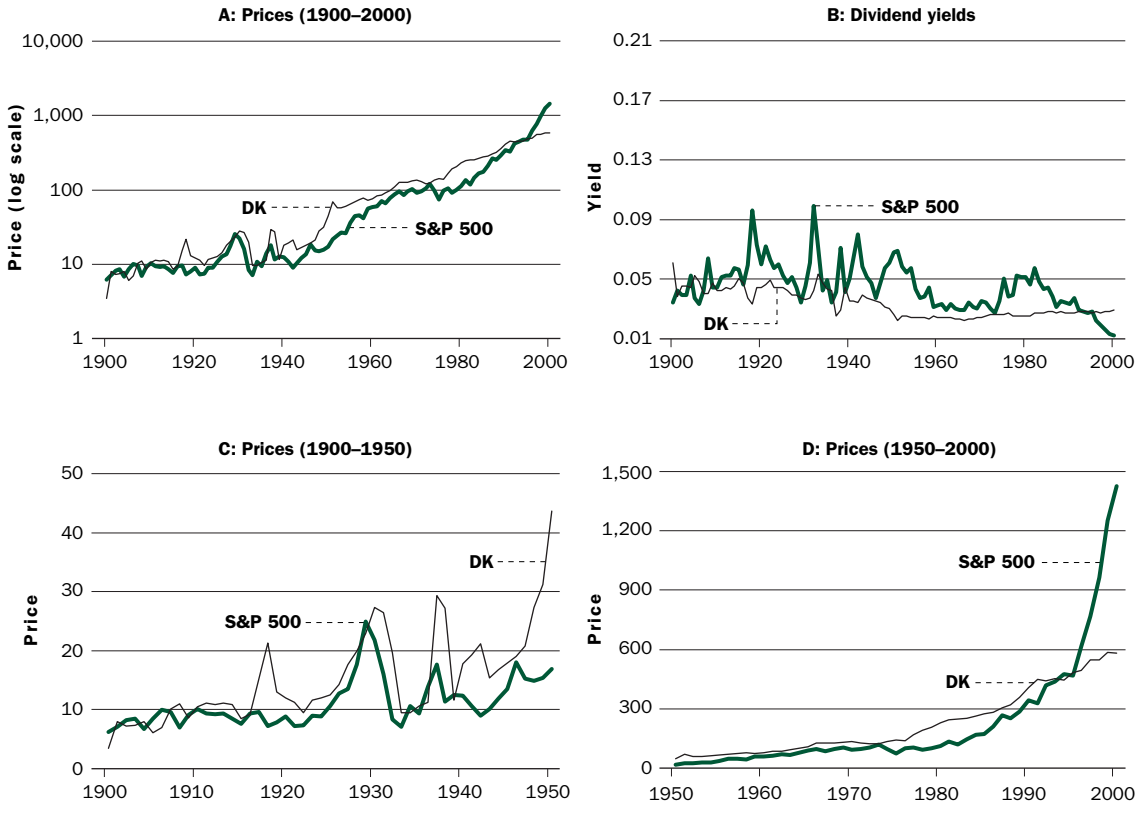
boom times like the 1920s, the 1960s, the 1980s, and the 1990s.

The Donaldson-Kamstra Gordon growth model. Donaldson and Kamstra (1996) further extend the Gordon model, imposing even fewer assumptions on the fundamental valuation equation than the Markov Gordon growth models and using statistical models of discounted dividend growth rates. The Donaldson-Kamstra model permits more flexible modeling of autocorrelation in growth rates than do simple Markov models. In the language of practitioners, this autocorrelation affects the fade rate: the speed at which company growth converges to its long-run stable growth rate (see, for instance, Wilson 2000). The greater the autocorrelation, the slower the fade to the long-run growth rate and the higher the value of a company enjoying temporarily high growth.

Why should one worry about autocorrelation? Take a simple example, a firm facing two equally likely scenarios for future discount rates. In one scenario, the discount rate decreases from its past average of 8 percent to a new average of 6 percent; in the other, the average rate increases to 10 percent. Once changed, the average rate remains fixed forever. The expectation before the rate change is for an average rate of 8 percent, just as in the past. Suppose dividend growth is expected to be 4 percent and the most recent dividend was $1. The Gordon growth model, applied blindly, would yield a price of $1/(0.08 – 0.04), or $25 per share. However, if interest rate changes are recognized as permanent (an extreme form of autocorrelation), then Gordon prices could be calculated separately for each scenario, and the two prices could be averaged to get a price that accounts for autocorrelation. The low discount rate case yields a price of $1/(0.06 – 0.04), or $50 per share, and the high discount rate case yields a price of $1/(0.10 – 0.04) or $16.67 per share, for an average price of roughly $33.33. Accounting for the autocorrelation dramatically changes the price estimate, increasing it by 30 percent. Although it is easy to adjust the Gordon model for a simple scenario like this one, the Donaldson and Kamstra technique makes it possible to perform extremely complex scenario analysis that is not feasible with simpler methods, such as scenarios in which the discount rate never settles to a constant, the dividend growth

FIGURE 5

The Donaldson-Kamstra Model for the S&P 500 Index



rate also moves around, and the two rates influence each other probabilistically.

Over the last 130 years the average annual value of discounted dividend growth rates has been 0.965 based on an equity premium of 3 percent, a premium recent research supports.[14]

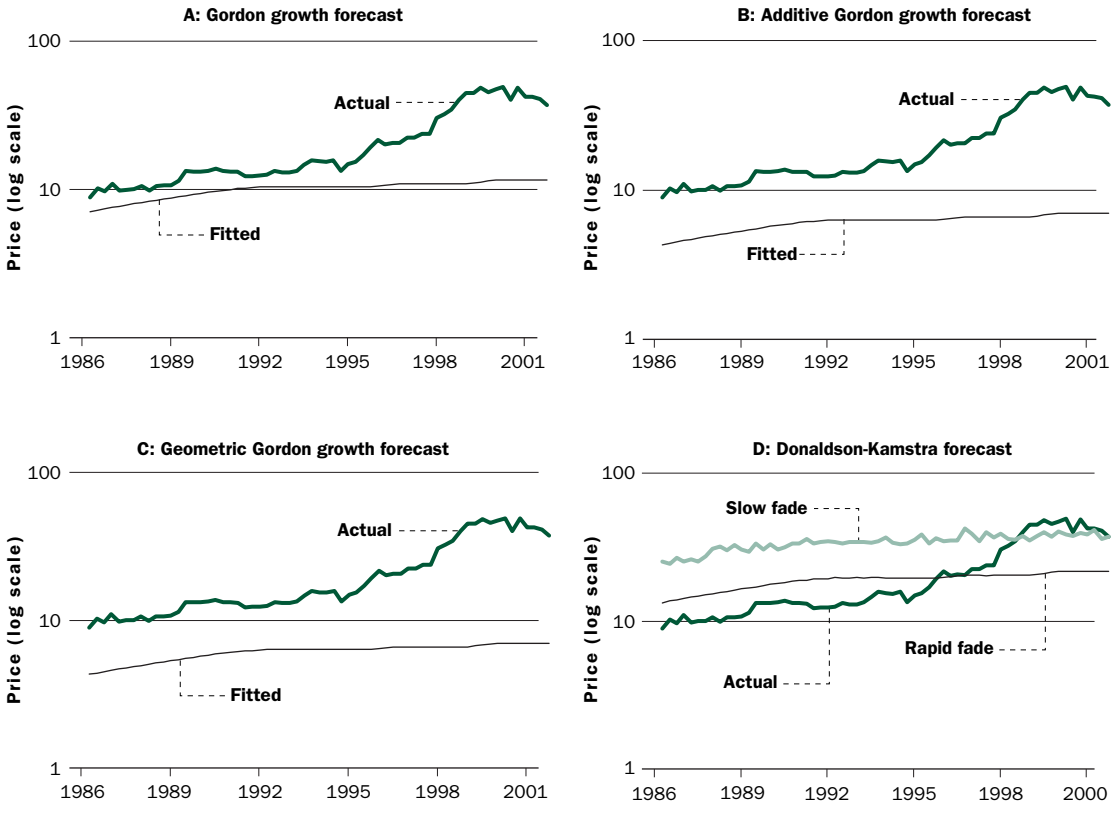
Figure 5 presents the price and dividend yield for the Donaldson-Kamstra (DK) model versus the market price and dividend yield for the period 1900 to 2000. The DK model was implemented to produce out-of-sample price estimates as the Gordon models were. The forecasts of discounted dividend growth rates are based on the last year of rates.[15] Applying the DK model to the S&P 500 index annual data produces much less evidence of surprisingly high market prices although the late 1990s still exhibit higher prices than the DK model's price forecasts. The dividend yields in panel B also provide evidence of excessively volatile market price movements in the last fifty years.

The ability of the DK model to capture much more market volatility, including the booms of the 1920s, the 1960s, and the 1980s, highlights the importance of accounting for the slow fade rate of dividend growth rates and discount rates. The continued failure to capture the height of the 1990s boom still leaves evidence of surprisingly high prices during the late 1990s. There is, however, still the question of a modeling failure; a large spike in prices could still be rationalized by a decrease in the fade rate during the 1990s.

14. The discounted dividend growth rate equals one plus the dividend growth rate divided by one plus the discount rate. This value should be close to, but less than, one. The equity premium is the extra return generated by stock market equity over relatively risk-free Treasury bills. The 3 percent premium is supported by, for instance, Fama and French (2002), Jorion and Goetzmann (1999), Jagannathan, McGrattan, and Scherbina (2001), and Donaldson, Kamstra, and Kramer (2003).
15. This forecasting model is an autoregressive model of order 1. The logarithm of discounted dividend growth rates was modeled for this exercise. The average value of the coefficient on the AR(1) term in the model was approximately 0.26, implying a fairly rapid fade rate. In as little as four years the impact from a change in the discounted dividend growth rate is expected to have virtually no remaining impact. An unexpected 10 percent increase in this growth would fade to less than 0.1 percent by year five. For implementation details, see Kamstra (2001) and Donaldson and Kamstra (forthcoming).

FIGURE 6

Forecasts of BellSouth Share Prices Based on Dividend-Forecasting Models



Note: Forecasts from the Donaldson-Kamstra model in panel D include models based on a rapid and a slow fade rate of growth in cash flows calibrated to the S&P 500 index over the last 100 years.

Application of Gordon growth models to BellSouth. It is interesting to apply the Gordon models based on dividends to the earlier example of BellSouth and explore how these models perform compared to relative valuation. A shortfall of the relative valuation approach is that a truly comparable firm may be difficult to find; great errors in valuation may follow an unwise choice of comparable firm. A model that does not look at prices, such as the Gordon growth models described earlier, should be immune to this problem.

The Gordon growth, additive and geometric Markov Gordon growth, and DK model price forecasts displayed in Figure 6 are formed using the same calibration used for the S&P 500 (BellSouth is, after all, a S&P 500 firm)—an average annual discount rate of 11 percent and an equity premium of 3 percent—and the same timing conventions used to form the relative value forecasts (so that all forecasts are out-of-sample).[16] For the DK model, price forecasts can be formed with the model described for the S&P 500 index, using the average annual estimated fade rate of dividend growth experienced by the S&P 500. This average fade rate is only an estimate, and plausible fade rates include slower and faster rates. These slower and faster rates allow bracketing high and low estimates of the share price. The low fade rate indicates a very slow reversion of growth to the long-term mean growth, implying high prices for fast-growing firms and low prices for firms that have experienced below-average or negative growth. A high fade rate indicates a very rapid reversion of growth to historic levels, so that price is not moved much by unexpected high or low growth.[17]

The simple Gordon growth model and the additive and geometric Gordon growth models all perform poorly, capturing neither the overall level of the share price nor the dramatic rise in share value in the late 1990s. Again, this performance is dramatic evidence of either irrational price setting or model failures. Allowing the Gordon models to incorporate larger dividend growth rates or smaller discount rates does not fix this problem—the

Gordon prices are not variable enough regardless of these settings. The DK model captures the average price level, reinforcing the notion that accounting for the fade rate matters, but the magnitude of the rise and fall of prices is not captured. The market prices start from below the lower bracket DK price and rise above the upper bracket DK price. Forecast prices are also much less volatile than actual market prices even when the fade rate is slow, the case for which we expect to see the most dramatic price swings. Clearly, relative valuation is capturing something these fundamental valuation methods fail to include. Hackel and Livnat (1996, 9) and others argue that dividends may be unreliable for assessing firm value because of institutional constraints on firm managers to smooth dividends over time. So a third possibility is that the models and the market prices are fine and that the problem is simply one of overly smoothed dividends. The next section presents discussions of formal extensions of classic dividend valuation models that allow the use of earnings, or sales, or other nondividend accounting numbers to value companies.

Augmenting Dividend Discounting Models

Kamstra (2001) extends dividend discount models like the Gordon growth model to firms that do not pay out dividends and incorporates nondividend information like earnings or sales figures into fundamental valuation of firms that do pay out dividends. The basic premise of this work is to incorporate the proceeds from share liquidation into the cash flows that are used to value the firm, accounting for the reduction in future growth of cash flows from this liquidation of shares. Share liquidation refers to selling a fraction of the stock holdings in a portfolio of stock. This sale generates immediate cash flow but reduces potential cash flows into the future. For instance, if a shareholder sells 3 percent of his shares this year, he will reduce his dividend flow next period from his remaining shares by 3 percent as well as reduce the cash from further liquidations because his portfolio will be 3 percent smaller next period.

A reasonable question is, How might share liquidation help one value a firm? It is well known that dividends are typically set low enough that the dividend payments can be maintained through economic downturns, leading them to be lowered only rarely and to inaccurately reflect future prospects for the firm, as argued by Hackel and Livnat (1996) and others. Augmenting dividends with the proceeds of share liquidation—say, to produce a yield equal to 3 percent of the sales yield—should produce valuation rules that more accurately reflect future prospects. Accounting for the share liquidation produces valuation formulas that are still tied to fundamentals of cash flow paid to investors even if the liquidation rule is itself calibrated to firm sales, not firm dividends.

The Kamstra method extends dividend discount models like the Gordon growth model to firms that do not pay out dividends and incorporates nondividend information like earnings or sales figures into fundamental valuation of firms that do pay out dividends.

A wealth of other work has, of course, been done on valuing zero-dividend firms. Among these studies are approaches that extend formal dividend discounting to zero-dividend firms relying on techniques similar to those used in option-pricing (see Bakshi and Chen 1998; Schwartz and Moon 2000, 2001), approaches that replace dividends with earnings and payout ratios or sales and profit margins, and, of course, relative valuation methods.[18]

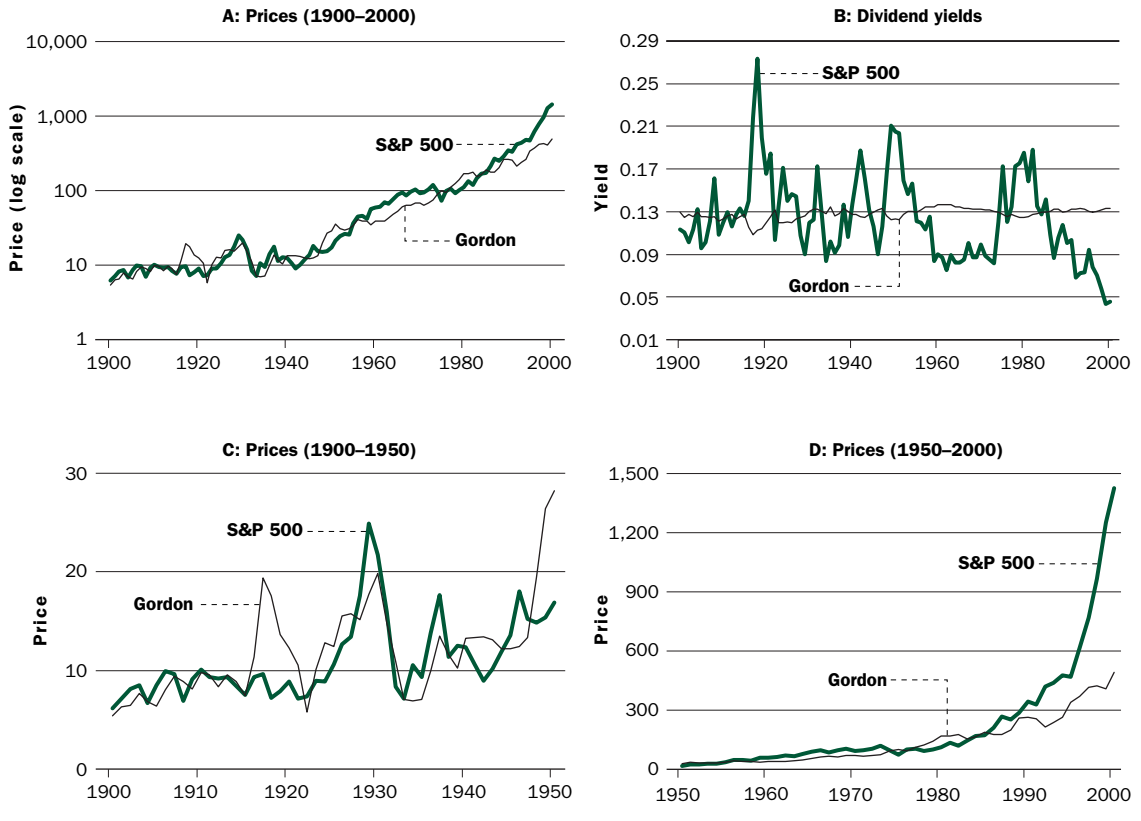
As the share liquidation rule of Kamstra (2001) uses past prices (to form the yield ratio), depending on how this rule is implemented it can have much in

16. The data range is shorter in Figure 6 than that displayed for relative valuation in Figure 1 because of the need to use greater lags of the data to form forecasts.
17. The high-fade-rate model was implemented by taking the average fade rate AR(1) parameter estimate of 0.26 and the standard deviation estimate of this parameter of 0.09 and subtracting two standard deviations from the parameter estimate, leaving a fade rate parameter of approximately 0.08. An unexpected 10 percent increase in growth would fade to less than 0.1 percent by year three for this parameter setting. The low-fade-rate model was implemented by taking the average fade rate AR(1) parameter of 0.26 and adding two standard deviations to it, producing a fade rate parameter of approximately 0.44. An unexpected 10 percent increase in growth would now take over seven years to fade to less than 0.1 percent.

    Bracketing price estimates can also be formed for the other Gordon growth models, but these price estimates are fairly small shifts up and down from the forecasts presented.
18. See Damodaran (1994, 244–48) for an example using sales and Sharpe and Alexander (1990, 474–76) for an example using earnings.

FIGURE 7

The Gordon Growth Model with Augmented Dividends



common with relative valuation. For instance, if one were valuing private equity one would not have past earnings yields to provide an expected earnings yield. In this context, one would pick an expected yield ratio by looking at the earnings-to-price ratio of similar but publicly traded firms, an approach borrowed from relative valuation.

An advantage of a share liquidation rule for valuation over relative valuation is that the fade rate in cash flows and discount rates can be incorporated with share liquidation as outlined in Kamstra (2001) while the relative value model ignores fade rates. The relative value model, taken at face value, assumes that the sales yield (or whatever yield is being considered, say, the earnings yield) will remain constant forever while Kamstra provides a method that makes this yield trend to some long-run stable level.

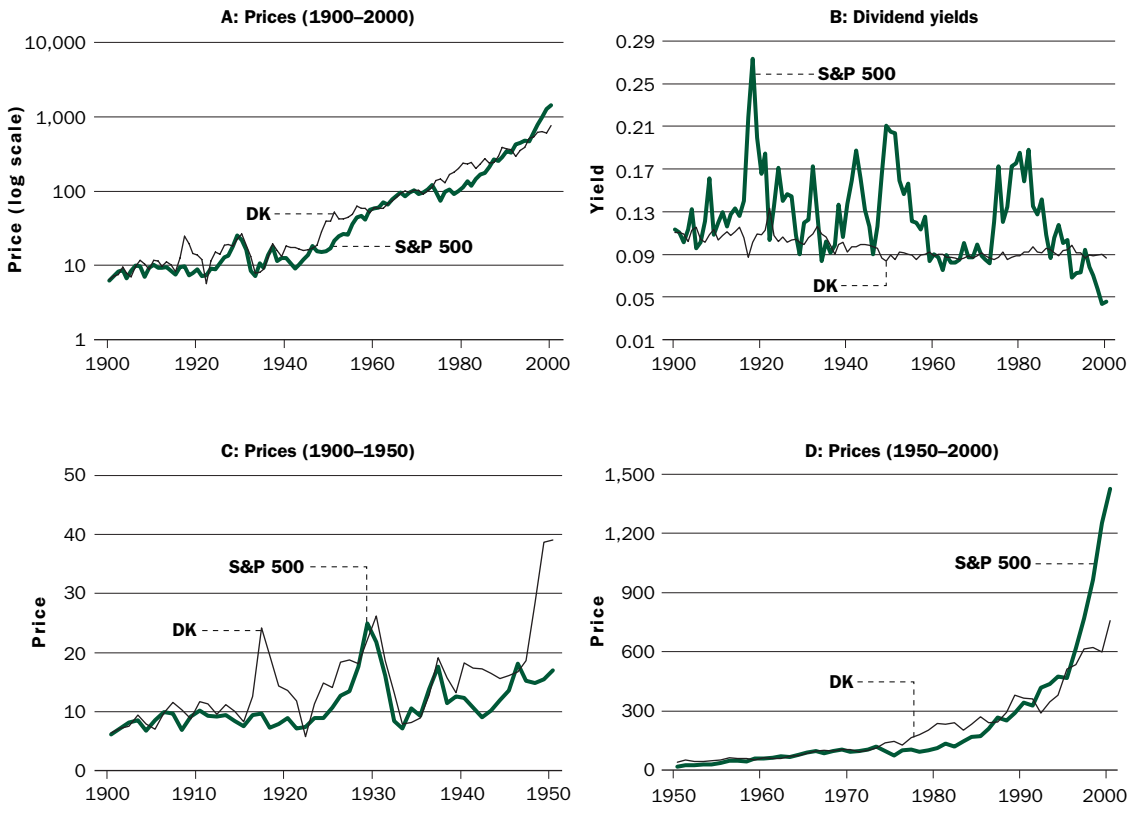
A disadvantage of the Kamstra method compared to the relative valuation method is that the long-run stable level of the yield ratio must be specified, and if this level is chosen incorrectly it can bias price forecasts. Also, the dependence (the fade rate) of the yield ratio and cash flow growth as well as the discount rate must be modeled, and typically the companies this method would be applied to would not have sufficient history to properly estimate fade rates based on own-company data. To implement the Kamstra method on such firms, a calibration to the S&P 500 index will be performed here, similar to that described above.

Compared to the techniques using option-pricing tools, the Kamstra method is simpler to apply though very similar in spirit. The approaches that replace dividends with earnings or sales have in common with relative valuation the disadvantage of not accounting for the fade rate of growth and discount rates and the advantage of simplicity over the Kamstra method.

Another valuation approach for zero-dividend firms is scenario analysis, the strategy of forecasting possible cash flows that a company might generate and computing the fair value of that company under the various scenarios. For instance, if there is a 50 percent chance that a company will be worth $5 per share and a 50 percent chance that it will be worth

FIGURE 8

The Donaldson-Kamstra Model with Augmented Dividends



$15 per share, then a price of $10 per share would be expected. This approach often combines elements of relative valuation and discounted cash flow analysis. Great skill and a great deal of detailed institutional knowledge of the firm and its industry are required to implement this valuation technique.[19]

I will restrict myself in this review to techniques I have already used, techniques that allow a narrow set of information for implementation and are therefore reasonably straightforward to apply.

Application to the S&P 500 index. In the case of the S&P 500 index with share liquidation set to equal accounting earnings, the total cash flow to the investor will equal the dividends paid plus earnings. The growth rate of this cash flow over the last 130 years equals 4.9 percent, and the annual yield ratio (see the appendix for the definition of this term) averages 8 percent. This information, together with the average annual discount rate (11 percent, as described above), allows us to produce Gordon prices, which are displayed in Figure 7.

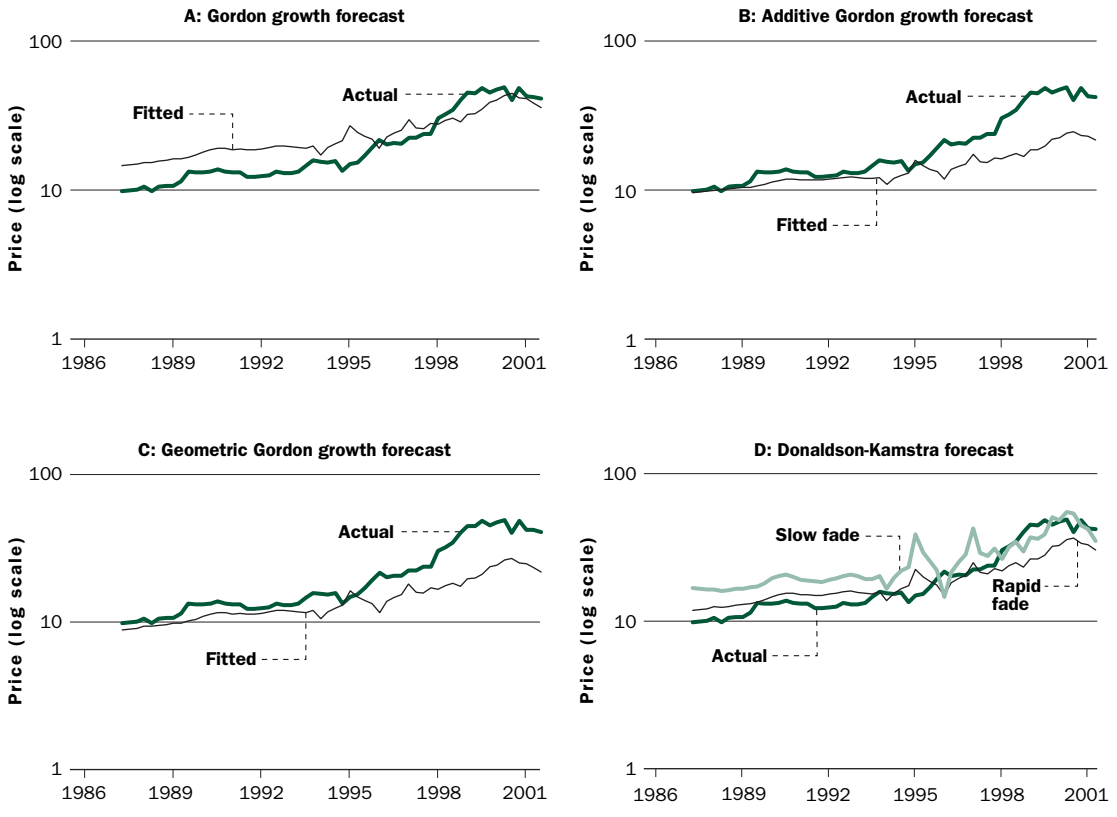
One can also produce DK prices based on the discounted cash flow growth rates. Based on earnings plus dividends and an equity premium of 3 percent, the average S&P 500 discounted growth rate over the last 130 years has been 0.974.[20] Figure 8 presents price and cash flow (dividends plus earnings) yields for the DK model versus the actual market price and yield for the period 1900 to 2000.[21]

Both the simplest Gordon growth model (Figure 7) and the DK model perform remarkably better when conditioned on the extra information provided by earnings. With the added consideration of earnings, the Gordon model captures most of the price rise and decline of the 1920s and tracks prices up to 1990 very well. In addition, the DK model now captures

19. See Wilson (2000) and Copeland, Koller, and Murrin (2000) for extended discussions that outline implementation details.
20. The discounted cash flow growth rate equals one plus the cash flow growth rate divided by one plus the discount rate. This value should be, on average, close to but less than one.
21. Results are presented for the slow-fade-rate DK model, calibrated as described above.

FIGURE 9

Forecasts of BellSouth Share Prices with Dividends Augmented by Share Liquidation Based on Earnings



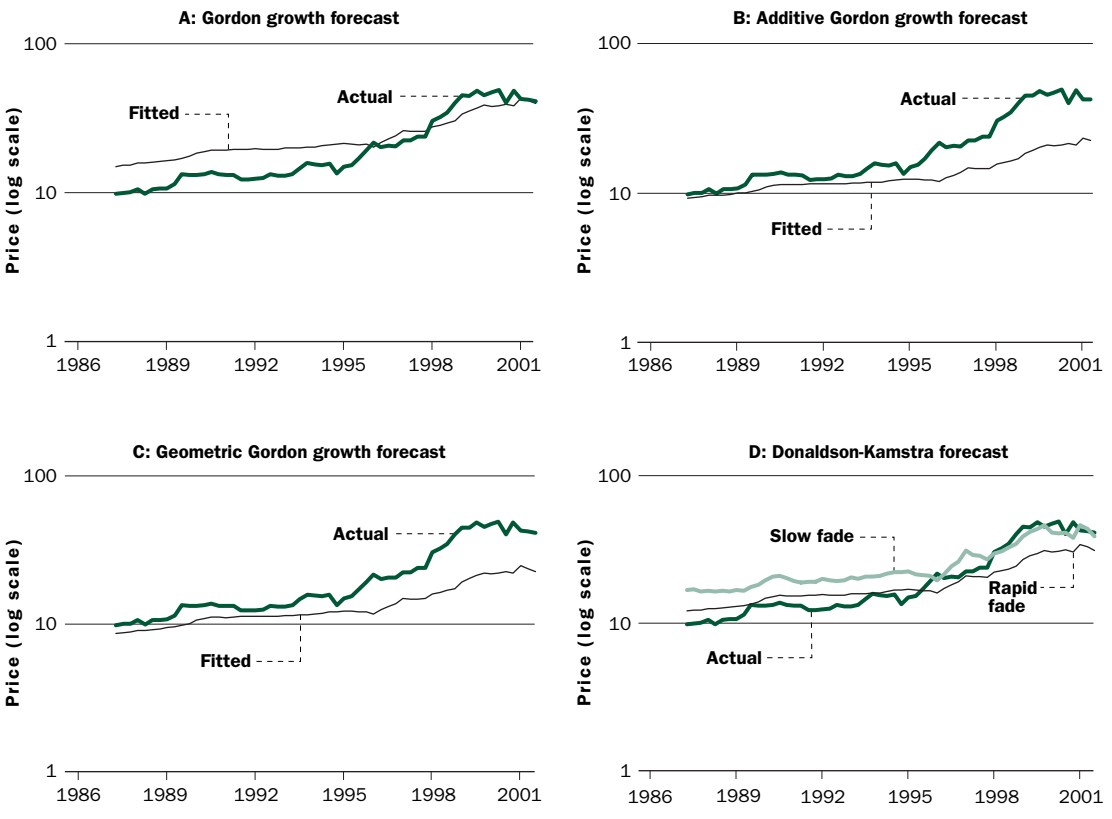
the timing of the turning point around the peak of the market in 1929 whereas both models peaked several years late when only dividends were used for pricing. Results from the additive and geometric Markov Gordon models are very similar to the basic Gordon model and are thus not presented here. The DK model forecasts prices and yields better than the Gordon growth model does, but the market yield ratio remains much more variable than can be explained by this model of fundamentals, and the market price at the end of the 1990s is approximately double what is forecast. Also worth noting is that some of the largest and most persistent deviations of market prices from forecast prices have occurred during periods of war, World War I and World War II in particular. This pattern highlights the fact that any algorithmic forecast based on a very restricted set of information can produce forecast prices that are less than reliable.

Application to BellSouth. It is also possible to apply these models to Bell South, augmenting its dividend payments with share liquidation based on either earnings, sales, or book value of equity. Figures 9, 10, and 11 show the logarithm of the price of BellSouth shares and of the forecast share price from the Gordon growth, additive and geometric Markov Gordon growth, and DK models.[22] Dividends are augmented with a stream of cash from liquidating shares equal to approximately 3 percent of the share price, calibrated to either earnings, sales, or book value of equity, and adjusting downward the growth of this cash stream to take account of this liquidation of shares.[23] Again, all the models borrow from the calibrations for the S&P 500 index, including setting the average discount rate to 11 percent and the equity premium to 3 percent and using the same timing conventions so that the forecasts are out-of-sample. The DK model results include forecasts from the slow- and rapid-fade-rate models, providing bracketing forecasts that one would expect to contain the actual market price.

The additive and geometric Gordon growth models never perform particularly well, regardless of the liquidation rule, but the classic Gordon growth

FIGURE 10

Forecasts of BellSouth Share Prices with Dividends Augmented by Share Liquidation Based on Book Value of Equity

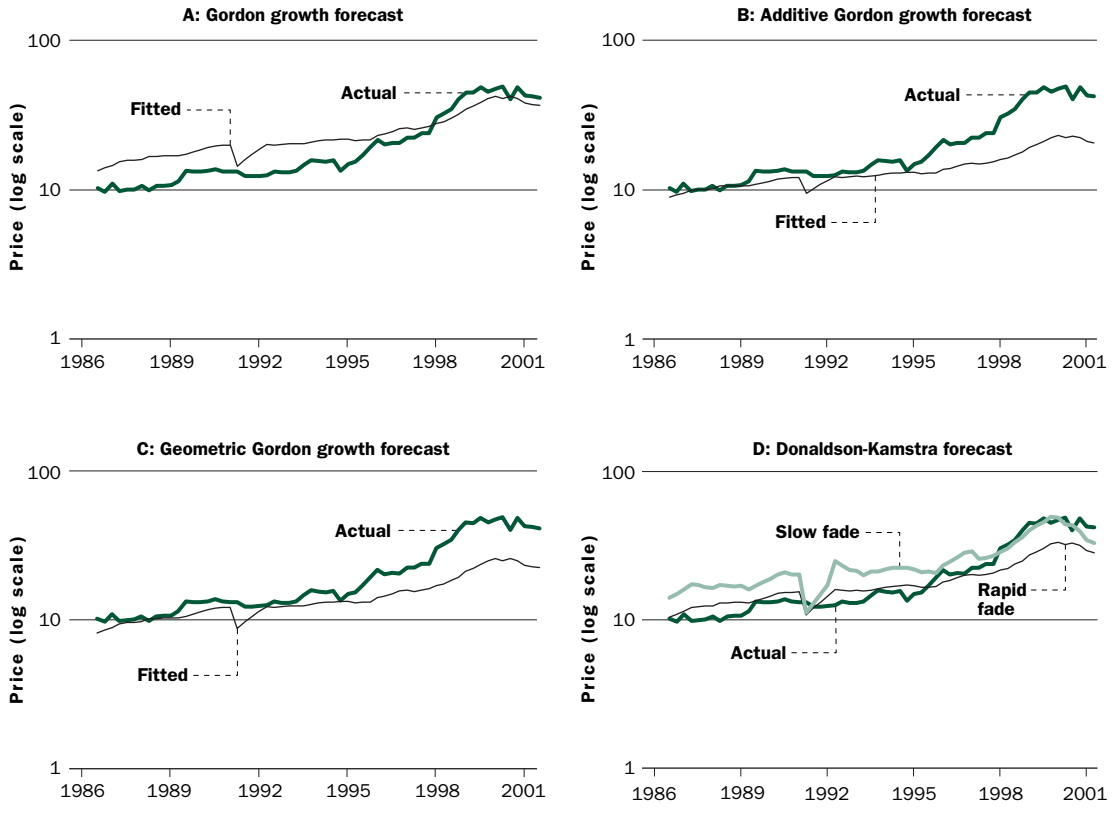


model and the DK model perform reasonably well with a liquidation rule based on sales or book value of equity. The classic Gordon model picks up much of the price rise and some of the decline over the period considered. The price bracket formed by the rapid- and slow-fade-rate DK models augmenting dividends with liquidation based on sales or book value indicates that the market price of BellSouth was often within a reasonable range of values, although a case for prices being somewhat high in 1999 and 2000 can still be made. Valuation based on augmenting dividends with an earnings-calibrated liquidation rule tends to have more false movements and random volatility, but this is a subjective judgment. The better performance when using book value or sales mirrors the relative valuation pattern found when using book value or sales for BellSouth and suggests at least two things. First, sales are more informative than earnings, at least for BellSouth over the last twenty years or so. Second, it is more difficult to argue that the price bubble observed in BellSouth stock valuation over the last three years was irrational—much of the up and down movement can be explained by changes in cash flows associated with high growth in sales, book value, and earnings.[24]

22. The data range is shorter here than those displayed in earlier figures because more lags of the data were needed to form forecasts.
23. The exact rule used when calibrating to sales was to liquidate a fraction of holdings equal to 3 percent multiplied by the most recently observed sales multiplied by the average price-to-sales ratio over the preceding year, not including the most recent quarter. The calibrations based on earnings and on book value of equity were performed similarly.
24. As the share liquidation scheme outlined here does make use of last year's sales, earnings, or book yield to calibrate liquidation, however, an argument can be made that a bubble was built into the "fundamental" price estimates generated. A share liquidation scheme based on the average yield over a longer period, as long as twenty years, does dampen the price rise in the late 1990s. Qualitatively, however, the evidence still supports the no-bubble view.

FIGURE 11

Forecasts of BellSouth Share Prices with Dividends Augmented by Share Liquidation Based on Sales



Application to high-growth firms. High-growth firms are particularly interesting for valuation exercises because such firms rarely pay out cash to shareholders, except perhaps to make share repurchases. The analysis will next consider Microsoft, Sun Microsystems, Starbucks, and Yahoo because all these firms are prominent members of the new economy, all have experienced very rapid growth, and all have had extreme share price fluctuations over the last several years. If the share liquidation scheme of Kamstra (2001) is used, these firms can be valued by traditional dividend-discounting models. Because all the Gordon growth models produce similar results, the discussion will focus on only the additive Gordon model and the DK model.

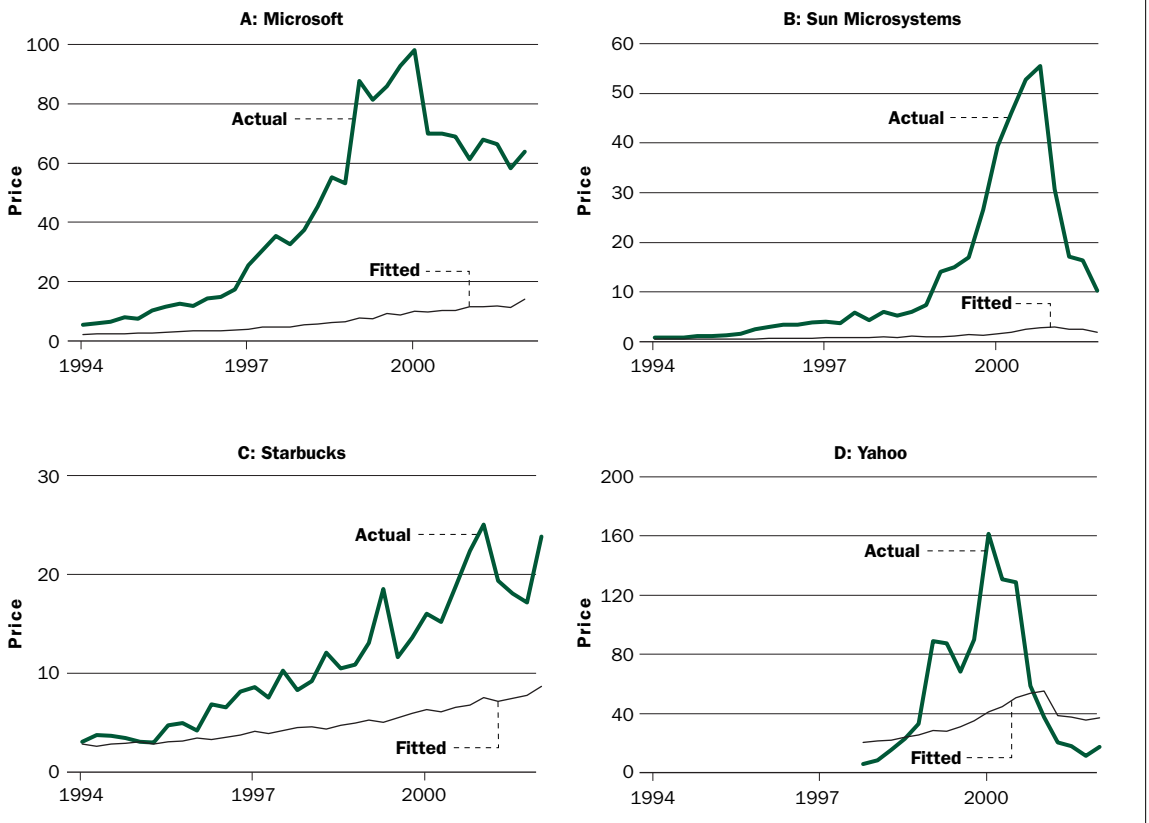
Figures 12 and 13 present forecasts from the additive Markov Gordon growth model and the DK model, respectively, based on a stream of cash from liquidating shares equal to approximately 3 percent of the share price, calibrated to sales. The calibrations used were identical to those used for BellSouth.[25] The DK model results include forecasts from the slow- and rapid-fade-rate models, providing bracketing forecasts that one would expect to contain the actual market price. In Figure 13 the prices are presented in logarithms to compress the scale for easier viewing.

The additive Gordon growth model (and indeed any Gordon model that ignores the fade rate) provides forecasts that are wildly at odds with the market prices for these four stocks. Even at the end of the sample, the last quarter of 2001, all but Yahoo still appear overvalued by the market. Given that the discount rate was calibrated to the S&P 500 index, even these prices are likely generous because these four stocks are arguably riskier than the average S&P 500 firm.[26] These plots of market prices versus fundamentals appear to strongly support the notion of a bubble in tech stock prices.

In contrast, the evidence from Figure 13 and the DK model prices—prices that take into account the fade rate of growth—does not support the notion of a bubble in the prices of these four stocks. By this method, Starbucks and Yahoo even appear to have

FIGURE 12

Additive Gordon Growth Forecasts of Share Prices of Microsoft, Sun Microsystems, Starbucks, and Yahoo Using Share Liquidation Based on Sales



been somewhat undervalued given the high growth rate of sales that each experienced over the last five years or so while Microsoft and Sun Microsystems display market prices that generally lie within the brackets formed by the slow- and rapid-fade-rate models. It should be noted that the bracketing forecasts generate a wide range of "reasonable" prices. Also, it should be noted that share value is estimated with models calibrated to the average S&P 500 firm, but an investor arguably faces more risk holding these four stocks than holding the average S&P 500 firm. Scenario analysis would factor in several different possible outcomes for all these stocks, including outright bankruptcy, that would lower the price estimates, possibly a great deal for Yahoo.

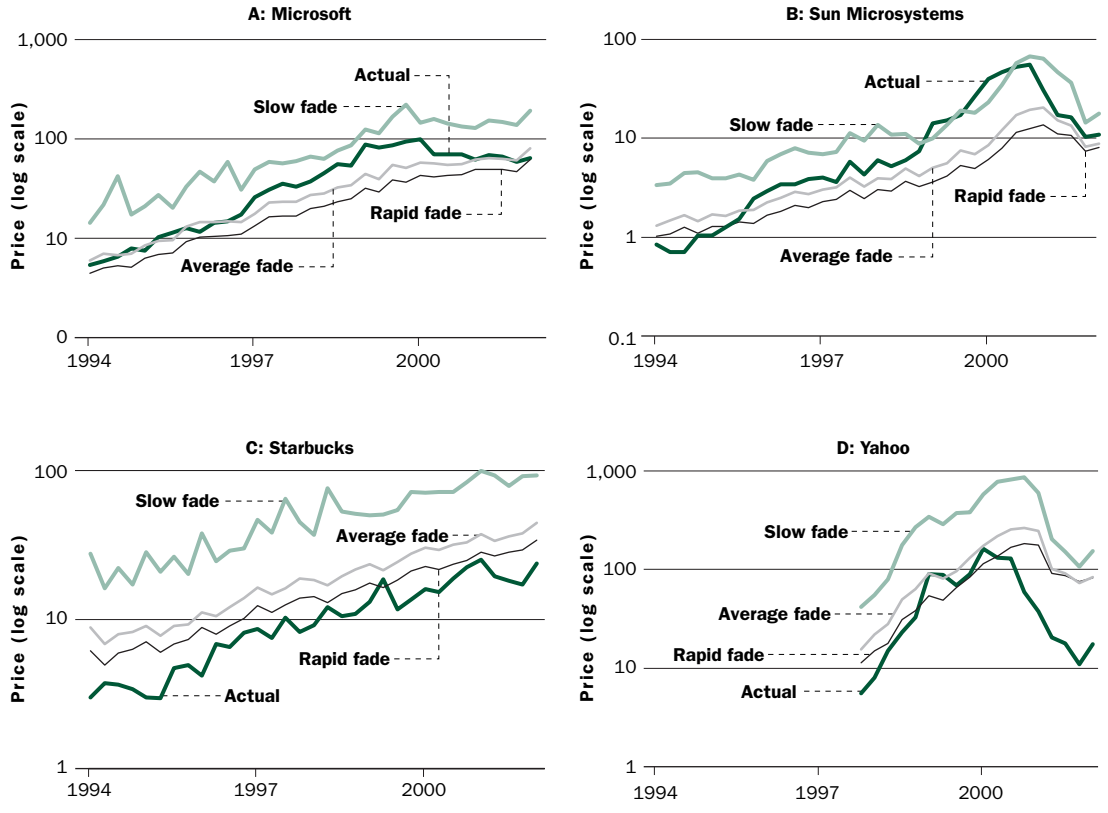
Conclusions

The pricing of stock market equity is one of the oldest problems in finance, but it is only in the last few decades that formal models have been developed to answer some of the most pressing questions. Many algorithms used to price equity are based on discounting cash flows accruing to the investor, though some methods base valuation on relative standing (that is, similar companies with similar balance sheets should be priced similarly, have similar price-to-sales ratios, similar price-to-earnings ratios, and so on) or a mix of discounting and relative valuation through scenario analysis. Valuation techniques based on the Gordon growth model, relative valuation, and the

25. The single exception is to the liquidation rule used in the DK and additive Markov Gordon growth models. Instead of using the average yield ratio from the past year, the average yield ratio is formed from the entire history of the stock. This rule is used because the high-growth stocks have much more volatile yield ratios than BellSouth has. Using only the past year produces similar results, with more exaggerated price movements forecast.
26. These four firms are all high-beta firms—that is, their stock returns are correlated with the overall market return but exhibit higher volatility than the overall market return.

FIGURE 13

Donaldson-Kamstra Bracketing Forecasts of Share Prices of Microsoft, Sun Microsystems, Starbucks, and Yahoo Using Share Liquidation Based on Sales



Note: Various parameterizations of the Donaldson-Kamstra model include a model calibrated to the average annual fade rate of the S&P 500 index over the last 100 years and models based on a rapid and a slow fade rate of growth in cash flows.

valuation method of Kamstra (2001) have been focused on here.

To demonstrate these methods in practice, they have been applied to pricing BellSouth shares, the S&P 500 index, and a few new-economy stocks. Pricing BellSouth using sales and sales growth is consistent with its dramatic rise and recent decline in price; this method is also appropriate for a small group of high-growth stocks, including Microsoft, Sun Microsystems, Starbucks, and Yahoo.

Fundamental models, however, have more trouble explaining the price movements of the overall market. Whether this failure to explain the overall market in the late 1990s is a shortcoming of these models or the kind of information used to price the index (earnings rather than, say, sales) or of an assumption of market rationality is not resolved here. Perhaps the most important point to take from this review is that algorithmic valuation techniques provide, at best, a rough starting point for firm valuation.

APPENDIX

Technical Details

Fundamental Valuation

The fundamental valuation equation is

$$(A1)\quad P_t = \mathfrak{E}_t\left\{\frac{P_{t+1}+D_{t+1}}{1+r_t}\right\},$$

where P_t is the price of the stock at the beginning of time period t, D_{t+1} is the per share dividend paid on the stock, r_t is the rate at which payments are discounted, and $\mathfrak{E}_t$ denotes the expectation of the future price and dividends conditional on what is known at the end of period t.

Solving equation A1 forward (substituting out for the future prices with future dividends) yields the textbook result that the market price equals the expected discounted value of future dividends.

$$(A2)\quad P_t = \mathfrak{E}_t\left\{\frac{D_{t+1}}{1+r_t}\right\} + \mathfrak{E}_t\left\{\frac{D_{t+2}}{(1+r_t)(1+r_{t+1})}\right\} + \mathfrak{E}_t\left\{\frac{D_{t+3}}{(1+r_t)(1+r_{t+1})(1+r_{t+2})}\right\} + \ldots$$

where "..." means "and so on."

The Gordon Growth Model

Define the growth rate of dividends from the beginning of period t to the beginning of period $t + 1$ as $g_t \equiv (D_{t+1} - D_t)/D_t$, so that $D_{t+1} = D_t(1 + g_t)$, $D_{t+2} = D_t(1 + g_t)(1 + g_{t+1})$, and so on, and rewrite equation A2 as follows:

$$(A3)\quad P_t = D_t\mathfrak{E}_t\left\{\frac{1+g_t}{1+r_t}\right\} + D_t\mathfrak{E}_t\left\{\frac{(1+g_t)(1+g_{t+1})}{(1+r_t)(1+r_{t+1})}\right\} + D_t\mathfrak{E}_t\left\{\frac{(1+g_t)(1+g_{t+1})(1+g_{t+2})}{(1+r_t)(1+r_{t+1})(1+r_{t+2})}\right\} + \ldots$$

Assume constant discount rates $r_{t+i} = r$ and constant growth rates of dividends $g_{t+i} = g$ for all values of i and with $g < r$. Substituting r and g into equation A3 and applying results from infinite series yields the classic Gordon price defined as P_t^G:

$$(A4)\quad P_t^G = D_t\left[\frac{1+g}{r-g}\right] \text{ or } P_t^G = \frac{D_{t+1}}{r-g}.$$

The Additive and Geometric Markov Gordon Growth Models

The additive Markov Gordon growth model is

$$(A5)\quad P_t^{ADD} = D_t/r + [1/r + (1/r)^2](q^u - q^d)\Delta,$$

where q^u is the proportion of the time the dividend increases, q^d is the proportion of the time the dividend decreases, and $\Delta = \Sigma_{t=2}^{T}|D_t - D_{t-1}|/(T-1)$ is the average absolute value of the level change in the dividend payment.

The geometric Markov Gordon growth model is

$$(A6)\quad P_t^{GEO} = D_t\left[\frac{1+(q^u-q^d)\Delta^{\%}}{r-(q^u-q^d)\Delta^{\%}}\right],$$

where $\Delta^{\%} = \Sigma_{t=2}^{T}|D_t - D_{t-1}|/D_{t-1}|/(T-1)$is the average absolute value of the percentage rate of change in the dividend payment.

The Donaldson-Kamstra Gordon Growth Model

Define the discounted dividend growth rate from the beginning of period t to the beginning of period $t + 1$ as $y_t \equiv (1 + g_t)/(1 + r_t)$ where again g_t equals $(D_{t+1} - D_t)/D_t$ (the dividend growth rate) and r_t is the discount rate. Rewrite equation A3 as follows:

$$(A7)\quad P_t = D_t[\mathfrak{E}_t\{y_t\} + \mathfrak{E}_t\{y_t y_{t+1}\} + \mathfrak{E}_t\{y_t y_{t+1} y_{t+2}\} + \ldots].$$

The Donaldson and Kamstra (1996) method estimates the price by generating thousands of possible values of y_t, y_{t+1}, y_{t+2}, ..., y_{t+I} (values of y out into the distant future, I periods from the present) and calculating

$$PV = D_t[y_t + y_t y_{t+1} + y_t y_{t+1} y_{t+2} + \ldots + y_t y_{t+1} y_{t+2} \ldots y_{t+I}]$$

for each, averaging these values of PV. Although this sum indexed by the parameter I should, technically, include all future values of y to infinity, if I is large enough there is only a very small truncation error. Donaldson and Kamstra (forthcoming) have found values of I = 400 to 500 for annual data to suffice. What distinguishes this method from other Gordon growth models is the way y_t is generated. Donaldson and Kamstra suggest time series models for y_t that have autoregressive patterns of dependence, a forecastable process.

The Augmented Dividend Case

Define A_t as the total cash an investor receives from her stock holdings in a particular company, including the payments the company makes to the investor (dividends paid by the company) and any proceeds the investor receives as a result of selling shares in the company (to other investors). Define V_t as the accounting variable (earnings, total asset value, sales, etc.) that will be used to calibrate investor share liquidations and notice that $A_t = D_t + V_t$, where again D_t is dividends. Define A's growth rate as $g_t^a \equiv (A_{t+1} - A_t)/A_t$.

APPENDIX (continued)

Define $f_t = \alpha V_t/P_t$ where α is set to determine the yield. (For instance, if the firm pays no dividends, V_t is earnings, and one wants to extract from one's portfolio a yield equal to the earnings yield, one would set α equal to 1; if V_t was annual sales, one would set α to 7 percent to extract a yield of roughly 7 percent, as annual sales per share equals price per share for many firms, based on Compustat annual data for S&P 500 firms over the last twenty years.) Define the average yield ratio f_t as f, the average cash payment growth rate g_t^a as g^a, and the average discount rate r_t as r. Then Kamstra (2001) derives the Gordon price with augmented dividends to be

$$(A8) \quad P_t^{G,v} = A_t\left[\frac{1+g^a}{r-g^a+f(1+g^a)}\right].$$

For the zero-dividend case, it can be shown that r must equal g^a so that equation A8 reduces to $P_t^{G,v} = V_t/f$.[1] This formula is the relative value model. Knowing what yield (f) to expect, say, from knowing what yields similar firms generate and knowing what V is for the firm one is valuing reveals what the price of the firm should be. For example, if the firm is generating earnings of \$1 per share and similar firms have an earnings yield of 5 percent, then the firm should have a value of \$1/0.05 = \$20.

The Donaldson and Kamstra (1996) model was also extended in Kamstra (2001). Define $y_t^a = (1-f_t)(1 + g_t^a)/(1 + r_t)$ and rewrite equation A7 as

$$(A9) \quad P_t = A_t\boldsymbol{\varepsilon}_t\{y_t^\alpha + y_t^\alpha y_{t+1}^\alpha + y_t^\alpha y_{t+1}^\alpha y_{t+2}^\alpha + \ldots\}.$$

1. See, for instance, Ohlson (1991) for a discussion in the context of the growth of earnings when firms pay out less than 100 percent of earnings.

REFERENCES

Bakshi, Gurdip, and Zhiwu Chen. 1998. Stock valuation in dynamic economies. University of Maryland unpublished paper.

Ball, Ray. 1992. The price-earnings anomaly. *Journal of Accounting and Economics* 15:319–45.

Barsky, Robert B., and J. Bradford DeLong. 1993. Why does the stock market fluctuate? *Quarterly Journal of Economics* 108 (May): 291–311.

Basu, Sanjoy. 1977. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient markets hypothesis. *Journal of Finance* 32, no. 3:291–311.

Bauman, W. Scott, and Robert E. Miller. 1997. Investor expectations and the performance of value stocks versus growth stocks. *Journal of Portfolio Management* 23, no. 3:57–68.

Beaver, William, and Dale Morse. 1978. What determines price-earnings ratios? *Financial Analysts Journal* 34, no. 4:65–76.

Brealey, Richard, Stewart Myers, Gordon Sick, and Ronald Giammarino. 1992. *Principles of corporate finance*. Toronto: McGraw-Hill Ryerson, Ltd.

Brooks, Robert, and Billy Helms. 1990. An N-stage, fractional period, quarterly dividend discount model. *Financial Review* 25 (November): 651–57.

Burgstahler, David C., and Ilia D. Dichev. 1997. Earnings, adaptation and equity value. *Accounting Review* 72, no. 2:187–215.

Campbell, John Y., and Albert S. Kyle. 1993. Smart money, noise trading and stock price behavior. *Review of Economics Studies* 60 (January): 1–34.

Campbell, John. Y., and Robert. J. Shiller. 1998. Valuation ratios and the long-run stock market outlook. *Journal of Portfolio Management* 24, no. 6:11–26.

Chiang, Raymond, Ian Davidson, and John Okuney. 1997. Some theoretical and empirical implications regarding the relationship between earnings, dividends and stock prices. *Journal of Banking and Finance* 21 (January): 17–35.

Copeland, Tom, Tim Koller, and Jack Murrin. 2000. *Valuation: Measuring and managing the value of companies*. 3d ed. New York: John Wiley and Sons.

Damodaran, Aswath. 1994. *Damodaran on valuation*. New York: John Wiley and Sons.

Donaldson, R. Glen, and Mark Kamstra. 1996. A new dividend forecasting procedure that rejects bubbles in asset prices. *Review of Financial Studies* 9 (Summer): 333–83.

——. Forthcoming. Estimating and testing fundamental stock prices: Evidence from simulated economies. In *Computer-intensive econometrics*, edited by D. Giles. New York: Marcel Dekker.

Donaldson, R. Glen, Mark Kamstra, and Lisa Kramer. 2003. Stare down the barrel and center the crosshairs: Targeting the ex ante equity premium. Federal Reserve Bank of Atlanta Working Paper 2003-4, January.

Estep, Preston W. 1985. A new method for valuing common stocks. *Financial Analysts Journal* 41, no. 6:26–33.

Fama, Eugene F., and Kenneth French. 1995. Size and book-to-market factors in earnings and returns. *Journal of Finance* 50, no. 1:131–55.

——. 2002. The equity premium. *Journal of Finance* 57, no. 2:637–59.

Farrell, James L. 1985. The dividend discount model: A primer. *Financial Analysts Journal* 41, no. 6:16–25.

Feltham, Gerald A., and James A. Ohlson. 1995. Valuation and clean surplus accounting for operating and financial activities. *Contemporary Accounting Research* 11, no. 2:689–731.

Garber, Peter. 1990. Famous first bubbles. *Journal of Economic Perspectives* 4 (Spring): 35–54.

Gordon, Myron. 1962. *The investment, financing and valuation of the corporation*. Homewood, Ill.: Irwin.

Hackel, Kenneth S., and Joshua Livnat. 1996. *Cash flow and security analysis*. Chicago: Irwin Publishing.

Hawkins, David F. 1977. Toward an old theory of equity valuation. *Financial Analysts Journal* 33, no. 6:48–53.

Hurley, William J., and Lewis D. Johnson. 1994. A realistic dividend valuation model. *Financial Analysts Journal* 50, no. 4:50–54.

——. 1998. Generalized Markov dividend discount models. *Journal of Portfolio Management* 24, no. 1:27–31.

Jaffe, Jeffrey, Donald B. Keim, and Randolph Westerfield. 1989. Earnings yields, market values, and stock returns. *Journal of Finance* 44, no. 1:135–48.

Jagannathan, Ravi, Ellen R. McGrattan, and Anna Scherbina. 2001. The declining U.S. equity premium. NBER Working Paper W8172.

Jorion, Philippe, and William N. Goetzmann. 1999. Global stock markets in the twentieth century. *Journal of Finance* 54, no. 3:953–80.

Kamstra, Mark. 2001. Rational exuberance: The fundamentals of pricing firms, from blue chip to "dot com." Federal Reserve Bank of Atlanta Working Paper 2001-21, November.

Kindleberger, Charles P. 1978. *Manias, panics, and crashes: A history of financial crisis*. New York: Basic Books.

Kirby, Chris. 1997. Measuring the predictable variation in stock and bond returns. *Review of Financial Studies* 10, no. 3:579–630.

Lee, Charles M.C., James Meyers, and Bhaskaran Swaminathan. 1999. What is the intrinsic value of the Dow? *Journal of Finance* 54, no. 5:1693–1741.

Leibowitz, Martin L. 1997. Franchise margins and the sales-driven franchise value. *Financial Analysts Journal* 53, no. 6:43–53.

——. 1999. P/E forwards and their orbits. *Financial Analysts Journal* 55, no. 3:33–47.

Ohlson, James A. 1991. The theory of value and earnings and an introduction to the Ball-Brown analysis. *Contemporary Accounting Research* 8, no. 1:1–19.

——. 1995. Earnings, book values, and dividends in equity valuation. *Contemporary Accounting Research* 11, no. 2:661–87.

Penman, Stephen H. 1996. The articulation of price-earnings ratios and market-to-book ratios and the evaluation of growth. *Journal of Accounting Research* 34, no. 2:235–57.

Penman, Stephen H., and Theodore Sougiannis. 1998. A comparison of dividend, cash flow and earnings approaches to equity valuation. *Contemporary Accounting Research* 15, no. 3:343–83.

Peters, Donald J. 1991. Using PE/growth ratios to develop a contrarian approach to growth stocks. *Journal of Portfolio Management* 17, no. 3:49–51.

Preinreich, Gabriel A.D. 1938. Annual survey of economic theory: The theory of depreciation. *Econometrica* 6:219–41.

Rappaport, Alfred. 1986. The affordable dividend approach to equity valuation. *Financial Analysts Journal* 42, no. 4:52–58.

Rubinstein, Mark. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics* 7, no. 2:407–25.

Schwartz, Eduardo S., and Mark Moon. 2000. Rational pricing of Internet companies. *Financial Analysts Journal* 56, no. 3:62–75.

——. 2001. Rational pricing of Internet companies revisited. *Financial Review* 36 (November): 7–26.

Sharpe, William F., and Gordon J. Alexander. 1990. *Investments.* 4th ed. Englewood Cliffs, N.J.: Prentice Hall.

Shiller, Robert. 1989. *Market volatility*. Cambridge, Mass.: MIT Press.

Sorensen, Eric H., and David A. Williamson. 1985. Some evidence on the value of dividend discount models. *Financial Analysts Journal* 41, no. 6:60–69.

White, Eugene. 1990. The stock market boom and crash of 1929 revisited. *Journal of Economic Perspectives* 4 (Spring): 67–84.

Wilcox, Jarrod W. 1984. The P/B-ROE valuation model. *Financial Analysts Journal* 40, no. 1:58–66.

Wilson, Barney. 2000. Valuing zero-income stocks: A practical approach. *Practical Issues in Equity Analysis*, 75–80. Association for Investment Management and Research.

Yao, Yulin. 1997. A trinomial dividend valuation model. *Journal of Portfolio Management* 23, no. 4:99–103.