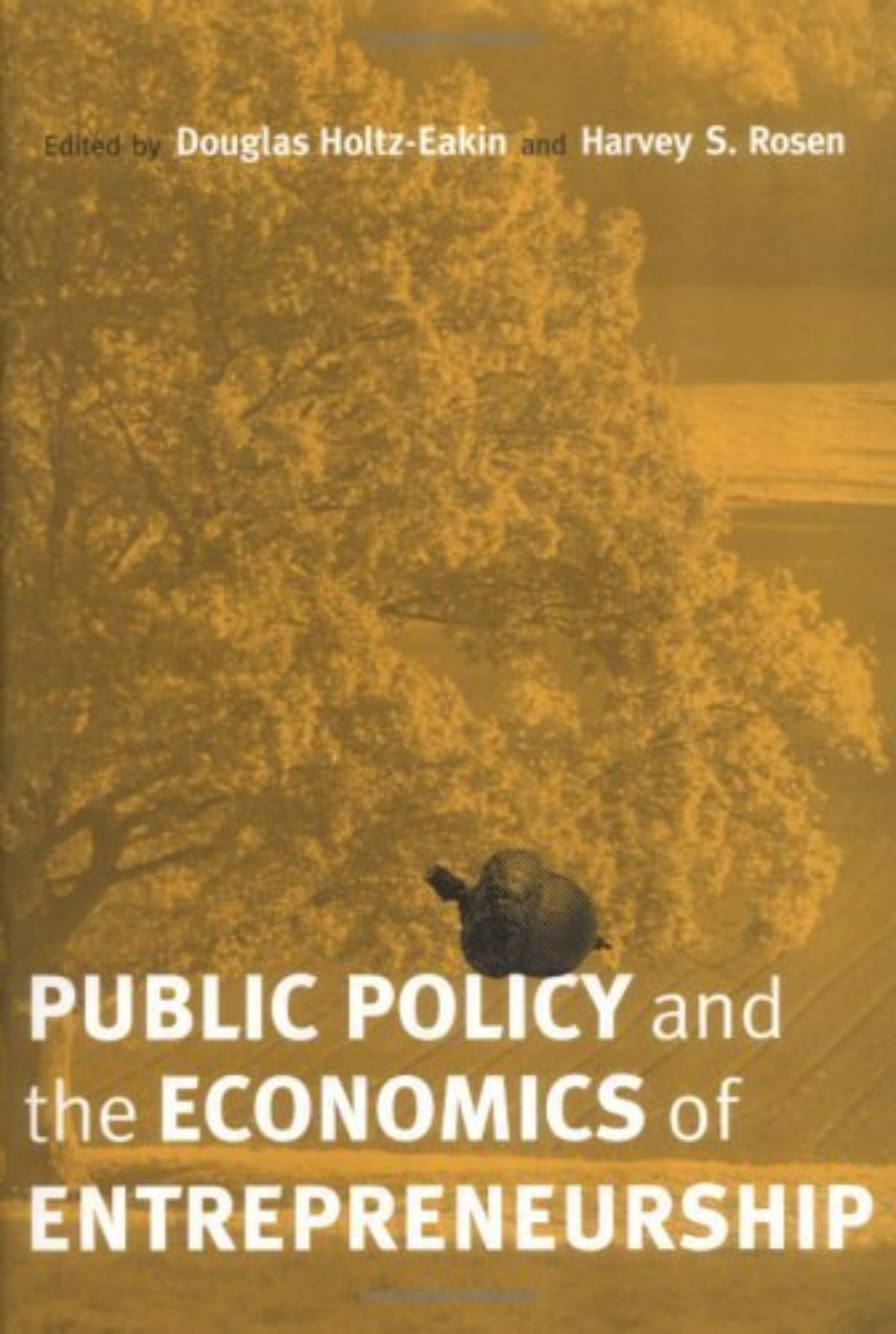
Edited by Douglas Holtz-Eakin and Harvey S. Rosen



PUBLIC POLICY and the ECONOMICS of ENTREPRENEURSHIP

Public Policy and the Economics of Entrepreneurship

Public Policy and the Economics of Entrepreneurship

edited by Douglas Holtz-Eakin
and Harvey S. Rosen

The MIT Press
Cambridge, Massachusetts
London, England

Set in Palatino on 3B2 by Asco Typesetters, Hong Kong. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Public policy and the economics of entrepreneurship / edited by Douglas Holtz-Eakin and Harvey S. Rosen.
p. cm.
Papers presented at a conference held at Syracuse University in April 2001.
Includes bibliographical references and index.
ISBN 0-262-08329-9 (hc. : alk. paper)
1. Entrepreneurship—Congresses. 2. Entrepreneurship—Government policy—United States. 3. Small business—Government policy—United States. 4. Income distribution—United States. I. Holtz-Eakin, Douglas. II. Rosen, Harvey S.
HB615.P83 2004
338′.04′0973—dc21 2003053963

10 9 8 7 6 5 4 3 2 1

Contents

Introduction

In recent years, entrepreneurs have been the focus of considerable discussion among both academics and policy makers. In part, this fascination has reflected the belief that entrepreneurship is a way to obtain upward social and economic mobility. Indeed, much of the literature on entrepreneurship focuses on its benefits to individuals—increases in standard of living, flexibility in hours, and so forth.

However, a good deal of the policy interest derives from the presumption that entrepreneurs provide economy-wide benefits in the forms of new products, lower prices, innovations, and increased productivity. How large are these effects? In a working paper titled Entrepreneurship and Economic Growth: The Proof Is in the Productivity (Center for Policy Research, Syracuse University, 2003), Douglas Holtz-Eakin and Chihwa Kao used a rich panel of state-level data to quantify the relationship between productivity growth (by state and by industry) and entrepreneurship. Specifically, they applied vector autoregression techniques to panel data to determine whether variations in the birth rate and the death rate for firms are related to increases in productivity. They found that shocks to productivity are quite persistent. Thus, to the extent that policies directly raise labor productivity, these effects will be long lasting. Their analysis also suggested that increases in the birth rate of firms lead, after some lag, to higher levels of productivity—a relationship reminiscent of Schumpeterian creative destruction.

In light of such evidence on the economy-wide benefits of entrepreneurship, a critical question is what stance public policy should take. To address this, a group of economists gathered at Syracuse University in April 2001 to discuss issues relating to entrepreneurship and policies to encourage it. This volume contains the papers presented at that conference. Briefly summarized in the remainder of this introduction, they

fall naturally into three main categories: Policies to Encourage Entrepreneurial Activity, Entrepreneurs in Unexpected Places, and Entrepreneurship and Inequality.

Policies to Encourage Entrepreneurial Activity

These days, in the public mind the archetypal entrepreneur is the owner of a small high-tech company. In his chapter, Josh Lerner reviews the motivation behind governmental efforts to finance such firms. Lerner emphasizes the complex environment in which venture capitalists operate. Small high-tech firms are inherently risky. To make matters worse, there are severe information asymmetries—even when business plans are intensively scrutinized, it is difficult for investors to know for sure whether their money is being used sensibly. While various mechanisms exist to help venture capitalists deal with these problems, making the right decisions is very hard. As Lerner documents, they often pick losers.

If it is hard for self-interested venture capitalists to get it right, can the government do better? Economists tend to be wary of the public sector's involvement in such situations. Lerner sets forth and evaluates two arguments for a government role in venture capital markets. The first is that public venture capital programs may play a role by certifying firms to outside investors; the second is that these programs may encourage technological spillovers. However, Lerner cautions that, while it is possible for government officials to identify winners, decisions about which firms to finance still may be based on political rather than economic criteria. Lerner suggests a number of ways to improve the performance of public venture capital efforts, one of which is that public decision makers should closely scrutinize the amount of funding a company has received from prior government sources.

Craig Perry and Harvey Rosen examine another policy focused on entrepreneurs, this one through the federal income tax system. They note that the self-employed are allowed to deduct their health-insurance expenses while wage earners are not. The purpose of this subsidy is to induce the self-employed to purchase medical insurance and hence enjoy better health. However, the link between insurance and health status is not as obvious as it might seem. Some argue that lifestyle issues may ultimately be more important than purchases of medical services. Alternatively, less risk-averse individuals may prefer to eschew health insurance and deal with health expenses out of pocket.

Perry and Rosen investigate whether the relative lack of medical insurance among the self-employed has a detrimental effect on their health. Using cross-sectional data collected in 1996, they find that it does not. For virtually every subjective or objective measure of health status, the self-employed and wage earners are statistically indistinguishable. Further, Perry and Rosen argue that this phenomenon is not due to the fact that individuals who select into self-employment are healthier than wage earners, other things being the same. Hence, the implicit subsidy for health insurance may be an example of a public policy targeted at entrepreneurs that does not have much of an effect.

Whereas the Lerner and Perry-Rosen chapters look at public policies that are targeted directly at entrepreneurs, the chapter by Sandra Black and Philip Strahan reminds us that policies that do not focus explicitly on entrepreneurs can nevertheless have a substantial effect on entrepreneurial activity. Black and Strahan note that the banking industry has experienced major changes over the past 25 years, in part because of changes in regulatory policy. For example, in the early 1980s, ceilings on interest rates were to a large extent removed, allowing banks to compete more vigorously for funds. During the same period, restrictions on banks' ability to expand into new markets were lifted by state initiatives allowing branching across the state and cross-state ownership of bank assets. One consequence of these changes was nationwide consolidation in banking, without any reduction of competition in local banking markets. Using data from the mid 1970s to the mid 1990s, Black and Strahan show that these changes in the structure of banking led to increased lending, and that this increase in the supply of bank loans fueled an increase in the rate of growth of new businesses. In short, although banking deregulation was not driven by a goal of increasing entrepreneurship, it nevertheless generated that spillover.

Entrepreneurs in Unexpected Places

There is a tendency to assume that entrepreneurs carry on their innovative activities only within small businesses. The next two chapters, though, remind us that entrepreneurs operate in a variety of environments, and the policies that are appropriate for encouraging entrepreneurship may depend on the type of organization in which the entrepreneur operates. Frank Lichtenberg's chapter examines a kind of innovation that takes place primarily within large corporations. Lichtenberg notes that what distinguishes the pharmaceutical industry

from other industries is the extent of the government's direct control over innovation. For example, new drugs have to be approved by the government, which requires that they be proven to be safe and effective.

One of the most striking issues Lichtenberg discusses is the relationship between the market value of a firm and its investment in research and development. He notes that econometric studies of R&D indicate that firms invest more when their market value is high, other things being the same. And the market value of firms is based on the *expected* present discounted value of their future net cash flows. Hence, government proposals that are not even ultimately implemented can affect R&D to the extent there is a positive probability that they will be enacted and that they will affect future revenues. Lichtenberg argues that through this mechanism the threat of President Clinton's healthcare reform reduced R&D investment by about 8.8 percent between September 1992 and October 1993. This episode points to the importance of expected economic policy as well as actual policy when one is assessing how government affects entrepreneurial activity.

The chapter by Joseph J. Cordes, C. Eugene Steuerle, and Eric Twombly moves us even farther from traditional notions of entrepreneurship. Indeed, as the authors note, "nonprofit entrepreneurship" seems at first to be an oxymoron. They point out, though, that many successful nonprofit organizations owe their beginning to individuals who exhibited the energy and creativity that we think of as characterizing entrepreneurs.

Cordes, Steuerle, and Twombly begin by painting a statistical portrait of the nonprofit sector and showing that its growth has been driven by the creation of new organizations. Turning to the theory of nonprofit organizations, they note that one important attribute of nonprofit institutions is the "nondistribution constraint": any surplus earned by an entrepreneur cannot be returned to the entrepreneur. The distribution constraint is important because it signals to people that the purpose of the enterprise truly is to do good, and not to serve as a mechanism for disguising entrepreneurial profits. This signal provides an incentive for individuals to contribute to the enterprise. As Cordes, Steuerle, and Twombly note, this phenomenon puts government policies that prevent employees of nonprofit organizations from receiving "excessive" compensation in a new light. Not only do such policies serve the obvious function of preventing abuses of the tax-exempt status of nonprofits; they also provide a legal framework that helps

make the nondistribution constraint credible. And the more credible the constraint, the easier it is for the nonprofit entrepreneurs to raise funds.

Entrepreneurship and Inequality

As we noted above, entrepreneurship is commonly viewed as a good thing not only because of its putative salutary effects on a nation's income, but also because of the distribution of that income. The notion is that entrepreneurship increases income mobility, particularly for minorities. But is it true? This is the question investigated by Robert Fairlie in his chapter. Fairlie uses data from the 1979–1998 National Longitudinal Surveys to examine the earning patterns of young African-American and Hispanic entrepreneurs and to make comparisons to their wage-earning counterparts. He finds some evidence suggesting that young self-employed Hispanic men experience faster earnings growth than young Hispanic wage earners. Young African-American entrepreneurs experience faster earnings growth than young African-American wage earners, but the differences are not statistically significant. Fairlie finds no significant differences at all between the earnings growth of female entrepreneurs and wage earners, but this may be due to small sample sizes. In addition, he finds that minority business owners generally experience more unemployment than wage earners, African-American business owners being the main exception.

Taken together, Fairlie's results provide some limited evidence that entrepreneurship provides a better route for economic advancement among African-American and Hispanic men than wage earning. The evidence for the contribution of self-employment to the economic mobility of African-American and Hispanic women is less promising.

Closely related to income mobility is the distribution of wealth. In particular, some claim that a substantial component of the observed inequality in the distribution of wealth is a consequence of successful entrepreneurship—entrepreneurs who succeed end up with a big portion of the pie. (Think of Bill Gates.) To the extent that this is true, policies aimed at reducing wealth inequality might have undesirable effects on entrepreneurs' incentives to work and save. Carolyn Moehling and Richard Steckel offer a case study of the links between entrepreneurship and the wealth distribution. They use a unique set of data that links information from the 1850–1910 federal censuses to property tax records in the state of Massachusetts. This was a period

in which Massachusetts experienced rapid industrialization and economic growth as well as rising wealth inequality. Moehling and Steckel examine how the distribution of wealth over this period was related to the fraction of the population engaged in entrepreneurial activity, to the share of wealth held by entrepreneurs, and to the inequality in wealth among entrepreneurs. They find that the self-employed held a disproportionate share of wealth in late-nineteenth-century Massachusetts, just as the self-employed do today. But the rise in wealth inequality in the decades leading up to 1900 appears to have been due primarily to growing disparities in the distribution of wealth among those who were not self-employed. To the extent that a similar pattern exists today, the implications for policies to redistribute wealth are rather different than they would be if growing inequality were due to changes in the distribution of wealth between entrepreneurs and non-entrepreneurs.

Taken together, the chapters in this volume demonstrate that entrepreneurship is a many-faceted phenomenon. Designing policy toward entrepreneurs is commensurately complicated. Nevertheless, the standard theoretical and empirical tools of economics can inform both the positive and the normative issues related to public policy toward entrepreneurs.

Public Policy and the Economics of Entrepreneurship

1

When Bureaucrats Meet Entrepreneurs: The Design of Effective "Public Venture Capital" Programs

Josh Lerner

The federal government has played an active role in financing new firms, particularly in high-technology industries, since the Soviet Union's launch of the Sputnik satellite. In recent years, European and Asian nations and many U.S. states have adopted similar initiatives. While these programs' precise structures have differed, the efforts have been predicated on two shared assumptions: (i) that the private sector provides insufficient capital to new firms and (ii) that the government either can identify investments which will ultimately yield high social and/or private returns or can encourage financial intermediaries to do so. In contrast to other government interventions designed to boost economic growth, such as privatization programs, these claims have received little scrutiny by economists.

The neglect of these questions is unfortunate. While the sums of money involved are modest relative to public expenditures on defense procurement or retiree benefits, these programs are very substantial when compared to contemporaneous private investments in new firms. Several examples underscore this point:

- The Small Business Investment Company (SBIC) program led to the provision of more than $3 billion to young firms between 1958 and 1969, more than three times the total private venture capital investment during these years (Noone and Rubel 1970).
- In 1995, the sum of the equity financing provided through and guaranteed by federal and state small business financing programs was $2.4 billion, more than 60 percent of the amount disbursed by traditional venture funds in that year (Lerner 1999). Perhaps more significantly, the bulk of the public funds went to early-stage firms (e.g., those not yet shipping products), which in the past decade had accounted for only about 30 percent of the disbursements by independent venture capital funds.

• Some of America's most dynamic technology companies received support through the SBIC and Small Business Innovation Research (SBIR) programs while still privately held entities, including Apple Computer, Chiron, Compaq, and Intel (Lerner 1999).

• Public venture capital programs have also had a significant impact overseas: e.g., Germany has created about 800 federal and state government financing programs for new firms over the past two decades, which provide the bulk of the financing for technology-intensive start-ups (Organization for Economic Cooperation and Development 1996).

Table 1 summarizes these programs in more detail. This chapter attempts to address this gap, discussing the major challenges that these programs face.

Government programs in this arena have been divided between those efforts that directly fund entrepreneurial firms and those that encourage or subsidize the development of outside investors. In this chapter, I will focus on "public venture capital" initiatives: programs that make equity or equity-like investments in young firms, or encourage other intermediaries to make such investments. In some such programs, such as the Advanced Technology Program and the Small Business Innovation Research programs discussed below, the funds are provided as a contract or outright grant.

While these efforts have proliferated, a consensus as to how to structure these programs remains elusive. While the design of regulatory agencies has been extensively studied from a theoretical and empirical perspective, little work has been done as to how to structure these programs to ensure their greatest effectiveness and to avoid political distortions. As we discuss below, a number of these programs appear predicated on a premise that is at odds with what we know about the financing process: that technologies in entrepreneurial firms can be evaluated in the absence of the consideration of the business prospects of the firm.[1]

This chapter will provide an overview of the motivations for these public efforts, as well as a brief consideration of design questions.

Venture Capitalists and the Financing Challenge

The initial reaction of a financial economist to the argument that the government needs to invest in growth firms is likely to be skepticism. A lengthy literature has highlighted the role of financial intermediaries

in alleviating moral hazard and information asymmetries. Young high-technology firms are often characterized by considerable uncertainty and informational asymmetries, which permit opportunistic behavior by entrepreneurs. Why one would want to encourage public officials instead of specialized financial intermediaries (venture capital organizations) as a source of capital in this setting is not immediately obvious.

The Challenge of Financing Young High-Technology Firms

Before discussing the role of government agencies, it is important to appreciate the challenges that financing young firms pose. I will thus begin by reviewing the types of conflicts that can emerge in these settings.

Jensen and Meckling (1976) demonstrate that agency conflicts between managers and investors can affect the willingness of both debt and equity holders to provide capital. If the firm raises equity from outside investors, the manager has an incentive to engage in wasteful expenditures (e.g., lavish offices) because he does not bear their entire cost. Similarly, if the firm raises debt, the manager may increase risk to undesirable levels. Because providers of capital recognize these problems, outside investors demand a higher rate of return than would be the case if the funds were internally generated.

Even if the manager is motivated to maximize shareholder value, informational asymmetries may make raising external capital more expensive or even preclude it entirely. Myers and Majluf (1984) and Greenwald, Stiglitz, and Weiss (1984) demonstrate that equity offerings of firms may be associated with a "lemons" problem (Akerlof 1970). If the manager is better informed about the investment opportunities of their firms than the investors and acts in the interest of current shareholders, then the manager issues new shares only when the company's stock is overvalued. Indeed, numerous studies have documented that stock prices decline upon the announcement of equity issues, largely because of the negative signal sent to the market.

These information problems have also been shown to exist in debt markets. Stiglitz and Weiss (1981) show that if banks find it difficult to discriminate among companies, raising interest rates can have perverse selection effects. In particular, the high interest rates discourage all but the highest-risk borrowers, so the quality of the loan pool declines markedly. To address this problem, banks may restrict the amount of lending rather than increase interest rates.

Table 1
U.S. public venture capital initiatives, 1958–2000. The table summarizes programs sponsored by state and federal organizations in which equity investments or equity-like grants were made into privately held companies, or into funds that made such investments. If a program had multiple names, we report the name as of 2000. If a program was terminated before 2000, we record its name at the time of termination. If an organization sponsoring a program changed its name, or if responsibility for the program was transferred between organizations, we record the name of the sponsoring organization as of 2000. If the program was terminated before 2000, we record the sponsoring organization at the time of termination.

Sponsoring organization	Program name	Brief description	Span
Small Business Administration	Small Business Investment Company Program	Provides capital to federally sponsored funds that make debt and equity investments in growth firms.	1958–2000
Department of Commerce	State Technical Services Program	Supported various government programs to help high-technology companies (especially new firms).	1965–1969
Department of Housing and Urban Development Model Cities Administration	Venture Capital Development Assistance	Demonstration projects in selected cities financed businesses begun by residents of targeted neighborhoods.	1967–1971
At least 30 states	At least 43 state venture funds or SBIC programs	Make investments into funds supporting new enterprises, which often focus on high-technology firms.	1970–2000
Department of State Agency for International Development	At least 13 developing country venture funds	Provided loans to financial intermediaries that made equity and debt investments in new enterprises in over 30 countries.	1971–1993
Small Business Administration	Specialized Small Business Investment Company Program	Provides capital to federally sponsored funds that make debt and equity investments in growth firms owned by disadvantaged individuals.	1972–2000
Department of Commerce National Bureau of Standards	Experimental Technology Incentives Program	Catalyzed new public programs across agencies to encourage industrial research and venture capital.	1972–1979
National Science Foundation	Federal Laboratories Validation Assistance Experiment	Funded assessments by national laboratory personnel of prototype products and processes developed by entrepreneurs.	1972–1975

National Science Foundation and Small Business Administration	Innovation Centers Experiment	Provided assistance to high-tech entrepreneurs through incubation centers, subsidies, and technical assistance.	1973–1981
Department of Energy Office of Energy-Related Inventions	Energy Related Inventions Program	Provides financing to individual inventors and small firms to commercialize energy-conserving discoveries.	1975–2000
Small Business Administration	Small Business Development Centers Program	Funds university-based centers to assist small businesses and encourage technology transfer.	1976–2000
Department of Commerce	Corporations for Innovation Development Initiative	Designed to fund state and regional corporations to provide equity financing to new firms. Only one such corporation was funded.	1979–1981
Department of Commerce Minority Business Development Agency	Technology Commercialization Program	Financed minority technology-oriented entrepreneurs, as well as centers to assist such entrepreneurs.	1979–1982
At least 15 states	At least 107 business incubators	Provide office and manufacturing space, support services, and often financing to start-up businesses.	1980–1996
Eleven federal agencies	Small Business Innovation Research Program	Provides awards to small technology-oriented businesses. (Also predecessor programs at 3 agencies, 1977–1982.)	1982–2000
Department of Energy Office of Energy Research	At least 6 contractor-organized venture funds	Make equity investments in spin-offs from national laboratories. (Funds organized by prime or sub-contractors at laboratories with Department's encouragement.)	1985–2000
At least 30 states	State Small Business Innovation Research Programs	Makes SBIR-like grants, often in conjunction with federal SBIR awards.	1987–2000
Department of Commerce National Institute of Standards and Technology	Advanced Technology Program	Awards grants to develop targeted technologies to firms and consortia. Some emphasis on small businesses.	1988–2000
Department of Defense Defense Advanced Research Projects Agency	Experimental venture capital investment program	Designed to make investments in private high-technology firms in exchange for equity or royalties. Program only made one investment.	1989–1991

Table 1
(continued)

Sponsoring organization	Program name	Brief description	Span
Department of State Agency for International Development	Enterprise Fund Program	Oversees 12 federally funded venture funds investing in Eastern Europe, the former Soviet Union, and Africa.	1990–2000
Overseas Private Investment Corporation	Venture capital fund guarantees	Guarantees full or partial return of capital to investors in at least 16 private venture funds in developing countries.	1990–2000
Department of Housing and Urban Development Community Relations & Involvement Office	Tenant Opportunity Program	Funds new businesses and other initiatives by public housing residents (other aspects of program had begun in 1987).	1993–2000
Department of Energy Office of the Undersecretary	Defense Programs Small Business Initiative	Provides funding, technological assistance, and national laboratory access to small high-technology businesses.	1993–2000
Eleven federal agencies	Small Business Technology Transfer Program	Finances cooperative research projects between small high-technology firms and nonprofit research institutions.	1994–2000
Department of Defense Cooperative Threat Reduction Program	Defense Enterprise Fund	Finances an independent venture fund investing in defense conversion projects in the former Soviet Union.	1994–2000
Department of the Treasury	Community Development Financial Institutions Fund	Invests in and provides assistance to community development venture capital and loan funds.	1995–2000
Department of Defense	"Fast Track" Program	Provides 4:1 matching funds for private financing raised by SBIR awardees.	1995–2000
Department of Agriculture Rural Business and Cooperative Development Service	Intermediary Relending Program (as amended)	Permits program managers to guarantee returns of investors in rural venture funds.	1997–2000
Central Intelligence Agency	In-Q-It	Invests in information technology-related companies whose products may have national security applications.	1999–2000

These problems in the debt and equity markets are a consequence of the information gaps between the entrepreneurs and investors. If the information asymmetries could be eliminated, financing constraints would disappear. Financial economists argue that specialized financial intermediaries can address these problems. By intensively scrutinizing firms before providing capital and then monitoring them afterwards, they can alleviate some of the information gaps and reduce capital constraints.

Responses by Venture Capitalists

The financial intermediary that specializes in funding young high-technology firms is the venture capital organization. The first modern venture capital firm, American Research and Development (ARD), was formed in 1946 by MIT president Karl Taylor Compton, Harvard Business School professor Georges F. Doriot, and local business leaders. A small group of venture capitalists made high-risk investments in emerging companies that were formed to commercialize technology developed for World War II. The success of the investments ranged widely: almost half of ARD's profits during its 26-year existence as an independent entity came from its $70,000 investment in Digital Equipment Company (DEC) in 1957, which grew in value to $355 million. Because institutional investors were reluctant to invest, ARD was structured as a publicly traded closed-end fund and marketed mostly to individuals (Liles 1977). The few other venture organizations begun in the decade after ARD's formation were also structured as closed-end funds.

The first venture capital limited partnership, Draper, Gaither, and Anderson, was formed in 1958. Imitators soon followed, but limited partnerships accounted for a minority of the venture pool during the 1960s and the 1970s. Most venture organizations raised money either through closed-end funds or small business investment companies (SBICs), federally guaranteed risk capital pools that proliferated during the 1960s. While investor demand for SBICs in the late 1960s and the early 1970s was strong, incentive problems ultimately led to the collapse of the sector.[2] The annual flow of money into venture capital during its first three decades never exceeded a few hundred million dollars and usually was substantially less.

The activity in the venture industry increased dramatically in the late 1970s and the early 1980s. Industry observers attributed much of the

shift to the U.S. Department of Labor's clarification of ERISA's "prudent man" rule in 1979. Before that year, the Employee Retirement Income Security Act (ERISA) limited pension funds from investing substantial amounts of money in venture capital or other high-risk asset classes. These years also saw the emergence of the limited partnership as the dominant organizational form for venture funds. Financial economists argue that these structures can alleviate the incentive and valuation problems often encountered in publicly traded funds. (See, e.g., Gompers and Lerner 1999b.)

The subsequent years saw both very good and trying times for venture capitalists. On the one hand, during the 1980s and the 1990s venture capitalists backed many of the most successful high-technology companies, including Apple Computer, Cisco Systems, Genentech, Netscape, and Sun Microsystems. A substantial number of service firms (including Staples, Starbucks, and TCBY) also received venture financing. At the same time, commitments to the venture capital industry were very uneven. The annual flow of money into venture funds increased by a factor of ten during the early 1980s, peaking at just under 6 billion 1996 dollars. From 1987 through 1991, however, fund raising declined steadily, reflecting the low returns from overinvestment in certain sectors.[3] Over the past decade, the pattern has been reversed. In 2000, a record year for fund raising, nearly $70 billion was raised by venture capitalists. This process of rapid growth and decline has created a great deal of instability in the industry. (These data are from Gompers and Lerner 2001.)

To address the information problems that preclude other investors in small high-technology firms, the partners at venture capital organizations employ a variety of mechanisms. Business plans are intensively scrutinized: of those firms that submit business plans to venture capital organizations, historically only 1 percent have been funded (Fenn, Liang, and Prowse 1995).

In evaluating a high-technology company, the venture capitalists employ several criteria. To be sure, the promise of the firm's technology is important. But this evaluation is inexorably linked with the evaluation of the firm's management. Venture capitalists are well aware that many promising technologies do not ultimately fill market needs. As a result, most place the greatest emphasize on the experience and flexibility of the management team and the size of the potential market. Even if the market does not evolve as predicted, with a sophisticated team the firm may be able to find an attractive opportu-

nity. The decision to invest is frequently made conditional on the identification of a syndication partner who agrees that this is an attractive investment (Lerner 1994). In exchange for their capital, the venture capital investors demand preferred stock with numerous restrictive covenants and representation on the board of directors.

Once the decision to invest is made, the venture capitalists frequently disburse funds in stages. Managers of these venture-backed firms often only raise a small fraction of the funds initially and are forced to return repeatedly to their financiers for additional capital in order to ensure that the money is not squandered on unprofitable projects. In addition, venture capitalists intensively monitor managers, often contacting firms on a daily basis and holding monthly board meetings during which extensive reviews of every aspect of the firm are conducted. (Various aspects of the oversight role played by venture capitalists are documented in Gompers and Lerner 1999b.)

It is important to note that, even with these many mechanisms, the most likely primary outcome of a venture-backed investment is failure, or at best modest success. Gompers (1995) documents that out of a sample of 794 venture capital investments made over three decades, only 22.5 percent ultimately succeeded in going public, the avenue through which venture capitalists typically exit their successful investments.[4] Similar results emerge from Huntsman and Hoban's (1980) analysis of the returns from 110 investments by three venture capital organizations. About one in six investments was a complete loss, while 45 percent were either losses or simply broke even. The elimination of the top-performing 9 percent of the investments was sufficient to turn a 19 percent gross rate of return into a negative return.

In short, the environment in which venture organizations operate is extremely difficult. Difficult conditions that have frequently deterred or defeated traditional investors such as banks can be addressed by the mechanisms that are bundled with the venture capitalists' funds. These tools have led to venture capital organizations emerging as the dominant form of equity financing for privately held technology-intensive businesses.[5]

Rationales for Public Programs

At the same time, there are reasons to believe that, despite the presence of venture capital funds, there still might be a role for public venture capital programs. In this section, I assess these claims. I highlight two

arguments: that public venture capital programs may play an important role by certifying firms to outside investors, and that these programs may encourage technological spillovers.

The Certification Hypothesis

A growing body of empirical research suggests that new firms, especially technology-intensive ones, may receive insufficient capital to fund all positive net present value projects due to the information problems discussed in the previous section.[6] If public venture capital awards could certify that firms are of high quality, these information problems could be overcome and investors could confidently invest in these firms.

As discussed above, venture capitalists specialize in financing these types of firms. They address these information problems through a variety of mechanisms. Many of the studies that document capital-raising problems examine firms during the 1970s and the early 1980s, when the venture capital pool was relatively modest in size. Since the pool of venture capital funds has grown dramatically in recent years (Gompers and Lerner 1998), even if small high-technology firms had numerous value-creating projects that they could not finance in the past, one might argue that it is not clear this problem remains today. While there may have once been a role for government certification, it may not still be there today.

A response to this argument emphasizes the limitations of the venture capital industry. Venture capitalists back only a tiny fraction of the technology-oriented businesses begun each year. In 2000, a record year for venture disbursements, just over 2,200 U.S. companies received venture financing for the first time.[7] Yet the Small Business Administration estimates that in recent years about 1 million new businesses have started up annually.[8] Furthermore, private venture funds have concentrated on a few industries: for instance, in 2000, fully 46 percent of the funding went to Internet-related companies. More generally, 92 percent of the funding went to firms specializing in information technology and health care. Thus, many promising firms in other industries are *not* attracting venture capitalists' notice, perhaps reflecting "herding" by venture capitalists into particular areas, a problem that finance theory suggests affects institutional investors (Devenow and Welch 1996). If government programs can identify and support technological areas that are neglected by venture capitalists, they might

provide the "stamp of approval" these high-potential, underfunded firms need to succeed.

But if government officials are going to address these problems, they will need to be able to overcome the many information asymmetries and identify the most promising firms. Otherwise, as de Meza (2002) argues, these efforts are likely to be counter-productive. Is it reasonable to assume that government officials can overcome these problems while private sector financiers cannot? Certainly, this possibility is not implausible. For instance, specialists at the National Institutes of Health or the Department of Defense may have considerable insight into which biotechnology or advanced materials companies are the most promising, while the traditional financial statement analysis undertaken by bankers would be of little value. In general, the certification hypothesis suggests that these signals provided by government awards are likely to be particularly valuable in technology-intensive industries where traditional financial measures are of little use.[9]

The Presence of R&D Spillovers

A second rationale emerges from the literature on R&D spillovers. Public finance theory emphasizes that subsidies are an appropriate response in the case of activities that generate positive externalities. Such investments as R&D expenditures and pollution control equipment purchases may have positive spillovers that help other firms or society as a whole. Because the firms making the investments are unlikely to capture all the benefits, public subsidies may be appropriate.

An extensive literature (reviewed in Griliches 1992 and Jaffe 1996) has documented the presence of R&D spillovers. These spillovers take several forms. For instance, the rents associated with innovations may accrue to competitors who rapidly introduce imitations, developers of complementary products, or to the consumers of these products. Whatever the mechanism of the spillover, however, the consequence is the same: the firm invests below the social optimum in R&D.

After reviewing a wide variety of studies, Griliches estimates that the gap between the private and social rate of return is substantial: the gap is probably equal to between 50 percent and 100 percent of the private rate of return. While few studies have examined how these gaps vary with firm characteristics, a number of case-based analyses (Jewkes et al. 1958; Mansfield et al. 1977) suggest that spillover problems are particularly severe among small firms. These organizations

may be particularly unlikely to effectively defend their intellectual property positions or to extract most of the rents in the product market.

Limitations of "Public Venture Capital" Programs

Even if spillover problems are substantial or government officials can successfully identify promising small firms, these efforts may not solve these financing problems. An extensive political economy and public finance literature has emphasized the distortions that may result from government subsidies as particular interest groups or politicians seek to direct subsidies in a manner that benefits themselves. As articulated by Olson (1965) and Stigler (1971), and as formally modeled by Peltzman (1976) and Becker (1983), the theory of regulatory capture suggests that direct and indirect subsidies will be captured by parties whose joint political activity, such as lobbying, is not too difficult to arrange (i.e., when "free riding" by coalition members is not too large a problem).

These distortions may manifest themselves in several ways. One possibility (Eisinger 1988) is that firms may seek transfer payments that directly increase their profits. Politicians may acquiesce in such transfers in the case of companies that are politically connected. A more subtle distortion is discussed by Cohen and Noll (1991) and Wallsten (1996): officials may seek to select firms based on their likely success and fund them regardless of whether the government funds are needed. In this case, they can claim credit for the firms' ultimate success even if the marginal contribution of the public funds was very low.

The presence of these distortions is likely to vary with program design. Consider the case of the SBIR program. The Small Business Innovation Development Act, enacted by Congress in July 1982, established the SBIR program. The program mandated that all federal agencies spending more than $100 million annually on external research set aside 1.25 percent of these funds for awards to small businesses. When the program was reauthorized in 1992, Congress increased the size of the set-aside to 2.5 percent. In 1997, this represented annual funding of about $1.1 billion.

While the eleven federal agencies participating in the program are responsible for selecting awardees, they must conform to the guidelines stipulated by the act and the U.S. Small Business Administration

(SBA). Awardees must be independently owned, for-profit firms with fewer than 500 employees, at least 51 percent owned by U.S. citizens or permanent residents. Promising proposals are awarded Phase I awards (originally no more than $50,000, today $100,000 or smaller), which are intended to allow firms to determine the feasibility of their ideas. (Typically about ten Phase I applications are received for every award made.) Approximately one-half of the Phase I awardees are then selected for the more substantial Phase II grants. Phase II awards of at most $750,000 (originally, $500,000) are transferred to the small firm as a contract or grant. The government receives no equity in the firm and does not own the intellectual property that the firm develops with these funds.

In particular, one of the reasons that has been suggested for why the SBIR program is relatively effective (as documented in Lerner 1999) is that the decision makers are highly dispersed. In particular, the federal program managers are scattered across many sub-agencies and are responsible for many other tasks as well. Thus, the costs of identifying and influencing these decision makers are high. In programs where a central group makes highly visible awards, the dangers of political distortions are likely to be higher.

The Challenge of Program Design

An immense literature in regulatory economics and industrial organization has considered the structure of regulatory bodies. The different ways in which regulators can monitor and shape industry behavior—and Congress can in turn monitor the regulators—has been explored in detail. (For an overview, see Laffont and Tirole 1993.)

Other areas of interactions between government officials and firms, however, have been much less well scrutinized. Not only is the theoretical foundation much less well developed, but the empirical literature is at a much earlier stage. (For an overview of the current state of empirical research, see Klette, Moen, and Griliches 2000.) Thus, our observations must be necessarily tentative in nature.

The design of efforts to assist high-technology entrepreneurs in one program, the Advanced Technology Program (ATP) run by the Department of Commerce, was examined in Gompers and Lerner 1999a. The object of this program is to fund generic pre-commercial technology, whether developed by single firms or joint ventures. The

awards are made in the form of contracts, typically for sums between a few hundred thousand and several million dollars. Between its inception in 1990 and 1997, the program awarded nearly a billion dollars in research and development funding to approximately 300 technology-based projects conducted by American companies and industry-led joint ventures.

While the ATP program is not mandated to fund firms of any particular size, it has become a major funder of small businesses. From 1990 to 1997, 36 percent of ATP funding went to small businesses. An additional 10 percent went to joint ventures led by small businesses.

In particular, we asked how the public sector could interact with the venture community and other providers of capital to entrepreneurial firms in order to most effectively advance the innovation process. Reflecting the early state of knowledge and lack of a theoretical foundation, we did not analyze these challenging questions through a large-sample analysis. Rather, we relied on seven case studies of ATP firms, complemented by a review of the secondary literature.

As part of this analysis, we highlighted four key recommendations, which are likely to be more generally applicable to public venture capital programs. In this section, we will review each of these recommendations. I particularly highlight our final recommendation, which challenges the premise that technologies in entrepreneurial firms can be evaluated in the absence of the consideration of the business prospects of the firm.

First, there is a strong need for public officials to invest in building relationships with and an understanding of the U.S. venture capital industry. Financing small entrepreneurial firms is exceedingly challenging. The venture capital industry employs a variety of important mechanisms to address these challenges, which empirical evidence suggests are quite effective. Because of the magnitude and success of venture capital financing, it is important that administrators view their actions in the context of this financial institution.

A corollary to the first point is that public venture capital investments should be made with an eye to the narrow technological focus and uneven levels of independent investments. As noted above, venture investments tend to be very focused on a few areas of technology that are perceived to have great potential. Increases in venture fund raising—which are driven by factors such as shifts in capital gains tax rates—appear more likely to lead to more intense price competition for transactions within an existing set of technologies than to greater

diversity in the types of companies funded. (For a discussion of these patterns, see Gompers and Lerner 2000.) Administrators may wish to respond to these industries' conditions by (i) focusing on technologies which are not currently popular among venture investors and (ii) providing follow-on capital to firms already funded by venture capitalists during periods when venture inflows are falling.

A third point is that federal officials must appreciate the need for flexibility that is central to the venture capital investment process. Venture capitalists make investments in young firms in settings with tremendous technological, product market, and management uncertainties. Rather than undertaking the (often impossible) task of addressing all the uncertainties in advance, they remain actively involved after the investment, using their contractually specified control rights to guide the firm. These changes—which often involve shifts in product market strategy and the management team—are an integral part of the investment process. In our case studies, it appeared that ATP administrators too often view these shifts as troubling indications that awardees are deviating from their plan, rather than as a natural part of their evolution.[10]

Fourth, just as the venture capital community carefully analyzes the track record of entrepreneurs they are considering funding, government officials should examine the track record of the firms receiving public venture awards. As it is now, public venture capital programs are often characterized by a considerable number of underachieving firms.[11] In particular, certain company characteristics—attributes that may not be adequately considered in the selection process of these programs—appear to be highly correlated with a company's ability to achieve its research and commercialization goals. These include the experience of the management team, the presence of a clear product market strategy, and a strong desire to seek private financing. By devising new methods to search for such factors, government officials would be better able to distinguish between high-performing and underachieving firms.

Our research indicates that a prevalent characteristic among underachieving companies is the existence of research grants from numerous government sources, with few, if any, tangible results to show from previous R&D awards. Because a lack of results can easily be attributed to the high-risk nature of technology development, many of these companies can avoid accountability indefinitely. These government grant-oriented research organizations are able to drift from one federal

contract to the next. For such companies, it appeared that public venture capital funds were treated in exactly the same manner as other government research grants: it did not appear that ATP funding showed any notable returns or that the unique program goals were well served.

Adding to the problem is the fact that companies with substantial government grant experience appear to have several advantages over other firms when applying for future public awards. Past grants, regardless of project outcomes, help a company gain legitimacy in a particular area of research, as well as acquire the equipment and personnel needed to do future work. There is also a tendency for some government programs to try to "piggyback" on other government programs, hoping to leverage their grant dollars. In addition, firms gain considerable insight into the grant application process with each proposal they submit. These firms consequentially often have a greater chance of being awarded future government grants than other firms. The end result can be a stream of government funding being awarded to companies that consistently underachieve.

To level the playing field, our research suggests that public venture capital should more closely scrutinize the amount of funding a company has received from prior government sources. A greater number of underachieving firms could be weeded out if government officials conducted a more comprehensive evaluation of a company's past performance and examined the tangible progress attributable to each government grant the firm has received. Moreover, large inflows of prior government funding without significant product development may indicate that a particular company is unlike to generate significant commercialization of new technologies.

Another telltale characteristic of underachieving firms was the existence of factors outside the scope of the publicly funded projects that undermined their ability to successfully complete and later commercialize government-funded technology. Legal troubles, for instance, can divert substantial amounts of human and financial resources away from a company's R&D projects and even cause dramatic changes in the size and structure of the company. And when a firm is ready to commercialize its technology, the liability concerns associated with pending legal battles will often drastically impair the company's ability to attract venture capital investment dollars.

For early-stage companies, additional limiting factors frequently involve managers who lack experience in running small companies.

Although some of these managers may have accumulated business experience as consultants or as members of large organizations, the successful operation of early-stage companies can demand very different management skills. It thus comes as no surprise that when venture capitalists sink substantial funds in a company, they will often place their own hand-picked manager in charge—typically an individual who has already been successful in managing an early-stage company in a similar industry. Because much of the skills needed for managing startup companies comes through experience, the existence of managers who do not have this background can significantly undermine a company's ability to succeed.

In a broader context, each of these performance-undermining factors emphasizes the need for government officials to critically evaluate whether a particular company is a viable vehicle for accomplishing its commercialization goals. This goes far beyond a simple assessment of the feasibility of a business plan. In fact, many of these potentially limiting factors will not even be discussed in a company's written proposal to the government. It is tempting, of course, to attribute the failures resulting from such factors to the high-risk nature of the technology. But to a large extent, companies exhibiting a high potential for underachievement could be more thoroughly weeded out by placing a greater emphasis on these factors during the selection process. The R&D project itself may be high-risk, but the risks of turning the technology into a product should be minimized. Regardless of how innovative or enabling a technology may be, or how well a business plan is constructed, if these undermining factors are present, a company will be hard pressed to succeed. In short, the claim that technological projects can be assessed in entrepreneurial firms without consideration of business issues is profoundly mistaken.

A broader implication is that administrators of public venture capital programs must think carefully about the validity of the concept of "pre-commercial research" in an entrepreneurial setting. An extensive body of entrepreneurship research has highlighted the unpredictability of the entrepreneurial process. Very few entrepreneurs, whether in high- or low-technology settings, commercialize what they initially set to develop in their original time-frame. Rather, successful entrepreneurs gather signals from the marketplace in response to their initial efforts, and adjust their plans accordingly. Once they identify an opportunity, they move very rapidly to take advantage of it before major corporations can respond. Yet many federal agencies, leery of being seen

as "picking winners," push entrepreneurs to devote Advanced Technology Program funds to purely pre-commercial research. This may lead them to ignore an essential source of information: i.e., feedback from customers. Even more detrimental are those instances where a company—having identified an attractive commercial opportunity—is afraid to rapidly pursue it, lest they jeopardize their public funds (on which they are relying as a key source of financing) on the grounds that they are pursuing commercial research. While well intentioned, such policies may have the perverse effect of punishing success. One potential change would be to allow firms that rapidly commercialize publicly funded projects to use the funds to pursue another project.

Conclusions

This chapter has examined the design of public venture capital programs. Much is still to be learned about the design of these programs. While the literature on the design of regulatory agencies and the problem of political distortions in subsidy programs has yet to consider public venture capital programs in much depth, one can be optimistic that this will be a topic of increasing interest to researchers. With the help of these theoretical insights—as well as the willingness of program administrators to encourage dispassionate analyses of their strengths and weaknesses—our ability to say more about the design of these programs should grow.

That being said, the many difficulties suggest the need for caution in proceeding with these programs. Indeed, it has been suggested that public policy may be far more effective in encouraging venture capital activity by addressing the demand for such funds—through such steps as encouraging academic R&D and cutting the tax rates that entrepreneurs pay on capital gains—rather than by directly boosting the supply of such funds (Gompers and Lerner 1998). The many hazards that these public programs face, as discussed above, suggest why efforts to address directly the supply of venture financing may be ineffective.

Acknowledgments

This is based in part on conversations with Zoltan Acs, Ken Flamm, Paul Gompers, Doug Holtz-Eakin, Adam Jaffe, Bill Sahlman, Greg Udell, and Allan Young. Participants in the workshops at Harvard

University, Syracuse University, and the University of Warwick provided helpful comments. Parts of this article are adapted from Lerner 1998 and from Gompers and Lerner 1999a. Financial support was provided by Harvard Business School's Division of Research.

Notes

1. Several limitations—necessitated by the limited available space—should be acknowledged up front. First, I will focus on the experience of the United States. Second, I will focus on government efforts to directly finance young firms, rather than on those that subsidize venture capital organizations, as has been done in the Israeli Yozma program or the BioRegio effort in Germany.

2. In particular, many SBICs made investments in ineffective or corrupt firms. Observers noted that SBIC managers' incentives to screen or monitor portfolio firms was greatly reduced by the presence of government guarantees that limited their exposures to unsuccessful investments.

3. The measurement of the riskiness of venture investments pose many challenges, as Gompers and Lerner (1997) discuss. As a result, there has not been a satisfactory systematic effort to calculate the risk-adjusted return for private equity over this period.

4. A Venture Economics study (Ross and Isenstein 1988) finds that a $1 investment in a firm that goes public provides an average cash return to venture capitalists of $1.95 in excess of the initial investment, with an average holding period of 4.2 years. The next best alternative, a similar investment in an acquired firm, yields a cash return of only 40 cents over a 3.7-year mean holding period.

5. While evidence regarding the financing of these firms is imprecise, Freear and Wetzel's (1990) survey suggests that venture capital accounts for about two-thirds of the external equity financing raised by privately held technology-intensive businesses from private-sector sources.

6. The literature on capital constraints (reviewed in Hubbard 1998) documents that an inability to obtain external financing limits many forms of business investment. Hall (1992), Hao and Jaffe (1993), and Himmelberg and Petersen (1994) show that capital constraints appear to limit research-and-development expenditures, especially in smaller firms, though the limits may be less binding than those on capital expenditures. Holtz-Eakin, Joulfaian, and Rosen (1994a,b) discuss these constraints on the survival of entrepreneurial firms.

7. Statistics on venture capital financing are available at http://www.nvca.org.

8. See http://www.sba.gov.

9. Another possibility, of course, is that the government could provide certification without funding, e.g., by selecting a small number of firms each year for prizes. Whether these signals would be as credible or whether government officials would approach this assignment with sufficient seriousness remains open to question.

10. Of course, since the goal of the program is to fund companies that are developing socially beneficial technologies, there is a need for program officers to be alert for firms

that radically shift their objectives. For instance, one supercomputer firm devoted considerable resources after receiving an ATP award to developing an e-commerce program, at a time when such technologies were receiving extensive funding from independent venture capitalists.

11. The presence of "SBIR mills" that have won large numbers of awards by cultivating relationships with federal officials is a manifestation of this phenomenon in another federal program (Lerner 1999).

References

Akerlof, G. A. 1970. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84: 488–500.

Becker, G. S. 1983. A theory of competition among pressure groups for political influence. *Quarterly Journal of Economics* 98: 371–400.

Cohen, L. R., and R. G. Noll, eds. 1991. *The Technology Pork Barrel*. Brookings Institution.

de Meza, D. 2000. Overlending. *Economic Journal* 112 (477) (February): F17–F31.

Devenow, A., and I. Welch. 1996. Rational herding in financial economics. *European Economic Review* 40: 603–615.

Eisinger, P. K. 1988. *The Rise of the Entrepreneurial State: State and Local Economic Development Policy in the United States*. University of Wisconsin Press.

Fenn, G. W., N. Liang, and S. Prowse. 1995. The Economics of the Private Equity Market. Board of Governors of Federal Reserve System.

Freear, J., and W. E. Wetzel Jr. 1990. Who bankrolls high-tech entrepreneurs? *Journal of Business Venturing* 5: 77–89.

Gompers, P. A. 1995. Optimal investment, monitoring, and the staging of venture capital. *Journal of Finance* 50: 1461–1489.

Gompers, P. A., and J. Lerner. 1997. Risk and reward in private equity investments: The challenge of performance assessment. *Journal of Private Equity* 1, winter: 5–12.

Gompers, P. A., and J. Lerner. 1998. What drives venture capital fund raising? *Brookings Papers on Economic Activity: Microeconomics*: 149–192.

Gompers, P. A., and J. Lerner. 1999a. Capital Formation and Investment in Venture Markets. Report GCR–99–784, Advanced Technology Program, National Institutes of Standards and Technology, U.S. Department of Commerce.

Gompers, P. A., and J. Lerner. 1999b. *The Venture Capital Cycle*. MIT Press.

Gompers, P. A., and J. Lerner. 2000. Money chasing deals? The impact of fund inflows on the valuation of private equity investments. *Journal of Financial Economics* 55: 281–325.

Gompers, P. A., and J. Lerner. 2001. *The Money of Invention*. Harvard Business School Press.

Greenwald, B. C., J. E. Stiglitz, and A. Weiss. 1984. Informational imperfections in the capital market and macroeconomic fluctuations. *American Economic Review* 74 (2, Papers and Proceedings of the Ninety-Sixth Annual Meeting of the American Economic Association) (May): 194–199.

Griliches, Z. 1992. The search for R&D spillovers. *Scandinavian Journal of Economics* 94 (Supplement): S29–S47.

Hall, B. H. 1992. Investment and Research and Development at the Firm Level: Does the Source of Financing Matter? Working paper 409b, National Bureau of Economic Research.

Hao, K. Y., and A. B. Jaffe. 1993. Effect of liquidity on firms' R&D spending. *Economics of Innovation and New Technology* 2: 275–282.

Himmelberg, C. P., and B. C. Petersen. 1994. R&D and internal finance: A panel study of small firms in high-tech industries. *Review of Economics and Statistics* 76: 38–51.

Holtz-Eakin, D., D. Joulfaian, and H. S. Rosen. 1994a. Entrepreneurial decisions and liquidity constraints. *RAND Journal of Economics* 23: 334–347.

Holtz-Eakin, D., D. Joulfaian, and H. S. Rosen. 1994b. Sticking it out: Entrepreneurial survival and liquidity constraints. *Journal of Political Economy* 102: 53–75.

Hubbard, R. G. 1998. Capital-market imperfections and investment. *Journal of Economic Literature* 36: 193–225.

Huntsman, B., and J. P. Hoban Jr. 1980. Investment in new enterprise: Some empirical observations on risk, return, and market structure. *Financial Management* 9 (summer) 44–51.

Jaffe, A. B. 1996. Economic Analysis of Research Spillovers: Implications for the Advanced Technology Program. Report GCR 97-708 Advanced Technology Program, National Institute of Standards and Technology, U.S. Department of Commerce.

Jensen, M. C., and W. H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

Jewkes, J., D. Sawers, and R. Stillerman. 1958. *The Sources of Invention*. St. Martin's Press.

Klette, T., J. Moen, and Z. Griliches. 2000. Do subsidies to commercial R&D reduce market failures? Microeconometric evaluation studies. *Research Policy* 29: 471–495.

Laffont, J.-J., and J. Tirole. 1993. *A Theory of Incentives in Procurement and Regulation*. MIT Press.

Lerner, J. 1994. The syndication of venture capital investments. *Financial Management* 23, autumn: 16–27.

Lerner, J. 1998. "Angel" financing and public policy: An overview. *Journal of Banking and Finance* 22: 773–783.

Lerner, J. 1999. The government as venture capitalist: The long-run impact of the SBIR program. *Journal of Business* 72: 285–318.

Liles, P. 1977. *Sustaining the Venture Capital Firm*. Management Analysis Center.

Mansfield, E., J. Rapoport, A. Romeo, S. Wagner, and G. Beardsley. 1977. Social and private rates of return from industrial innovations. *Quarterly Journal of Economics* 91: 221–240.

Myers, S. C., and N. Majluf. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13: 187–221.

Noone, C. M., and S. M. Rubel. 1970. *SBICs: Pioneers in Organized Venture Capital*. Capital.

Olson, M. 1965. *The Logic of Collective Action*. Harvard University Press.

Organization for Economic Cooperation and Development. 1996. Venture capital in OECD countries. *Financial Market Trends* 63.

Peltzman, S. 1976. Towards a more general theory of regulation. *Journal of Law and Economics* 19: 211–240.

Ross, P. W., and S. Isenstein. 1988. *Exiting Venture Capital Investments.* Venture Economics, Inc.

Stigler, G. 1971. The theory of economic regulation. *Bell Journal of Economics* 2: 3–21.

Stiglitz, J. E., and A. Weiss. 1981. Credit rationing in markets with incomplete information. *American Economic Review* 71: 393–410.

Wallsten, S. J. 1996. The Small Business Innovation Research Program: Encouraging Technological Innovation and Commercialization in Small Firms. Unpublished paper, Stanford University.

2

The Self-Employed Are Less Likely to Have Health Insurance Than Wage Earners. So What?

Craig William Perry and
Harvey S. Rosen

A persistent public policy concern in the United States is that so many Americans—currently more than 39 million (U.S. Census Bureau 2001)—lack health insurance. Indeed, this was a major issue in the 2000 presidential campaign. Republican George W. Bush proposed a tax credit of up to $2,000 per family to help low-income workers buy insurance; Democrat Al Gore suggested expanding the federal-state health plan for children. Although their approaches differed considerably, both parties clearly viewed the lack of health insurance as a serious problem.

Within the ranks of the uninsured, the self-employed have been the objects of particular concern. Owners of small businesses do indeed have lower rates of health insurance than wage earners. Only 69 percent of those under 63 years of age had any coverage in 1996 as compared with 81.5 percent of wage earners, according to our tabulations from the Medical Expenditure Panel Survey. The principal public policy response to this situation has been to subsidize self-employed individuals' purchases of health insurance through the personal income tax. Starting in 1998, self-employed workers were allowed to deduct 45 percent of their health insurance premiums; this deduction grew to 60 percent in 2000 and 70 percent in 2002. Effective in 2003 the entire premium is deductible for health-insurance purchased through a self-employed person's business.[1]

Implicit in the support behind this type of policy is the assumption that health insurance affects health outcomes—if an individual has health insurance, he or she is more likely to be healthy. Certainly, at face value, this seems to make sense. Health insurance reduces the cost to individuals of a variety of medical services, increasing consumption of these services and presumably improving health, ceteris paribus. However, the link between insurance and health status is not as

obvious as it might seem. While most researchers agree that socioeconomic status has a significant effect on health, some argue that insurance does little to contribute to these differentials.[2] Some have argued, for example, that lifestyle issues may ultimately be more important than purchases of medical services (Fuchs 1998). Alternatively, relatively less risk-averse individuals may prefer to eschew health insurance and deal with health expenses out of pocket. Thus, it is not obvious whether the health of the self-employed suffers because of their relative lack of health insurance. In fact, we know of no research that looks at the link between insurance status and health for the self-employed. The purpose of this chapter is to investigate whether the lack of health insurance among the self-employed has a detrimental effect on their health. The centerpiece of the study is a statistical analysis of differences between the self-employed and wage earners in a variety of health status measures.

Previous Literature

The determinants of health status have been the subject of a number of studies. A central issue in this literature is the effect of income or wealth on health. The general finding is that there is a positive relationship in the data between health status and economic resources.[3] See, for example, Menchik 1993, Ettner 1996, Smith and Kington 1997, and Smith 1999.[4] While these studies look at the effects of a variety of other economic and demographic characteristics on health, none examines possible health differences between the self-employed and wage earners.

Two related literatures are relevant to this chapter. First is the health insurance demand literature, in which several studies have noted that the tax treatment of insurance differs for wage earners and the self-employed, and take advantage of this fact to estimate the price elasticity of demand for insurance (Monheit and Harvey 1993; Gruber and Poterba 1994). Their results show that lowering the effective price of insurance does indeed increase the probability that a self-employed individual will buy insurance. The question remains, however, whether having the insurance makes any difference to their health.

The second literature focuses on links among health insurance, health services utilization, and health outcomes. Currie and Gruber (1995) examine health insurance eligibility, utilization, and children's health. They find that utilization increases with insurance eligibility,

but has no effect on a set of paternal-reported health status measures. They do not consider differences between the children of wage earners and the self-employed, or the health status of adults more generally. Ross and Mirowsky (2000) examine whether medical insurance helps explain differences by socioeconomic status in health. They find that, after controlling for socioeconomic status and base-line health, private insurance is not associated with good health outcomes and that public insurance is actually associated with *worse* health. We regard this finding with a degree of skepticism, since unobservable heterogeneity may be driving the results. Meara (2001) finds that the most important determinants of low birth weights are the health behaviors of the mother, rather than the availability of public insurance. Similarly, a key finding of the RAND Health Insurance Experiment is that the type of insurance an individual possesses has a significant effect on the utilization of health care, but only minor effects on health status (Newhouse 1993). But for the self-employed, even the link between insurance and utilization of medical services is rather weak. Perry and Rosen (2004) show that the differential use of health services between the self-employed and wage earners is less than one would expect on the basis of their differential insurance rates.

In short, when we consider previous papers focusing on the connections among health insurance, medical services utilization, and health outcomes, the self-employed make only a few appearances. In particular, there is no work on what is arguably the central policy question here: does the lack of health insurance among the self-employed lead to worse health outcomes for them? Further, the literature on the link between insurance and health outcomes in other contexts creates no presumption that the answer to this question is necessarily yes.

So far, we have ignored a question that all empirical analyses in this literature have to confront: Just how does one characterize health outcomes? The World Health Organization defines health as "a state of complete physical, mental, and social well-being, not merely the absence of disease or infirmity" (Newhouse 1993, p. 183). Clearly, no single number can capture every aspect of an individual's health. In the literature, basically two types of measures are used, subjective and objective.

Subjective measures rely on answers to questions such as the following one, which comes from the March 1996 Supplemental Current Population Survey: "Would you say your health in general is: (1) Excellent (2) Very good (3) Good (4) Fair (5) Poor?" Clearly, "healthy"

can mean different things to different people. For example, some smokers might consider themselves to be in excellent health, despite the fact that they cough incessantly. Similarly, some obese individuals might be unaware of their health risks. In the same way, some individuals may under-rate their health status when compared to other individuals whom they see as being very healthy, such as professional athletes. Nevertheless, it is well documented that self-reported measures of health have excellent explanatory power in predicting mortality rates. As Idler and Benyamini (1997, p. 21) note in their comprehensive survey of the literature on self-reported health measures, "over two dozen studies have been published in the U.S. and international literature that test the association between simple, global health assessments and mortality in the samples used: Most find a significant, independent association that persists when numerous health status indicators and other relevant covariates are included."[5]

Objective measures tend to rely on descriptions of behavior or diseases that are, in principle, observable. For example, another question from the March Current Population Survey asks "Do you have a health problem or disability which prevents you from working or which limits the kind or amount of work you can do?" The advantage of this type of measure is that the interpretation of responses is relatively simple—either an individual has a limitation or condition or does not (although even here one can imagine that a condition that would keep one person from working might not keep another person away from the job).

Neither type of measure is obviously superior to the other. As noted below, our data contain subjective as well as objective measures, and we analyze both. The hope is that we will find consistent results on the relationship between self-employment on health status regardless of the type of measure used. In addition to information on physical health, there are some self-reported mental health data, which we also discuss.

Data

Description

Our basic strategy is to see how differences in insurance coverage between the self-employed and wage earners translate into differences in health outcomes. This strategy requires information on an individ-

ual's insurance coverage and health status, as well as a set of exogenous characteristics that might influence health and insurance outcomes. We draw upon the Household Component of the 1996 Medical Expenditure Panel Survey (MEPS). The MEPS consists of approximately 22,000 respondents, in 9,500 families. In the survey, the respondents were asked a series of questions relating to their demographic characteristics, insurance coverage, employment status, and health. We exclude individuals with missing information on insurance, health, and education. In addition, we drop from the sample any people younger than 18 or older than 62. Those under 18 are unlikely to have developed a strong attachment to the labor market, and the decisions of those over 62 are complicated by impending retirement. All of these exclusions left a group of 8,986 individuals, of whom 1,088 (12 percent) were self-employed. This figure corresponds fairly closely to other estimates of the self-employment rate (U.S. Census Bureau 1998, p. 412).

As was noted in the preceding section, a major issue in a study like this is how to measure health outcomes. The MEPS contains both self-reported and objective characterizations of individuals' health status, and we examine both. The subjective measures include self-reported ratings for both general physical and mental health. The objective measures include information regarding individuals' physical limitations and whether or not they have a variety of medical conditions (including cancer and cardiac problems).

Preliminary Analysis

In table 1 we examine differences in health status and insurance by employment status. For each variable, columns 1, 2, and 3 show the means for the entire sample, for wage earners, and for the self-employed, respectively. The fourth column displays the t-statistic associated with the hypothesis that the means in columns 2 and 3 are equal.

The insurance variable in the first row is a dichotomous variable generated in the MEPS that takes a value of 1 if the individual has health insurance coverage and 0 otherwise. Specifically, the variable equals one if the individual is covered under Medicare, Medicaid, CHAMPUS/CHAMPVA,[6] or other public hospital/physician or private hospital/physician insurance. (Note that an individual is considered covered if the source of insurance is a spouse.) Sixty-nine percent of the self-employed in the sample have insurance, versus 81.5 percent

Table 1
Summary statistics: insurance and health status by employment status. Each entry in columns 1, 2, and 3 shows the proportion of the relevant group that had each condition within the last year. Figures in parentheses are standard errors. Column 4 shows t-tests on the differences in means in columns 2 and 3.

	1 Entire sample	2 Wage earners	3 Self-employed	4 Test statistic of difference in means between columns 2 and 3
Insurance	0.800 (0.00422)	0.815 (0.00437)	0.690 (0.0140)	90.717
Healthy	0.930 (0.00270)	0.928 (0.00290)	0.938 (0.00734)	−10.092
Mentally healthy	0.968 (0.00186)	0.967 (0.00201)	0.975 (0.00472)	−10.445
Any physical limitations	0.137 (0.00363)	0.135 (0.00385)	0.148 (0.0108)	−10.124
Priority condition	0.131 (0.00371)	0.129 (0.00391)	0.149 (0.0118)	−10.704
Cancer	0.00242 (0.000541)	0.00245 (0.00058)	0.00219 (0.00155)	0.152
Viral infection	0.0190 (0.00150)	0.0182 (0.00156)	0.0252 (0.00518)	−10.446
Headache	0.0206 (0.00156)	0.0208 (0.00167)	0.0186 (0.00447)	0.447
Cardiac condition	0.0272 (0.00179)	0.0271 (0.00189)	0.0284 (0.00550)	−0.238
Upper respiratory infection	0.105 (0.00337)	0.106 (0.00359)	0.0985 (0.00986)	0.679
Respiratory disease	0.0479 (0.00235)	0.0490 (0.00252)	0.0394 (0.00644)	1.284
Skin disease	0.0363 (0.00206)	0.0366 (0.00219)	0.0339 (0.00599)	0.412
Intestinal disorder	0.0496 (0.00239)	0.0509 (0.00256)	0.0394 (0.00644)	1.513
Arthritis	0.0230 (0.00165)	0.0218 (0.00170)	0.0328 (0.00590)	−2.100
Observations	8,986	7,898	1,088	

Table 2
Insurance source by employment status (conditional on having insurance). Figures in each cell are means, with standard errors in parentheses. All means are computed conditional on having insurance.

	1 Sample	2 Wage earners	3 Self-employed
CHAMPUS/CHAMPVA	0.0198 (0.00160)	0.0199 (0.00170)	0.0187 (0.00478)
Medicaid	0.0344 (0.00210)	0.0353 (0.00225)	0.0262 (0.00564)
Any public insurance	0.0687 (0.00291)	0.0683 (0.00307)	0.0722 (0.00914)
Medicare	0.00292 (0.000621)	0.00267 (0.000629)	0.00498 (0.00249)
Private	0.953 (0.00243)	0.955 (0.00254)	0.941 (0.00829)
Private employer group	0.850 (0.00412)	0.875 (0.00403)	0.635 (0.0170)
Private non-group	0.0439 (0.00236)	0.0273 (0.00199)	0.183 (0.0137)
Holder private insurance	0.717 (0.00519)	0.741 (0.00534)	0.514 (0.0176)
Holder private group insurance	0.645 (0.00551)	0.690 (0.00563)	0.270 (0.0157)
Holder private non-group insurance	0.00570 (0.000867)	0.00490 (0.000851)	0.0125 (0.00392)

of wage earners. From column 4, this difference is significant at all conventional levels—a result that is consistent with previous research (Holtz-Eakin, Penrod, and Rosen 1996; Health Insurance Association of America 2003).

As was suggested above, insurance can come from a variety of sources. Table 2 examines whether wage earners and the self-employed differ with respect to where their insurance comes from, conditional on having insurance. Column 1 reports the conditional proportions of the entire sample with each type of insurance; columns 2 and 3 present the conditional proportions for wage earners and the self-employed, respectively.

The first five rows reveal that, conditional on being insured, wage earners and the self-employed are equally likely to have public or private insurance, as well as to have coverage offered through the military. However, rows 6 and 7 indicate substantial differences between the two groups with respect to the type of private coverage:

87.5 percent of covered wage earners have private employer provided group coverage; the corresponding figure for the self-employed is only 63.5 percent. Similarly, over 18 percent of the self-employed have private non-group coverage, while only 2.7 percent of wage earners have non-group coverage. These results are consistent with the notion that the self-employed are unable to form or become part of groups that purchase insurance.

Another striking finding from table 2 is that the self-employed are significantly less likely than wage earners to be the holders of their policies. Only 51.4 percent of the self-employed, as compared to 74.1 percent of wage earners, are the policy holders for private insurance. Further, only 27 percent of the self-employed are the policy holders for group insurance policies, while 69 percent of wage earners are. These findings remind us of the importance of viewing insurance as a *family* rather than an *individual* issue. The fact that an entrepreneur cannot obtain his own insurance does not necessarily mean that he has to go uninsured.

In any case, the central issue is whether their relative lack of insurance affects the health status of the self-employed. As noted, the MEPS provides a subjective health evaluation based on the individual's response when asked to rate his or her health as poor, fair, good, very good, or excellent. Consistent with earlier literature, we use this information to create the dichotomous self-reported health variable HEALTHY, which takes a value of 1 if the individual is in good, very good, or excellent health, and 0 otherwise. Individuals were also asked to evaluate their mental health; in analogy to the physical health variable, we create a dichotomous variable MHEALTHY, which equals 1 if the individual reports himself being in good, very good, or excellent health, and 0 otherwise. The figures reported in rows 2 and 3 of table 1 indicate that one cannot reject the hypothesis that the mean values of both HEALTHY and MHEALTHY are the same for the two groups. Despite the differential in their rates of insurance coverage, the self-employed and wage earners have about the same subjective perceptions of physical and mental health.

To complement this discussion of subjective health status measures we examine several objective measures. Row 4 examines a dichotomous variable that takes a value of 1 if the individual has any physical limitations[7] and is equal to 0 otherwise. There appear to be no differences between the self-employed and wage earners in the likelihood of having physical limitations.

The MEPS also asks individuals a series of questions about specific medical conditions. To keep things manageable, we condense the conditions data into ten categories: cancer, viral infection, headaches, cardiac condition, upper respiratory infection, respiratory disease, skin disease, intestinal disorder, and arthritis.[8] The MEPS also indicates whether or not the individual has a "priority condition," defined as any of a number of serious medical conditions. These include AIDS, diabetes, emphysema, high cholesterol, hypertension, arthritis, gall bladder disease, stomach ulcers, back problems, Alzheimer's disease, and depression. A glance down column 4 of table 1 indicates that only for the case of arthritis is there a significant difference between the self-employed and wage earners; the self-employed are slightly more likely to have arthritis. From a statistical point of view, however, it is no surprise that if one examines a substantial number of effects, one of them comes up significant. In short, table 1 indicates that in spite of their low insurance rates, the self-employed appear generally as healthy as wage earners. Still, a number of different factors are known to influence health and some of them could be correlated with self-employment status. Hence, while these results are suggestive, we now turn to a multivariate approach.

Multivariate Analysis

The Setup

The univariate comparisons in table 1 suggest that self-employed individuals are just as healthy as wage earners, despite their lower propensity to have medical insurance. In this section we estimate conventional probit models to investigate whether this finding is robust to the inclusion of variables other than self-employment status that might influence an individual's health.

Focusing first on self-reported physical health status, we assume that the probability that the individual is healthy is given by

$$\text{Prob}(\text{HEALTHY}_i > 0) = \text{F}[\beta X_i + \delta \text{SE}_i],$$

where X_i is a vector of observable demographic characteristics, SE_i is a dichotomous variable equal to 1 if the individual is self-employed and 0 otherwise, and F[ ] is the cumulative normal distribution.

To estimate the model, we need to decide what to include in the X vector. We attempt to use only variables that are very likely to be

exogenous to health. Age is included because health tends to deteriorate with age. Previous research has also suggested that a quadratic function of age may be appropriate; therefore, we include the square of age. Education affects one's lifestyle and environment as well as the ability to pay for care (Taubman and Rosen 1982; Ruhm 2000); thus, there is a set of dichotomous variables for education level. In addition, some evidence suggests that race may be a factor in health status (Smith and Kington 1997). To allow for this possibility, we include a set of race dichotomous variables. Similarly, it has been documented that health status can vary by region (Preston and Taubman 1994); therefore, we use a set of indicator variables for the region of the country in which a person lives.[9]

Further, we enter a dichotomous variable for the individual's sex, because previous research has suggested that men and women differ in their probability of having various health conditions (Verbrugge 1985), and in the way they perceive their health (Idler and Benyamini 1997). Finally, we include a dichotomous variable for marital status and a continuous variable for family size—number of adults plus dependents. Previous research has suggested that marital status is correlated with differing levels of stress, which might affect health status (Taubman and Rosen 1982); similar reasoning would suggest that it is reasonable to include family size as well.

Our specification omits certain variables that have appeared as covariates in several previous studies of health status. A number of papers, for example, include household income. (See, for example, Ross and Mirowsky 2000 and McDonough et al. 1997.) There is indeed a substantial literature documenting the links between income and health status, but the direction of causality is not known. (See, for example, Deaton and Paxson 1999 and Ettner 1996.) To the extent that individuals' incomes are low because they are in poor health, then income is an endogenous variable and should be excluded from the reduced form.[10]

Insurance is another variable that sometimes appears in models of health status (Ross and Mirowsky 2000). But, as Gruber (2000, p. 46) noted, "insurance coverage itself may be a function of health status, leading to endogeneity bias in estimates of the effects of insurance on health." It is not clear whether there are any compelling instruments for either income or insurance status in this context, and we therefore exclude them. While this makes it difficult to attach a structural inter-

pretation to the results, it does increase the likelihood of obtaining consistent parameter estimates.

Table 3 lists and presents summary statistics for the right-hand-side variables just discussed, and for a few additional characteristics that are used in subsequent analyses. For each variable, the first column shows the mean value for the entire sample; the second and third columns exhibit the means for the self-employed and wage earners, respectively. The fourth column has t-tests on the differences in the means between columns 2 and 3. The table suggests that, in certain respects, the self-employed and wage earners are similar—levels of educational attainment, family size, and distribution across regions are roughly the same. The self-employed are more likely to be white, male, and married with a spouse present. Further, the self-employed tend to be older (5.4 years) on average than wage earners. They also have higher incomes ($3,000 per year) and work longer hours. These findings all echo previous research (Fairlie and Meyer 1999; Hamilton 2000).

An important question is whether there is unobservable heterogeneity with respect to health status. Do the self-employed differ systematically from wage earners in their underlying health in ways that cannot be captured by the covariates in table 3? Specifically, might there be unobservable variables that drive both health status and the likelihood of becoming self-employed? For example, perhaps very healthy, energetic people have the "animal spirits" that lead them to become entrepreneurs. Alternatively, perhaps people who are too ill to hold jobs as employees decide to become self-employed.

Previous research with other data sets suggests that, in fact, there is no selection along these lines. Holtz-Eakin, Penrod, and Rosen, hereafter cited as HPR, employ both the Survey of Income and Program Participation (SIPP) and the Panel Study of Income Dynamics (PSID) data to examine transitions from wage earning to self-employment (HPR 1996). Both data sets indicate that health status is not a good predictor of whether a wage earner will become self-employed in the future or not, ceteris paribus. In the next section we use the MEPS to update and extend the HPR study. We examine both transitions from wage earning into self-employment and from self-employment into wage earning, and, like HPR, we find no selection on the basis of health status. While these findings cannot definitively exclude the possibility of unobservable heterogeneity, they certainly provide no

Table 3
Summary statistics: individual characteristics by employment status. Figures in each cell are means, with standard errors in parentheses. Except for family size, age, age squared, wage, and hours per week, all variables are dichotomous. They are equal to 1 if the individual is in the category, and 0 otherwise. Column 4 is a t-test of the difference between columns 2 and 3.

	1 Entire sample	2 Wage earners	3 Self-employed	4 t-test
Education				
No degree[a]	0.128 (0.00353)	0.130 (0.00378)	0.118 (0.00977)	1.133
GED	0.0425 (0.00213)	0.0437 (0.0023)	0.0340 (0.00550)	1.483
High school diploma	0.502 (0.00527)	0.505 (0.00563)	0.481 (0.0152)	1.507
B.A.	0.176 (0.00402)	0.176 (0.00429)	0.177 (0.0116)	−0.113
Master's	0.0591 (0.00249)	0.0575 (0.00262)	0.0708 (0.00778)	−1.743
Ph.D.	0.0154 (0.00130)	0.0125 (0.00125)	0.0358 (0.00564)	−5.873
Other degree	0.0763 (0.00280)	0.0753 (0.00297)	0.0836 (0.00840)	−0.967
Race				
Other	0.0414 (0.00210)	0.0419 (0.00225)	0.0377 (0.00578)	0.656
Black	0.122 (0.00346)	0.129 (0.00378)	0.0726 (0.00787)	5.354
White[a]	0.835 (0.00391)	0.828 (0.00425)	0.890 (0.00950)	−5.168
Region				
Northeast	0.192 (0.00416)	0.190 (0.00441)	0.208 (0.0123)	−1.408
Midwest	0.230 (0.00444)	0.233 (0.00476)	0.207 (0.0123)	1.901
South	0.35 (0.00504)	0.356 (0.00539)	0.313 (0.0141)	2.771
West[a]	0.227 (0.00442)	0.221 (0.00467)	0.272 (0.0135)	−3.745
Other				
Age	38.8 (0.117)	38.2 (0.125)	43.6 (0.300)	−15.41
Age squared	1631 (9.30)	1580 (9.82)	2001 (25.9)	−14.96
Family size	3.11 (0.0165)	3.11 (0.0175)	3.13 (0.0492)	−0.45

Table 3
(continued)

	1 Entire sample	2 Wage earners	3 Self-employed	4 t-test
Male	0.526 (0.00527)	0.511 (0.00563)	0.635 (0.0146)	−7.686
Married with spouse in house	0.620 (0.00512)	0.601 (0.00551)	0.756 (0.0130)	−9.877
Income	26473 (225.3)	26119 (229.4)	29049 (827.3)	−4.247
Hours worked per week	37.91 (0.145)	37.37 (0.140)	41.76 (0.614)	−9.948

a. omitted from right-hand side of regression models.

support for the notion that people who select into self-employment are systematically different with respect to health-related attributes.

Basic Results

Above, we used the differences in insurance status between wage earners and the self-employed as a kind of base line against which to measure differences in health status. In analogy, we begin the multivariate analyses with an examination of the probability of being insured, and then turn to the various indicators of health status.

Insurance Coverage

The results are reported in column 1 of table 4, which presents the marginal effect of each of the variables on the probability of having insurance coverage. Notably, the coefficient on the self-employed variable (SE) is both negative (−0.194) and statistically significant (standard error = 0.0173). Since 81.5 percent of the wage earners have insurance, this implies that the self-employed are 25 percent less likely to be insured, even after controlling for demographic characteristics.

While not the primary focus of this chapter, the other coefficients in column 1 merit some discussion. The coefficients on the age variables indicate insurance coverage increases throughout the entire relevant range of ages. The male variable's coefficient suggests that men are 3.6 percentage points less likely to be insured than women. Consistent with previous research (Institute for the Future 2000, p. 23), the coefficients on the education variables indicate that, relative to individuals

with no high school degree, people with more education have higher coverage rates.

Table 4 also reveals that family composition affects an individual's insurance status. Ceteris paribus, the likelihood of having coverage falls by 1.4 percentage points with each additional person in the family. Further, married persons are 13.8 percentage points more likely to have coverage than single individuals. Since spouses often act as sources of insurance, this result is not surprising.

The coefficients on the race variables tell an interesting story. Notably, the coefficient on the black variable indicates that blacks are 2.8 percentage points less likely to have coverage than whites (the omitted group), other things being the same. Members of the "other" category, which consists of Asian-Americans, Eskimos, and Native Americans, are 5.1 percentage points less likely to have insurance than whites.

There are substantial regional effects. Northeasterners are about 3.0 percentage points more likely to have insurance than those in the west (the omitted category), while midwesterners are 5.1 percentage points more likely. People who live in the south are about as likely to have insurance as those who live in the west.

Health Status

With the insurance results in hand, we now turn to the various health measures available in the MEPS. Column 2 of table 4 reports the results for the self-reported health measure. The coefficient on the self-employment variable is small and insignificantly different from 0–0.0119, with standard error 0.00705. There is no statistically discernible difference in subjective evaluations of health between wage earners and the self-employed. Insofar as the self-employed are 25 percent less likely to have health insurance, this finding contradicts the notion that their lack of insurance translates into worse health outcomes.

Before examining the remaining health indicators, we discuss the coefficients of the other variables in column 2. The linear and quadratic age variables are individually significant, but taken together, they are jointly significant, with a chi-squared statistic of 52.3. Together they imply that the probability of being healthy declines throughout the age range. The dichotomous variables for education reveal that health outcomes tend to improve with education, a finding that is consistent with previous research (Ross and Mirowsky 2000). Family size, marital status, and location have no statistically discernible effect on the self-reported health status measure. However, black individuals are 2 per-

Table 4
Probit estimates for insurance coverage and for measures of general health status. The coefficients give the marginal effects of the associated right-hand-side variable on the probability of being covered by insurance in column 1, and on the probabilities of assessing oneself as healthy, assessing oneself as mentally healthy, having any physical limitations, and having a priority condition, in columns 2, 3, 4, and 5, respectively. The standard errors appear in parentheses. The coefficients give the marginal effects of the associated right-hand-side variable on the probability of being covered by insurance (column 1), and on the probabilities of assessing oneself as being healthy, assessing oneself as being mentally healthy, having any physical limitations, and having a priority condition, in columns 2, 3, 4, and 5, respectively. The standard errors appear in parentheses.

	1 Insurance status	2 HEALTHY	3 MHEALTHY	4 Any physical limitations	5 Priority condition
Self-employed	−0.194 (0.0173)	0.0118 (0.00706)	0.00625 (0.00509)	−0.0100 (0.0104)	0.00333 (0.0116)
Age	0.00642 (0.00256)	−0.00255 (0.00163)	−0.00312 (0.00115)	0.00739 (0.00236)	0.00710 (0.00240)
Age squared	−0.0000338 (0.0000325)	90.56×10^{-6} (0.0000199)	0.0000348 (0.0000142)	−0.0000323 (0.0000287)	−0.0000494 (0.0000293)
GED	0.0877 (0.0128)	0.0212 (0.00873)	0.00593 (0.00678)	0.0592 (0.0233)	0.0262 (0.0215)
H.S. diploma	0.199 (0.0114)	0.0667 (0.00710)	0.0291 (0.00508)	−0.00930 (0.0112)	−0.0167 (0.0126)
B.A.	0.190 (0.00701)	0.0645 (0.00452)	0.0237 (0.00367)	−0.0357 (0.0117)	−0.0566 (0.0117)
M.A.	0.173 (0.00576)	0.0561 (0.00420)	0.0226 (0.00368)	−0.0606 (0.0126)	−0.0522 (0.0136)
Ph.D.	0.165 (0.00585)	0.0570 (0.00478)	—	−0.0668 (0.0194)	−0.0533 (0.0207)
Other degree	0.151 (0.00727)	0.0531 (0.00455)	0.0239 (0.00340)	−0.0400 (0.0136)	−0.0246 (0.0152)
Family size	−0.0141 (0.00276)	−0.00100 (0.00183)	−0.00235 (0.00121)	−0.0173 (0.00277)	0.000624 (0.00293)
Black	−0.0275 (0.0133)	−0.0194 (0.00870)	−0.00197 (0.00562)	−0.0345 (0.00991)	0.0138 (0.0134)
Other	−0.0506 (0.0238)	−0.0249 (0.0149)	0.00221 (0.00843)	−0.0447 (0.0154)	−0.0140 (0.0187)
Northeast	0.0296 (0.0118)	0.00828 (0.00740)	0.00764 (0.00484)	−0.0443 (0.00983)	−0.0317 (0.0105)
Midwest	0.0514 (0.0109)	0.0129 (0.00705)	0.00222 (0.00500)	0.00751 (0.0106)	0.0172 (0.0108)
South	0.00276 (0.0108)	0.00380 (0.00666)	0.00881 (0.00442)	−0.00585 (0.00962)	−0.0145 (0.00963)
Male	−0.0363 (0.00820)	0.0132 (0.00516)	0.00406 (0.00361)	−0.00502 (0.00723)	0.0262 (0.00762)
Married	0.138 (0.0106)	0.00213 (0.00603)	0.0164 (0.00455)	−0.0126 (0.00854)	−0.0218 (0.00896)
Log likelihood	−3,851	−2,166	−1,228	−3,374	−3,129
Observations	8,986	8,986	8,986	8,803	8,260

centage points less likely than whites to report that they are in good health. Men are 1.3 percentage points more likely to report that they are in good health than women. This finding must be interpreted with caution, because some researchers have suggested that men and women may use different processes to incorporate information into their self-assessments of health (Idler and Benyamini 1997, p. 26). Likewise, the results in column 3 of table 4 with respect to mental health must be taken with a grain of salt. While there is no statistically significant difference between the self-employed and wage earners in their perceived mental health status, one cannot be sure of the validity of this self-reported measure.

These reminders of possible problems with subjective health measures provide a natural segue to our analyses of the various objective health measures. We re-estimate the model for each of a series of such measures. Columns 4 and 5 of table 4 look at two summary measures of health: whether there are any physical limitations and whether the individual has a priority condition. As was the case with the subjective measure in column 2, there are no statistically discernible differences between the self-employed and wage earners in their propensity to be healthy. That is, the objective measures give exactly the same answer as the subjective measure.

This conclusion is reinforced by table 5, which presents results for seven specific health conditions. There is not one single condition that the self-employed are statistically more likely to have than wage earners. In short, even though the self-employed are 25 percent less likely to be insured than wage earners, their health does not appear to be any worse, ceteris paribus. Thus, concerns about their health do not seem to merit medical insurance subsidies to the self-employed.

Alternative Specifications

We subjected the model to a variety of different tests to examine whether the substantive results were sensitive to changes in specification.

Income

Previous research has shown that income is positively related to health status. The conventional explanation is that "the less well-to-do have access to less or lower quality medical care" (Smith 1999, p. 145). Recall that the tabulations above revealed that the self-employed have higher

average incomes than wage earners (on the order of $3,000). Perhaps, then, the fact that we find no health differences between wage earners and the self-employed is due simply to the fact that the self-employed have higher incomes. To allow for this possibility, we augment the canonical specification with family income.[11] Of course, as was noted above, income might be endogenous if, for example, healthier individuals are able to work more and earn higher incomes. For this reason, income was not included in the basic specifications in tables 4 and 5.

Column 1 of table 6 shows the self-employment coefficients only from the augmented probit models for the various health measures. The results indicate that including income on the right-hand side generally has no significant effect on the self-employment coefficients. Again, because of the potential endogeneity of income, these results should be interpreted with caution. Just the same, the inclusion of family income as a covariate reinforces the core result—wage earners and the self-employed appear equally healthy.

Hours

It is well documented that the compensation packages of part-time workers are less likely than those of full-time employees to include benefits such as medical insurance (Campling 1987). At the same time, there is reason to suspect that self-employment might be correlated with hours of work. In fact, the correlation in our data is 0.104. Hence, our estimates of the effects of self-employment on insurance coverage and utilization rates might be biased because of the failure to take into account differences in hours worked. Therefore, we augment the canonical specification with a set of dichotomous variables for hours worked per week. Of course, hours of work might itself be endogenous, since people who are ill may work fewer hours, ceteris paribus. That is why it was not included in the original specification.

The coefficients on the self-employment variables associated with this specification are reported in column 2 of table 6. A quick comparison with the results in tables 4 and 5 suggests that, for almost every health measure, the inclusion of the hours of work has barely any impact on the self-employment coefficient.

Utilization

Some previous research has used differences in the utilization of medical services to help explain disparities in health (Thomas et al. 1992). Certainly, ceteris paribus, one would expect medical service usage and

Table 5
Probit estimates for specific health status measures. Coefficients give marginal effects of associated right-hand-side variable on the probabilities of having various health conditions. Standard errors appear in parentheses.

	1 Viral infection	2 Headaches	3 Cardiac conditions	4 Upper respiratory infection	5 Respiratory disease	6 Skin disease	7 Intestinal disorder	8 Arthritis
Self-employed	0.00877 (0.0564)	−0.000817 (0.00466)	−0.00444 (0.00415)	0.00201 (0.0111)	−0.0110 (0.00674)	−0.00424 (0.00622)	−0.00738 (0.00729)	0.00634 (0.00516)
Age	−0.000114 (0.000959)	0.00195 (0.000943)	0.00276 (0.00112)	0.000126 (0.00218)	0.000249 (0.00148)	−0.00125 (0.00134)	−0.000659 (0.00150)	0.00197 (0.00102)
Age squared	-3.37×10^{-6} (0.0000121)	−0.0000264 (0.0000118)	−0.0000188 (0.0000131)	−0.0000291 (0.0000275)	-4.30×10^{-6} (0.0000182)	0.0000158 (0.0000164)	-3.44×10^{-6} (0.0000191)	−0.0000137 (0.0000122)
GED	−0.00397 (0.00753)	0.00178 (0.00806)	0.0105 (0.0108)	0.0107 (0.0206)	0.00956 (0.0149)	0.00196 (0.0120)	−0.0118 (0.0108)	−0.00266 (0.00808)
H.S. diploma	0.00395 (0.00561)	−0.00436 (0.00499)	0.00277 (0.00541)	0.0201 (0.0126)	0.00133 (0.00875)	0.000122 (0.00755)	−0.00146 (0.00810)	0.000878 (0.00550)
B.A.	0.00795 (0.00753)	−0.00245 (0.00521)	−0.00675 (0.00540)	0.0296 (0.0155)	0.0191 (0.0112)	0.00345 (0.00884)	−0.00440 (0.00883)	0.00509 (0.00668)
M.A.	0.00663 (0.00969)	−0.0110 (0.00445)	−0.00267 (0.00687)	0.0395 (0.0209)	0.0245 (0.0152)	0.0212 (0.0138)	−0.0100 (0.0104)	−0.00243 (0.00695)
Ph.D.	0.00837 (0.0161)	−0.0132 (0.00643)	−0.0108 (0.00770)	0.0395 (0.0338)	0.0230 (0.0236)	0.0131 (0.0201)	−0.0154 (0.0154)	−0.00662 (0.00932)
Other degree	0.00783 (0.00932)	0.000222 (0.00660)	−0.000133 (0.00737)	0.0317 (0.0192)	0.00468 (0.0123)	0.0137 (0.0121)	0.00362 (0.0114)	−0.00258 (0.00677)
Family size	−0.000462 (0.00111)	−0.00141 (0.00117)	0.00135 (0.00113)	0.00233 (0.00254)	−0.000173 (0.00181)	−0.000702 (0.00157)	0.000940 (0.00174)	0.000991 (0.00118)

Black	−0.00624	−0.00349	0.0157	0.00476	−0.0125	0.0122	−0.0118	0.00939
	(0.00428)	(0.00450)	(0.00714)	(0.0123)	(0.00732)	(0.00830)	(0.00744)	(0.00647)
Other	−0.000114	−0.0168	0.0216	−0.000291	0.000736	0.0117	−0.000961	0.0230
	(0.00762)	(0.00351)	(0.0121)	(0.0180)	(0.0124)	(0.0123)	(0.0117)	(0.0115)
Northeast	−0.00275	−0.00534	0.0148	−0.000274	0.00434	−0.00741	−0.0191	−0.00545
	(0.00413)	(0.00423)	(0.00652)	(0.0107)	(0.00778)	(0.00575)	(0.00611)	(0.00427)
Midwest	−0.00521	−0.000210	0.00443	0.0136	0.00751	−0.00557	−0.00577	0.00669
	(0.00361)	(0.00414)	(0.00506)	(0.00997)	(0.00714)	(0.00537)	(0.00616)	(0.00475)
South	−0.00494	0.00483	0.00556	−0.00819	0.00405	−0.00862	−0.00432	−0.00435
	(0.00373)	(0.00416)	(0.00466)	(0.00912)	(0.00651)	(0.00509)	(0.00591)	(0.00395)
Male	0.00193	−0.00432	0.0125	0.0210	0.0166	0.00706	0.00866	−0.000256
	(0.00298)	(0.00291)	(0.00338)	(0.00697)	(0.00490)	(0.00425)	(0.00481)	(0.00308)
Married	−0.00374	−0.00316	0.000207	0.00420	0.000163	0.00581	−0.00244	−0.00155
	(0.00340)	(0.00361)	(0.00388)	(0.00784)	(0.00543)	(0.00462)	(0.00557)	(0.00367)
Log likelihood	−768	−813	−978	−2,736	−1,573	−1,280	−1,609	−877
Observations	8,260	8,260	8,260	8,260	8,260	8,260	8,260	8,260

Table 6
Self-employment effects in alternative specifications. These are the coefficients on the self-employment dichotomous variables from the probit equations of tables 4 and 5 augmented with a continuous variable for family income (column 1), with a set of dichotomous variables for hours worked (column 2), and with a continuous variable for number of doctor visits (column 3). Coefficients are marginal effects on the respective probabilities, and figures in parentheses are standard errors.

	1 Income	2 Hours	3 Doctor visits
HEALTHY	0.0108 (0.00705)	0.0119 (0.00727)	0.0110 (0.00705)
MHEALTHY	0.00532 (0.00505)	0.00831 (0.00495)	0.00568 (0.00512)
Any physical limitations	−0.00923 (0.0105)	−0.0199 (0.0103)	−0.00910 (0.0104)
Priority condition	0.00172 (0.0117)	0.00271 (0.0122)	0.00368 (0.0116)
Cancer	−0.000287 (0.000266)	−0.000720 (0.000483)	−0.000516 (0.000546)
Viral infection	0.00753 (0.00545)	0.00904 (0.00596)	0.00884 (0.00564)
Headaches	−0.000554 (0.00476)	0.00127 (0.00529)	−0.000966 (0.00461)
Cardiac condition	−0.00427 (0.00428)	−0.00417 (0.00430)	−0.00457 (0.00413)
Upper respiratory infection	0.00380 (0.0113)	0.00444 (0.0117)	0.00184 (0.0111)
Respiratory disease	−0.0112 (0.00676)	−0.0115 (0.00697)	−0.0110 (0.00672)
Skin disease	−0.00473 (0.00624)	−0.00547 (0.00623)	−0.00441 (0.00618)
Intestinal disease	−0.00590 (0.00747)	−0.00704 (0.00772)	−0.00751 (0.00727)
Arthritis	0.00684 (0.00520)	0.00671 (0.00538)	0.00610 (0.00509)

health status to be related; however, the direction of causation is unclear. In a demand function for health services, for example, one might include health status as an explanatory variable—healthier individuals require less health care. Alternatively, however, one could argue that people who consume more health-care services receive treatments that lead to better health. Therefore, including utilization rates of health-care services on the right-hand side of an equation explaining health status is problematic. That said, previous research indicates that self-employed individuals are less likely than wage earners to use many (but not all) types of medical-care services (Perry and Rosen 2004). Thus, to the extent that utilization *does* belong on the right-hand side, failure to take it into account may bias the estimates of the self-employment effects on insurance coverage and health status.

One common measure of health-care utilization is the number of doctor visits during the year. We therefore augmented the canonical specification with a continuous variable for number of doctor visits. The results, reported in column 3 of table 6, suggest that its inclusion has no serious impact on the self-employment coefficients. Thus, to the extent that utilization does belong in the model, it appears to have no effect on our substantive results.

Children

We have shown that the relative lack of health insurance among the self-employed does not appear to have a negative effect on their health. However, much of the recent concern over health insurance has focused on the needs of children. One could argue that a tax subsidy to the self-employed for purchases of health insurance is warranted if it helps improve their children's health. Do the children of the self-employed have worse health than the children of wage earners, ceteris paribus? We address this question by taking advantage of a set of parental reported and objective health measures in the MEPS. Three of these measures are based on the parents' responses to a series of statements about their children's health: "Child resists illness," "Child seems to be less healthy than other children," and "Child seems to catch diseases that are going around." The parent then responded on a scale from 1 to 4, where 1 meant "definitely false" and four meant "definitely true." We convert each answer into a dichotomous variable equal to 1 if the respondent's answer was indicative of the presence

of a health problem (a response of 1 or 2 to the first statement, and an answer of 3 or 4 to the second and third statements).

Earlier we cited research that indicated that adults' self-reported health reports are meaningful indicators of their health status. We know of no such research validating parents' assessments of their children's health. As Currie and Gruber (1995) note, such measures may be subject to directional bias based on contact with the health-care system. Further, there is some evidence that the number of illnesses a mother reports for her children is a function of her education (Currie and Thomas 1995).[12] Hence, while interesting, these parental evaluations must be viewed with caution.

The MEPS also has some more objective measures of children's health. For children 4 years of age and younger we have information on whether there are any limitations on their activities,[13] and for children 17 and under a set of condition variables similar to those we studied for adults. As before, it is useful to have as a base line an estimate of how self-employment affects the probability of being insured for the relevant population. We use the sample of families with children under 17 to estimate an equation for the probability that the children in the family were covered by some form of health insurance. On the right-hand side we include a dichotomous variable which takes a value of 1 if both parents were self-employed or if one parent was self-employed and the other did not work, and 0 otherwise. In addition, we include a vector of the child's characteristics including age, age squared, race, family size, sex, and region.

The results are reported in column 1 of table 7. They indicate that the children of the self-employed are about as likely as the children of wage earners to have insurance coverage—one cannot reject the hypothesis that the coefficient on the parent self-employment variable is 0. This is a striking contrast to the 19.4-percentage-point differential between the probabilities that self-employed and wage-earning adults have health insurance. Apparently, parents place a premium on having their children covered, a result that is certainly consistent with anecdotal evidence. For example, after a recent 40 percent spike in insurance premia for his two children, a wage earner named Eddie Williams observed:

> Of course you ask yourself why. You even wonder whether it's worth it to pay all that. The children are healthy. Seems like they've only gone to the doctor twice this year, both times for shots, which weren't even covered by the insur-

Table 7
Insurance and health status for children. This table shows the coefficient on the dichotomous variable for parents' self-employment status in each of a series of models estimated using as observations the children in the sample. Other covariates are child's age, race, sex, and region. The figures are the marginal effects from probit equations, with the standard errors in parentheses.

	Coefficient on parents' self-employment status
Insurance coverage	0.04003 (0.0322)
Does not resist illness well	0.0320 (0.0397)
Less healthy than others	−0.00729 (0.0348)
Catches diseases	0.000802 (0.0514)
Priority condition	0.0151 (0.0272)
Upper respiratory infection	0.0393 (0.0486)
Skin disease	−0.00927 (0.0177)
Intestinal disease	0.00457 (0.0287)

ance. But these are my kids we're talking about here. You never know what might happen. So we pay it. I wouldn't dream of them being without insurance. (Verhovek, *New York Times*, September 18, 2000)

In view of the lack of an insurance coverage differential, the rest of table 7 is rather anti-climactic. Analyses of both the parent-reported responses and the objective measures indicate that there are no statistically significant differences between the children of the self-employed and the children of wage earners. Concerns for the health of their children do not seem to provide adequate justification for subsidizing the health-insurance purchases of the self-employed.

Do Healthier People Become Self-Employed?

A potentially important problem mentioned earlier is that unobservable heterogeneity may be driving our results. Specifically, the concern is that underlying differences between the self-employed and wage earners with respect to health and the demand for health services may

not be captured by the covariates. One can imagine, for example, that people who are too ill to hold jobs as employees decide to become self-employed. Alternatively, it may be that healthy, energetic people have the "animal spirits" that lead them to become entrepreneurs. This latter possibility is particularly important in view of our finding that, in spite of their relatively low insurance rates, the self-employed do not suffer from adverse health outcomes relative to their wage-earning counterparts. Perhaps this result is due to the fact that the self-employed are healthier to begin with. We address this issue by examining transitions into and out of self-employment. Consider a group of wage earners during a given time period. If the probability that an individual transits to self-employment in the subsequent period is independent of his or her health status at the outset, then one can feel some confidence that selection into self-employment on the basis of health is not driving our results. On the other hand, if healthier individuals are more likely to make transitions into self-employment, the interpretation of our findings becomes problematic.

As was noted above, this issue has been studied previously by Holtz-Eakin, Penrod, and Rosen (1994). They employed both the Survey of Income and Program Participation (SIPP) and the Panel Study of Income Dynamics (PSID) to examine transitions from wage earning to self-employment. Both data sets indicate that, in a given year, those wage earners who become self-employed in the future are not statistically different in health status or health-care utilization from those who remain wage earners.[14] In the SIPP data, the health measures are combined days in bed during the last 4 months and a self-reported health-status variable. The utilization measures are combined nights in a hospital in the last 4 (and 12) months and the combined number of doctor visits in the last 4 (and 12) months. In the PSID, the health measures are hours of work lost due to illness and a self-reported health variable. The utilization measure is number of nights in the hospital during the year.

In this section we update and extend these results using the MEPS. We take advantage of the panel nature of the data set to examine transitions into and out of self-employment between rounds 1 and 5, corresponding to the period from January 1996 to January 1998.[15] The MEPS has two advantages in this context. First, it allows us to study the transitions of the same sample of individuals upon whom our results on self-employment and health are based. Second, these data

are more recent and provide richer information on utilization and health-care status than the data sets used by HPR.

During the two-year period, 145 individuals made the transition from wage earning to self-employment (from an initial group of 7,188 wage earners) and 138 left self-employment to become wage earners (from an initial group of 836 self-employed). The implied rates of entry (about 2 percent) and exit (about 16 percent) are similar to those that have been found in other data sets (see HPR 1996).

Self-Employment Transitions and Health Status

To begin, we examine transitions into self-employment by wage earners as a function of a variety of indicators of their health status. The sample consists of wage earners in January 1996, and we examine the probability that they are self-employed in January 1998, conditional on a set of demographic characteristics and their initial health status.[16] If one believed that our results were due to the fact that healthy people are particularly likely to enter self-employment, then, ceteris paribus, one would expect indicators for good health to increase the probability of transiting to self-employment, and vice versa. The results are reported in column 1 of table 8. The first row reveals that the coefficient on the self-reported measure of health status, HEALTHY, is statistically insignificant. Moving down the column, we see that the same holds true as well for every single specific health condition. In short, whether subjective or objective measures of health status are employed, the results in table 8 suggest no systematic tendency for healthier people to enter self-employment.

Column 2 of table 8 reports the results from a series of equations that examine transitions out of self-employment into wage earning. Here the sample consists of individuals who were self-employed in January 1996, and the left-hand-side variable is the probability that they were wage earners 2 years later. None of the health measures has any effect on the decision to exit self-employment except for the presence of headaches. Of course, insofar as the results from a dozen regressions are reported in column 2, it is not surprising to turn up at least one statistically significant health measure. But even if there is a true "headache effect," when taken in conjunction with the other results in table 8, it does not undermine the main message—health status does not appear systematically to influence decisions to enter or leave self-employment.

Table 8
Health effects on transitions into and out of self-employment. Each value in column 1 shows the marginal effect of the associated health condition on the probability of making a transition from wage earning to self-employment, ceteris paribus. The values in parentheses are standard errors. Each coefficient is generated from a probit model in which the left-hand-side variable is the probability that an individual who was a wage earner initially is self-employed 2 years later. The right-hand-side variables are those in table 4 in addition to the associated health variable. Each value in column 2 shows the marginal effect of the associated health condition on the probability of making a transition from self-employment to wage earning, ceteris paribus. For this column, the probit equation is estimated over the sample of individuals who were initially self-employed, and the left-hand-side variable is the probability of being a wage earner 2 years later.

	1 Probability of being self-employed conditional on having been a wage earner	2 Probability of being a wage earner conditional on having been self-employed
HEALTHY	0.00110 (0.00629)	0.0102 (0.0416)
MHEALTHY	−0.0116 (0.0116)	−0.0650 (0.0705)
Priority condition	−0.000820 (0.00367)	−0.0368 (0.0275)
Cancer	0.0712 (0.0786)	a
Viral infection	0.00217 (0.0102)	0.127 (0.0960)
Headaches	−0.00233 (0.00687)	−0.0861 (0.0231)
Cardiac condition	0.0177 (0.0132)	−0.0472 (0.0433)
Upper respiratory infection	0.00169 (0.00427)	0.00462 (0.0379)
Respiratory disease	0.000410 (0.00575)	0.00920 (0.0600)
Skin disease	0.00193 (0.00711)	−0.0203 (0.0555)
Intestinal disease	−0.00788 (0.00326)	0.0142 (0.0480)
Arthritis	0.00210 (0.00993)	0.0238 (0.0693)
Any physical limitations	0.00185 (0.00327)	−0.00888 (0.0268)
Observations	7,188	861

a. Not estimated because of perfect collinearity.

Self-Employment Transitions and Children's Health

We argued above that there appear to be few significant differences in health status between the children of the self-employed and those of wage earners. This raises a question analogous to the one just discussed: Does the health status of a person's children affect his or her decision to enter or exit self-employment? To examine this possibility, we estimate the same kind of transition equations as reported in table 8, but this time using parent-reported and objective measures of children's health. As in table 7, for each parent-reported measure, the associated dichotomous variable takes a value of 1 if the answer for any of a person's children is consistent with the presence of a health problem. Similarly, for each objective measure, the dichotomous variable is 1 if any child has the condition. The results are reported in table 9. In general, one cannot reject the hypothesis that the coefficients on the children's health variables are 0. The exceptions are the grab-bag "priority condition" variable and intestinal diseases (for entry into self-employment only). We are inclined to regard these as statistical anomalies, especially because the "priority condition" variable appears with the same sign in both the entry and exit equations. By and large, the main story told by the table is that self-employment transitions are not significantly affected by children's health.

Summary

This section has investigated the possibility of self-selection into or out of self-employment on the basis of health conditions. We find that, in general, a wage earner's health status does not predict whether he or she will be self-employed 2 years later, ceteris paribus. Similarly, a self-employed person's health status does not predict whether or not he or she will be a wage earner 2 years later. Neither does a child's health status predict whether the child's parent will make a transition into or out of self-employment.

In work not reported here for the sake of brevity, we also investigated whether an individual's initial utilization of health services is a predictor of transitions into or out of self-employment. These results, too, suggest that health issues are not related to the selection of employment mode.[17] On the basis of the available evidence, then, we conclude that our findings with respect to the lack of health differences between wage earners and the self-employed—despite the large

Table 9
Effects of children's health on transitions into and out of self-employment. Each value in column 1 shows the marginal effect of the associated child's health condition on the probability of a parent making a transition from wage earning to self-employment, ceteris paribus. The values in parentheses are standard errors. Each coefficient is generated from a probit model in which the left-hand-side variable is the probability that an individual who was a wage earner initially is self-employed 2 years later. The right-hand-side variables are those in table 4 in addition to the associated child's health variable. Each value in column 2 shows the marginal effect of the associated child's health condition on the probability of making a transition from self-employment to wage earning, ceteris paribus. For this column, the probit equation is estimated over the sample of individuals who were initially self-employed, and the left-hand-side variable is the probability of being a wage earner 2 years later.

	1 Probability of being self-employed conditional on having been a wage earner	2 Probability of being a wage earner conditional on having been self-employed
Does not resist illness well	−0.00238 (0.00309)	0.00526 (0.0272)
Less healthy than others	0.0116 (0.00689)	0.00760 (0.0508)
Catches diseases	−0.00287 (0.00331)	−0.00389 (0.0350)
Priority	−0.00884 (0.00323)	−0.0837 (0.0315)
Viral infection	0.00876 (0.00801)	0.0703 (0.0596)
Upper respiratory infection	0.00525 (0.00460)	−0.00282 (0.0371)
Respiratory disease	0.00739 (0.00738)	−0.0283 (0.0488)
Intestinal disease	−0.0102 (0.00218)	−0.0463 (0.0455)
Skin disease	−0.00737 (0.00436)	0.147 (0.0805)
Observations	7,029	836

differences in insurance coverage—is not due to the fact that relatively healthy people tend to select into self-employment.

Conclusion

Using data from the 1996 Medical Expenditure Panel Survey, we have analyzed differences between the self-employed and wage earners with respect to insurance coverage and health status. Our results suggest that the relative lack of health insurance among the self-employed has essentially no effect on their health or on the health of their children. This finding is robust to a number of reasonable changes in the specification of our statistical model. Further, we demonstrate that the result does not seem to be due to selection into self-employment on the basis of health status.

There are several possible explanations for this phenomenon. One is that the self-employed finance health care from sources other than insurance. Perhaps, for example, they self-insure, paying for medical care out of their incomes or accumulated saving. However, in other research we have shown that the out-of-pocket costs that the self-employed incur for health care do not differ much from those of wage earners, both in absolute terms and relative to income (Perry and Rosen 2004).[18] Another possibility is that access to health care is responsible for only a relatively small part of health, more important determinants being genetics, environment, and health behaviors (Institute for the Future 2000, p. 23). From this perspective, our results might be viewed as adding to a line of research which has shown, in a variety of other contexts, that the links between insurance coverage and health outcomes are weaker than one might imagine. (See Currie and Gruber 1995; Meara 2001; Jaestner, Joyce, and Racine 1999; Ross and Mirowsky 2000.) In any case, insofar as the self-employed do not suffer adverse health outcomes as a result of their relative lack of health insurance, targeting health-insurance subsidies at them may not be an appropriate public policy.

Acknowledgments

We are grateful to Princeton University's Center for Economic Policy Studies and the National Science Foundation for financial support of this research. We thank Douglas Holtz-Eakin for useful suggestions.

Appendix

The purpose of this appendix is to provide careful definitions of the various health-status variables employed in the text.

PRIORITY is set equal to 1 if an individual has any of the following conditions:

Long-term, life threatening conditions:

Cancer (of any body part): cancer, tumor, malignancy, malignant tumor, carcinoma, sarcoma, lymphoma, Hodgkin's disease, leukemia, melanoma, metastasis, neuroma, adenoma

HIV/AIDS: HIV, AIDS

Diabetes: diabetes, diabetes mellitus, high blood sugar, juvenile diabetes (Type I diabetes), adult-onset diabetes (Type II diabetes), diabetic neuropathy

Emphysema: emphysema, chronic obstructive pulmonary disease (COPD), chronic bronchitis (MUST use the word 'chronic', only for adults), Chronic obstructive bronchitis (MUST use the word 'chronic', only for adults), smokers cough

High Cholesterol: high cholesterol, high or elevated triglycerides, hyperlipidemia, hypercholesterolemia

Hypertension: hypertension, high blood pressure, ischemic heart disease, angina, angina pectoris, coronary artery disease, blocked, obstructed, or occluded coronary arteries, arteriosclerosis, myocardial infarction, heart attack

Stroke: stroke, cerebral hemorrhage, cerebral aneurysm, transient ischemic accident, transient ischemic attack, apoplexy, carotid artery blockage, arterial thrombosis in brain, blood clot in brain

Chronic, manageable conditions:

Arthritis: rheumatoid arthritis, degenerative arthritis, osteoarthritis, bursitis, rheumatism

Back Problems of Any Kind: back problems or pain of any kind, (lower or upper back), sore, hurt, injured, or stiff back, backache, 'vertebrae', 'lumbar', 'spine', or strained or pulled muscle in back, sprained back, muscle spasms, bad back, lumbago, sciatica or sciatic nerve problems disc problems: herniated, ruptured, dislocated, deteriorated, or mis-

aligned discs, 'spinal', back spasms, slipped, compressed, extruded, dislocated, deteriorated, or misaligned discs

Asthma: anything with the word 'asthma' or 'asthmatic'

Gall Bladder Disease: gall bladder disease, trouble, attacks, infection, or problems, gallstones

Stomach Ulcers: stomach ulcer, duodenal ulcer, peptic ulcer, bleeding ulcer, ulcerated stomach, perforated ulcer

Mental Health Issues

Alzheimer's Disease and Other Dementias: anything with the words 'Alzheimer's' or 'dementia', organic brain syndrome

Depression and Anxiety Disorders: depression (including severe, chronic, or major depression), dysthymia, dysthymic disorder, bipolar disorder, manic depression or manic depressive illness, anxiety attacks, panic attacks, anxiety, nerves, nervous condition, nervous breakdown

In the text we also discuss a number of specific health conditions (see table 3). They are defined as follows:

CANCER
Cancer of head and neck, esophagus, stomach, colon, liver and intrahepatic bile, lung/bronch/other intrathora, bone and intraconnective tissue, melanomas of skin, other non-epithelial, cancer of skin, breast, uterus, cervix, other female genital organs, prostate, bladder/kidney/renal pelvic, brain and nervous system

VIRAL INFECTION
Viral infection

HEADACHE
Headache, including migraines

CARDIAC CONDITION
Heart valve disorders, peri-, endo-, and myocarditis, cardiomyo, hypertension and hypertension with complications, acute myocardial infarction, coronary atheroscelrosis and other heart, nonspecific chest pain, pulmonary heart disease, other and ill-defined heart disease, conduction disorders, cardiac dysrhythmias, cardiac arrest, and ventricular fibrillation, congestive heart failure, nonhypertensive

UPPER RESPIRATORY INFECTION
Acute and chronic tonsillitis, acute bronchitis, other upper respiratory infections, chronic obstructive pulmonary disease

RESPIRATORY DISEASE
Lung disease due to external agents, other lower respiratory disease, other upper respiratory disease

SKIN DISEASE
Skin and subcutaneous tissue, other inflammatory conditions, chronic ulcer of skin, other skin disorders

INTESTINAL DISEASE
Intestinal infection

ARTHRITIS
Infective arthritis and osteomyelitis, rheumatoid arthritis and related disease, osteoarthritis and other non-traumatic joint disorders

Notes

1. See Internal Revenue Service Code section 162(1).

2. See, e.g., Ross and Mirowsky 2000. Sorlie et al. (1994) found that individuals covered by Medicare or Medicaid have 1.6 times the mortality rate of the uninsured, after controlling for age, sex, race, and income. We conjecture that this result might be due to unobservable heterogeneity—individuals who end up on Medicaid differ in important ways from those who do not, even after taking observable covariates into account. The possibility of a similar issue rises in our context, and we deal with it in some detail below.

3. However, Meer, Miller, and Rosen (2003) argue that the causal relationship running from wealth to health status disappears once the endogeneity of wealth is taken into account.

4. A distinct but closely related question is how inequality in income affects health outcomes; see, e.g., Deaton and Paxson 1999.

5. Additional confirmation of this finding is reported in Hurd and McGarry 1995.

6. CHAMPUS is a health benefits program designed to provide medical coverage for the dependents of active duty military servicemen/women. CHAMPVA is intended for dependents and survivors of severely disabled veterans.

7. The variable equals 1 if the respondent has had any activities of daily living, instrumental activities of daily living, or functional or sensory limitations in the past year.

8. The appendix to this chapter provides the details of these variables' construction.

9. The regional classifications correspond to those used by the Census Bureau.

10. As an experiment, we estimated our canonical model including income on the right-hand side. We found that while income was positively related to insurance coverage and health status, our substantive results did not change. In the same spirit, we also augmented the equation with dichotomous variables for the industry in which the individual worked. This, too, left our substantive results unchanged.

11. For this exercise, we drop observations for which total family income is below $5,000, operating on the assumption that measured income is not a good index of ability to pay.

Such families might either have substantial income in kind, or own businesses that create accounting losses.

12. For further evidence along these lines, see McCormick et al. 1993; Dadds, Stein, and Silver 1995.

13. We create a dichotomous variable equal to 1 if any child aged 4 or under in the family is limited in any way, including play activity, because of an impairment or a physical or mental health problem.

14. These results are cited in HPR 1996. For more detailed documentation, see National Bureau of Economic Research working paper 4880 (1994).

15. We also examined one-year transitions, and the results were essentially the same.

16. The employment status and demographic information were recorded at the beginning of 1996, and the health information was recorded in the middle of that year.

17. We examined whether the utilization of any of the following medical services was a good predictor of a transition into or out of self-employment: cholesterol exam, breast exam, blood pressure test, physical exam, flu shot, mammogram, prostate exam, doctor visit, hospital admission, and purchase of prescription medicine.

18. For a careful analysis of the financial effect of health insurance, see Levy 2002.

References

Campling, R. F. 1987. Employee Benefits and the Part-Time Worker. School of Industrial Relations Research Essay Series No. 13, Queen's University Industrial Relations Centre, Kingston, Ontario.

Currie, Janet, and Jonathan Gruber. 1995. Health Insurance Eligibility, Utilization of Medical Care, and Child Health. Working paper 5052, National Bureau of Economic Research.

Currie, Janet, and Duncan Thomas. 1995. Medical care for children: Public insurance, private insurance, and racial differences in utilization. *Journal of Human Resources* 30, winter: 135–162.

Dadds, Mark R., Ruth E. K. Stein, and Ellen Johnson Silver. 1995. The role of maternal psychological adjustment in the measurement of children's functional status. *Journal of Pediatric Psychology* 20: 527–544.

Deaton, Angus, and Christina Paxson. 1999. Mortality, Education, Income and Inequality among American cohorts. Working paper 7140, National Bureau of Economic Research.

Ettner, Susan L. 1996. New evidence on the relationship between income and health. *Journal of Health Economics* 15: 67–85.

Fairlie, Robert W., and Bruce D. Meyer. 1999. Trends in Self-Employment Among White and Black Men: 1910–1990. Working paper 7182, National Bureau of Economic Research.

Fuchs, Victor. 1998. Health, Government, and Irving Fisher. Working paper 6710, National Bureau of Economic Research.

Gruber, Jonathan. 2000. Medicaid. Working paper 7829, National Bureau of Economic Research.

Gruber, Jonathan, and James M. Poterba. 1994. Tax incentives and the decision to purchase health insurance: Evidence from the self-employed. *Quarterly Journal of Economics* 109, August: 701–733.

Hamilton, Barton Hughes. 2000. Does entrepreneurship pay? An empirical analysis of the returns to self-employment. *Journal of Political Economy* 108, June: 604–631.

Health Insurance Association of America. 2000. *Source Book of Health Insurance Data, 1999–2000*.

Holtz-Eakin, Douglas, John Penrod, and Harvey S. Rosen. 1996. Health insurance and the supply of entrepreneurs. *Journal of Public Economics* 62: 209–235.

Hurd, M., and K. McGarry. 1995. Evaluation of the subjective probabilities of survival in the health and retirement survey. *Journal of Human Resources* 30: S268–S292.

Idler, Ellen, and Yael Benyamini. 1997. Self-related health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior* 38, March: 21–37.

Institute for the Future. 2000. *Health and Health Care 2010—The Forecast, The Challenge*. Wiley.

Kaestner, Robert, Theodore Joyce, and Andrew Racine. 1999. Does Publicly Provided Health Insurance Improve the Health of Low-Income Children in the United States? Working paper 6887, National Bureau of Economic Research.

Levy, Helen. 2002. The Economic Consequences of Being Uninsured. Economic Research Initiative on the Uninsured working paper 12, University of Michigan.

McCormick, Marie C., Jeanne Brooks-Gunn, Kathryn Workman-Daniels, and George J. Peckham. 1993. Maternal rating of child health at school age: Does the Vulnerable Child Syndrome persist? *Pediatrics* 92, September: 380–388.

McDonough, Peggy, Greg J. Duncan, David Williams, and James House. 1997. Income dynamics and adult mortality in the United States, 1972 through 1989. *American Journal of Public Health* 87, September: 1476–1483.

Meara, Ellen. 2001. Why Is Health Related to Socioeconomic Status? Working paper 8231, National Bureau of Economic Research.

Menchik, Paul L. 1993. Economic status as a determinant of mortality among black and white older men: Does poverty kill? *Population Studies* 47: 427–436.

Meer, Jonathan, Douglas Miller, and Harvey S. Rosen. 2003. Exploring the Health-Wealth Nexus. Working paper 9554, National Bureau of Economic Research.

Monheit, Alan C., and P. Holly Harvey. 1993. Sources of health insurance for the self employed: Does differential taxation make a difference? *Inquiry (Rochester)* 30 (3) (fall): 293–305.

Newhouse, Joseph, and the Insurance Experiment Group. 1993. *Free for all? Lessons from the RAND Health Insurance Experiment.* Harvard University Press.

Perry, Craig W., and Harvey S. Rosen. 2004. Insurance and the utilization of medical services among the self-employed. In *Public Finances and Public Policy in the New Millennium*, ed. S. Cnossen.

Preston, S. E., and Paul Taubman. 1994. Socioeconomic differences in adult mortality and health status. In *Demography of Aging*, ed. L. Martin and S. Preston. National Academy Press.

Ross, Catherine E., and John Mirowsky. 2000. Does medical insurance contribute to socioeconomic differentials in health? *Milbank Quarterly* 78, no. 2: 291–321.

Ruhm, Christopher. 2000. Are recessions good for your health? *Quarterly Journal of Economics* 115, May: 617–650.

Smith, James P. 1999. Healthy bodies and thick wallets: The dual relation between health and economic status. *Journal of Economic Perspectives* 13, no. 2: 145–166.

Smith, James P., and Raynard S. Kington. 1997. Race, socioeconomic status, and health in late life. In *Racial and Ethnic Differences in the Health of Older Americans*, ed. L. Martin and B. Soldo. National Academy Press.

Sorlie, P. D., N. J. Johnson, E. Blacklund, and D. D. Bradham. 1994. Mortality in the uninsured compared with that in persons with public and private health insurance. *Archives of Internal Medicine* 154: 2409–2416.

Taubman, Paul, and Sherwin Rosen. 1982. Healthiness, Education, and Marital Status. Working paper 0611, National Bureau of Economic Research.

Thomas, Cynthia, Howard R. Kelman, Gary J. Kennedy, Chul Ahn, and Chun-yong Yang. 1992. Depressive symptoms and mortality in elderly persons. *Journal of Gerontology: Social Sciences* 47, no. 2: S80–S87.

Verbrugge, Lois M. 1985. Gender and health: An update on hypothesis and evidence. *Journal of Health and Social Behavior* 26, no. 3: 156–182.

Verhovek, Sam Howe. 2000. Frustration grows with cost of health insurance. *New York Times*, September 18.

3 Business Formation and the Deregulation of the Banking Industry

Sandra E. Black and
Philip E. Strahan

Before the 1980s, banks in the United States were subject to a wide range of regulations that limited activities, constrained pricing, and restricted the ability to expand both within and across state lines. Many of these regulations have a long history; for example, restrictions on bank branching originated in the nineteenth century. The legacy of this system lasted into the 1970s. In 1976 only 14 states permitted banks to open branches across the state; the other 36 states either restricted branching to the city in which a bank's head office was located, or prohibited branching altogether. Similarly, in 1976 no state permitted an out-of-state banking company to buy banks headquartered in the state (table 1).

Starting in the latter half of the 1970s, the U.S. banking system began to be reshaped, both by technological innovations and by the removal of many of these constraining regulations. In the early 1980s, for example, interest-rate ceilings were largely removed, allowing banks to compete more vigorously for funds. New technologies like the automated teller machine also enhanced competition within banking, and innovations such as the cash management account offered by non-bank financial companies enhanced competitive pressures from outside the industry. During the same period, restrictions on banks' ability to expand into new markets were lifted by state-level legislative initiatives allowing branching across the state and allowing interstate banking—that is, cross-state ownership of bank assets (Jayaratne and Strahan 1998). By the early 1990s, almost all states had removed their restrictions on branching and interstate banking. These changes were codified at the national level in 1996 when Congress passed the Interstate Banking and Branching Efficiency Act. Banks may now branch not only within states but also across state lines, and bank holding companies may buy banks anywhere in the United States.

Table 1
Trends in openness of the banking market. (A state is defined as permitting branching if banks may purchase branches anywhere across the state. A state is defined as permitting interstate banking if it allows out-of-state bank holding companies to buy its banks.)

	Number of states permitting statewide branching	Number of states permitting interstate banking
1976	14	0
1977	15	0
1978	16	0
1979	17	1
1980	18	1
1981	19	1
1982	21	1
1983	22	1
1984	23	5
1985	24	8
1986	28	18
1987	30	28
1988	35	39
1989	41	43
1990	42	45
1991	46	46
1992	48	48
1993	48	49
1994	49	50

These technological and regulatory changes enhanced the openness and competitiveness of banking markets and, at the same time, set the stage for rapid growth of expansion-minded banks. Table 2 summarizes the effects of these changes on the market share of small banks and on the concentration of the banking market. Nationwide consolidation in banking has been going on for many years, and, as the table shows, small banks have been losing ground consistently over the past 20 years. In the mid 1970s, banks with assets under $100 million (in 1993 dollars) held about 24 percent of all assets, while banks with under $500 million in assets held about 48 percent of the total. By the mid 1990s, these shares had fallen to 15 percent and 34 percent, respectively. Over the same time, there has been no hint that this consolidation has increased concentration, or retarded competition, in local banking markets.[1] As table 2 shows, the Hirfindahl-Hirschmann Index

Table 2
Trends in small bank market share and local concentration (unweighted average across states of the share of assets held by small banks and the local Herfindahl-Hirschmann Index). A small bank here is defined as a bank with $100 ($500) million in assets or less, in 1993 dollars. The local HHI equals the sum of squared deposit market shares across all banks operating in a Metropolitan Statistical Area (MSA). For states with more than 1 MSA, we average the local HHI across all MSAs, weighted by total deposits in each MSA.

	Share of assets held by banks with assets less than $100 million	Share of assets held by banks with assets less than $500 million	Local deposit-based HHI
1976	0.239	0.477	0.198
1977	0.237	0.472	0.192
1978	0.227	0.463	0.187
1979	0.225	0.464	0.184
1980	0.229	0.464	0.183
1981	0.235	0.466	0.186
1982	0.232	0.467	0.189
1983	0.223	0.466	0.190
1984	0.211	0.448	0.184
1985	0.203	0.435	0.192
1986	0.195	0.417	0.192
1987	0.182	0.402	0.196
1988	0.183	0.395	0.196
1989	0.180	0.385	0.191
1990	0.173	0.380	0.193
1991	0.170	0.373	0.198
1992	0.164	0.365	0.192
1993	0.157	0.358	0.196
1994	0.150	0.344	0.190

of concentration in local markets has remained very constant over this long period of deregulation. Banks have been expanding into new markets rather than combining forces with other banks in their old markets.

In this chapter, we study how these changes in the structure of the banking industry have affected the availability of bank credit and, as a consequence, have affected the rate of creation of businesses. We are motivated by the idea that bank lending is especially important for firms very early in their life cycle. Without credit, young firms starve and die. With it, they have a chance to grow and prosper. We show first that bank lending increased significantly after deregulation of

both restrictions on bank branching and restrictions on interstate banking. In addition, changes associated with deregulation (for example, the decline in the prevalence and market share of small banks) have been associated with increased lending.

We then link this increase in bank loan supply to the rate of growth in the number of new businesses—measured by the amount of newly incorporated businesses in each state. This builds on earlier work (Black and Strahan 2002) in which we focused on the direct relationship between bank structure and business creation; in this chapter, we turn our attention to the channel through which this mechanism works: bank lending. Our bottom line is that the technological and deregulatory changes in banking of the past 20 years have been good for entrepreneurs looking to start businesses. The growth in new incorporations is positively related to bank lending, and this positive association seems to reflect supply-side factors. Our estimates suggest that a one-standard-deviation increase in bank lending, an increase of about 10 percent, is associated with an increase of 2–3 percentage points in the growth rate of new incorporations.

How Finance Affects Business Formation

Liquidity Constraints and Business Formation

There is a large literature which suggests that finance is important to entrepreneurs looking to start businesses. Liquidity constraints place important roadblocks before potential entrepreneurs; individuals with more assets, for instance, are more likely to become self-employed and to succeed in small businesses. Evans and Jovanovic (1989) find that individuals with more assets are more likely to become self-employed. Holtz-Eakin, Joulfaian, and Rosen (1994a,b) find that individuals who have received large inheritances are more likely to succeed in running small businesses, and Holtz-Eakin and Rosen (1999) find that entrepreneurial activity in Germany is retarded relative to the United States by limited access to capital. Gentry and Hubbard (2000a) report that entrepreneurial households hold a substantial share of overall household wealth, and that nonbusiness assets helps predict the likelihood and success of entrepreneurial activity. Huck et al. (1999) find that new businesses rely heavily on credit from informal sources such as business contacts and family, and Avery et al. (1999) find that bank loans to small businesses tend to be personally guaranteed. Fairlie (1997)

Table 3
The importance of banks to small business. Source of data: 1993 *National Survey of Small Business Finance*. See Cole and Wolken 1995 for details.

Number of full-time equivalent employees	Percentage of small firms using			
	Any commercial bank service	A checking account	Any credit facility	A line of credit
0–1	81	90	42	16
2–4	90	97	55	23
5–9	93	98	67	32
10–19	96	99	76	40
20–49	97	99	78	53
50–99	96	99	86	56
100–499	99	99	88	60

finds a lower level of minority-owned businesses, in part because of minorities' lower levels of wealth.[2]

Competition, Banks, and Small-Business Lending

Many studies have shown that the creation of businesses is bounded by liquidity constraints, but there has been little work focusing on how the structure of the banking sector affects entrepreneurship. We do know that banks and banking services are important to small and young firms, which suggests a link between bank structure and business creation. Nearly 90 percent of even the smallest businesses use banking services. Most have a checking account, and almost half of businesses with fewer than two employees have a credit facility of some kind from a bank or other financial institution (table 3).

Not only do small businesses borrow from banks, but they also tend to concentrate their borrowing at a single bank with which they have a long-term relationship. The nature of these relationships is an important feature of small-business lending; long-term relationships enable banks to collect private information on the credit worthiness of small firms. Recent evidence suggests that the credit availability is enhanced when banks forge relationships with small businesses. Petersen and Rajan (1994) find that small firms that have established a relationship with a bank are less likely to use expensive trade credit. They find very weak effects on loan interest rates, however, suggesting that there may be some credit rationing for firms that have not established a banking

relationship. On the other hand, Berger and Udell (1995) find that small firms with banking relationships pay lower interest rates on one narrowly defined type of loans, the line of credit.[3] Our focus below will be on the effect of bank credit supply (i.e., quantity of loan growth) on the rate of formation of businesses. We are unable, however, to test for credit rationing, because we do not have that data to measure loan interest rates to businesses.

As we noted in our introduction, there has been a recent trend toward increased competition in the banking sector, and a number of studies have questioned how these developments toward increased competition will affect relationship lending. (For a review, see Boot and Thakor 2000.) Banks are no longer protected from competition by barriers to in-state branching and interstate banking. Moreover, non-bank financial institutions have become increasingly important providers of credit to new businesses. Competition makes it easier for borrowers to switch lenders, which can reduce the incentive to invest in relationships at the outset. On the other hand, Boot and Thakor (ibid.) argue that competition may raise the rewards to activities that allow lenders to differentiate themselves from other lenders, thereby raising the incentive to invest in relationships.

Developments toward greater competition have probably reduced the costs of providing credit on average. Conventional analysis of market power would clearly predict that more market openness and an expansion of the number of competitors should lead to reduced prices, making customers better off. In fact, Jayaratne and Strahan (1998) find declines in average loan prices of about 40 basis points after branching deregulation overall, although they do not look at lending to business. According to this simple view, entrepreneurial activity ought to be enhanced by increased competition in banking. This view, however, does not account for the importance of relationships in allowing banks and other lenders to extend credit to potential entrepreneurs.

Petersen and Rajan (1995) present a model in which market power *helps* new businesses by allowing banks to forge long-term relationships with them. They argue that with market power, banks can subsidize borrowers during some periods because they can extract rents during other times. In competitive markets, however, firms have access to alternative sources of credit, so banks cannot offer low prices early on because they lack the market power to recover those investments later. As evidence, they show that in concentrated banking markets

interest rates on bank loans tend to be higher as the length of a relationship increases, suggesting some intertemporal cross subsidization. In less concentrated markets, however, they find no effect of the length of a relationship on bank loan rates.[4] Bonaccorsi di Patti and Dell'Ariccia (2001) provide further evidence along these lines. They find that Italian firms that are more opaque (e.g., firms with fewer physical assets) may benefit more (or are harmed less) from concentrated banking markets than firms that are less opaque.

Other evidence is less supportive that competition reduces the incentive for banks to invest in private information and make relationship loans. While Cetorelli and Gambera (2001) find that industries that rely heavily on external finance grow faster in countries with concentrated banking systems than they do in countries with more open and competitive banking, they find a *negative* overall effect of banking concentration on economic growth. Fisman and Raturi (2000) use data from five African countries to show that trade credit is more prevalent when suppliers are in competitive industries.

In view of the uncertainty in both the theoretical and empirical literature, enhanced competition could plausibly help *or* hinder entrepreneurs' access to credit. Our empirical tests attempt to resolve this uncertainty by looking directly at how changes in the structure of the U.S. banking industry that enhanced the openness of markets and raised competitiveness have affected lending overall, and then how the associated changes in lending have affected business formation.

Consolidation and Entrepreneurial Activity

At the same time that it enhanced competition, deregulation and consolidation in banking have led to a decline in the importance of small banks (table 2). A number of recent studies have argued that small banks possess a better technology for relationship lending than large banks. Berger and Udell (1996), for example, argue that because of the importance of long-term financial relationships, the technology of lending to small businesses differs fundamentally from the technology of other types of lending. Larger firms with well-established track records may be able to borrow based on readily observable information. Similarly, most residential real estate as well as consumer lending is now based on credit scoring models. On the other hand, small-business (or "relationship") loans may require tighter control and oversight over loan officers by senior management than do loans

based on simple ratio analyses or credit scoring models. As a consequence, the complexity of large banks may lead to organizational diseconomies that make relationship loans more costly. They suggest that senior management of small banks can monitor lending decisions closely, so they can authorize more non-standard, relationship loans.[5]

The stylized fact that motivates this idea is that small banks hold a larger fraction of their assets in small-business loans than large banks do. However, this cross-sectional pattern may reflect small banks' inability to lend to large firms, rather than large banks' inability to lend to small firms. A small bank can remain well diversified only if it avoids large loans. Moreover, regulations restrict bank lending to a single borrower to 10 to 15 percent of capital (Spong 2000). So, for instance, regulations prevent a bank with $100 million in assets (a small bank) and $10 million in capital from making any loan greater than $1.5 million.

Since the cross-sectional relationship between bank size and small-business lending is difficult to interpret, a number of recent papers have estimated the effects of mergers and acquisitions on small-business lending. However, the results have been mixed. Some papers find that lending to small businesses increases when small banks are acquired, suggesting the increased scale increases a bank's willingness to make relationship loans, while others find declines in lending after mergers.[6]

Strahan and Weston (1998) argue that size-related diversification may offset the potential organizational diseconomies in relationship lending. Diamond (1984) shows theoretically that the costs associated with delegating the monitoring of borrowers from the principal (depositors) to the agent (the bank) decline with diversification because diversification makes the bank more transparent to the depositor. A large bank's superior ability to diversify credit risks across borrowers reduces the (agency) cost of lending to risky and opaque borrowers. Thus, large banks may be lower cost lenders generally than smaller banks.[7]

Finally, our earlier work (Black and Strahan 2002) focuses on the relationship between competition and consolidation on new incorporations and concludes that increased competition is associated with higher levels of new incorporations. In addition, consolidation appears to help entrepreneurs; states with more large banks experience a higher level of incorporations. These results suggest that the diversification

benefits of consolidation and greater bank size outweigh the possible advantages small banks may have in forming long-term relationships.

In this chapter, we focus more specifically on lending. If small banks really can provide relationship loans at lower cost than large ones, we ought to find that recent consolidation in banking, and the associated decline in small banks, has reduced bank lending that supports entrepreneurial activity and business formation. In contrast, if large banks are lower cost lenders than small ones overall, and if there are no important diseconomies in relationship lending, then we ought to see just the opposite.

Empirical Methods and Data

We start by estimating a reduced-form model to test how overall bank lending in a state depends on measures of the banking environment in that state. Because states deregulated their restrictions on branching and interstate banking at different times, we can estimate the regulatory effects on lending in a panel data set in which we control for state and time fixed effects. In addition, we also test how changes in banking structure affected lending. Our study runs over a long time period, from the mid 1970s until 1994, so that we can take advantage of the broad changes in banking emphasized earlier. The study ends in 1994 because banks began to operate across state lines after that year, making it impossible to measure our banking structure variables by state.[8]

After estimating the reduced-form model for lending, we then link bank lending to the rate of business formation in an instrumental variables regression and find that the large changes in banking structure over this period had large effects on lending.

The Reduced-Form Bank Lending Relationship

Our reduced-form model of bank lending includes both demand- and supply-side variables. On the demand side, we include both state and national measures of the business cycle. On the supply side, we include measures of the regulatory environment, measures of the structure of the banking industry (bank size and local market concentration), and measures of bank financial condition. To measure bank lending to businesses in a state and year, we sum all commercial and industrial

loans and commercial real estate loans made by all banks headquartered in the state. These data come from the fourth-quarter *Reports of Income and Condition* (the "Call Reports").

We capture the effects of state-level deregulation of restrictions on geographical expansion by including an indicator equal to 1 after a state permits branching by merger and acquisition within its borders, and another indicator equal to 1 after a state permits interstate banking (that is, after a state allows bank holding companies in other states to buy their banks).[9] The effects of other kinds of national deregulation that occurred during the period, such as removal of the Regulation Q interest-rate ceilings or the introduction of risk-based capital requirements, will be absorbed in the model by the annual fixed effects. In addition, common technological trends like the growth of ATMs will also be absorbed by these fixed effects.

In addition to looking directly at how deregulation, and the associated increase in market openness, affected lending, we also include the deposit Herfindahl-Hirschmann Index (HHI) as a measure of competition in local markets. The HHI is equal to the sum of the squared share of deposits held by each bank operating in a local market, defined as a Metropolitan Statistical Area (MSA). To go from the local level to the state level, we average the HHIs across all MSAs in a state, weighted by total deposits in each MSA. The information on deposits by MSA and bank are based on branch-level data from the FDIC's *Summary of Deposits*. Of course, one might prefer to use a loan-based measure of market concentration for these purposes, since we are interested in how competition affects lending. Unfortunately, unlike deposits, loan data are not available at the branch level, making it impossible to compute MSA-level market shares based on loans.

To test whether consolidation, and the associated decline in small banks' market share, has raised or lowered the rate of business formation through its effects on the supply of relationship loans, we include the share of total assets in a state held by small banks. In one set of specifications, we define a bank as "small" if it holds $100 million or less in assets (in 1993 dollars). In the other specification, we define a bank as "small" if it holds $500 million or less in assets. Data on bank size come from the year-end Call Reports.

We also consider whether the financial health of the banking industry affects the rate of business creation. In an environment where their liabilities are insured, weak banks have an incentive to look for risky

lending opportunities, such as lending to new businesses.[10] Depositors holding claims at poorly capitalized banks have little or no incentive to prevent this risk-seeking behavior. This moral-hazard problem became severe during the early and mid 1980s in the thrift industry here in the United States. In contrast, banks may reduce their risky lending (and hence businesses formation) in response to a "capital crunch." Partly in response to concerns about bank solvency, the Basle Accord of 1988 led to formal capital adequacy standards for all internationally active banks. The accord tightened capital standards and linked these standards explicitly to a bank's portfolio risk (Demsetz and Strahan 1995).[11] In addition, concern about banking and thrift solvency in the United States led to passage of the Financial Institutions Reform, Recovery and Enforcement Act in 1989 and the FDIC Improvement Act in 1991. Each of these laws tightened the regulation of financial institutions in the United States, in part to mitigate perceived problems with deposit insurance and financial institutions' propensity to take risks. The greater emphasis on capital regulations suggests that poorly capitalized banks may have lent less than well-capitalized banks to risky, small businesses.

To test how bank financial condition affects lending, we introduce two market share variables denoting the share of assets held by banks with different capital-asset ratios. First, we include the share of a state's assets held by critically undercapitalized banks—banks with a capital-asset ratio below 2 percent. Second, we include the share of assets held by banks that are weakly capitalized but not in immediate danger of failing—banks with a capital-asset ratio between 2 percent and 6 percent. Banks with a capital-asset ratio above 6 percent are omitted from the equation. The coefficients therefore measure how lending changes when a given share of assets moves from the >6 percent group to the group in question.

We also include variables to control for demand conditions. First, we use personal income growth in the state, collected from the Bureau of Economic Analysis, to account for business-cycle factors, along with two lags of this variable. Second, since better-educated people are more likely to start businesses, we include the share of workers in a state with a college degree or more. These data come from the March *Current Population Survey*.[12] Third, we include both state and time fixed effects.[13] For all of the regressions, all but one of our explanatory variables are measured as of the end of the year preceding the year in which

we measure the rate of business formation. The one exception is the personal income growth variable, which is measured during the same year as the dependent variable. We also include lags of this variable.

Business Formation and Bank Lending

We use new incorporations in each state and year from 1976 to 1994 as our measure of business formation.[14] This series comes from the individual states and is reported by Dun & Bradstreet. Of course, business incorporations is not a perfect proxy for the rate of business formation in a state; however, it is the best proxy available that is compiled on a consistent basis over a relatively long period.

Dun & Bradstreet also report a series on business starts that is an offshoot of their credit database. Since this series only goes back to 1985, it is not helpful in exploring how the changes in banking that began in the mid 1970s affected entrepreneurship and business formation. Nevertheless, we can use the starts data to test whether business incorporations provides a useful proxy for the rate of business formation in a state. Table 4 shows that new incorporations per capita and business starts per capita are consistently positively correlated with each other; the cross-state correlation ranged from a low of 0.58 in 1994 to a high of 0.72 in 1988. There is one important exception, however. The number of incorporations in Delaware is about 20 times the aver-

Table 4
Cross-state correlation between business starts per capita, new incorporations and new establishments.

	Correlation between		
	Starts and incorporations	Starts and new establishments	Incorporations and new establishments
1985	0.62	—	—
1986	0.64	—	—
1987	0.64	—	—
1988	0.72	—	—
1989	0.64	0.65	0.57
1990	0.62	0.52	0.52
1991	0.66	0.44	0.52
1992	0.61	0.41	0.54
1993	0.65	0.46	0.54
1994	0.58	0.55	0.54

age number of incorporations in the other states (per capita), while the number of starts in Delaware is very close to the average. This difference reflects favorable legal treatment of incorporations in that state. In addition, measures of banking structure in both Delaware and South Dakota are skewed by the presence of credit card banks in those states. We therefore drop both of these states from all of our regressions.[15]

As a further check on the data, we compared incorporations per capita and starts per capita with the number of new establishments per capita, which is available from the Small Business Administration starting in 1989. An establishment is not a firm; rather, it is an economic unit that employs people, such as a plant, a factory, or a restaurant. Nevertheless, we think that the number of new establishments ought to be highly correlated with the economic quantity that we are trying to observe—the rate of creation of businesses. Again, it is highly correlated with both incorporations and starts. From 1989 to 1994, the cross-state correlation between incorporations and new establishments ranges from 0.52 to 0.57, and cross-state correlation between starts and new establishments ranges from 0.41 to 0.65 (table 4). This suggests that using new incorporations in a state may be a good proxy for business formation.

Results

Reduced-Form Lending Results

Before turning to the results, table 5 reports summary statistics for the variables in our model. Growth in business incorporations averaged 4.66 percent per year, while the growth in bank loans to businesses averaged 8.67 percent and the growth in personal income averaged 8.26 percent. The higher income and loan growth rates reflect the fact that these variables are computed in dollar terms, so part of the increase reflects inflation.[16] The average for the post-branching indicator is 0.564, meaning that 56.4 percent of our state/year observations occurred after branching deregulation and the rest before deregulation. Similarly, the interstate banking indicator averages 0.42. The deposit HHI index averaged 0.191 during our sample; to understand what this means, consider that a local MSA with 5 equally sized banks would have an HHI of 0.2. The share of bank assets held by small banks averaged 0.203 when a "small bank" is defined as one with less than $100 million in assets, and 0.426 when a "small bank" is defined as one with

Table 5
Summary statistics. Business loans equals commercial and industrial loans plus commercial real estate loans. The Herfindahl-Hirschmann Index is the sum of squared market shares based on deposits for all MSAs in the state. For states with more than one MSA, we average this across MSAs weighted by depositors. The post-branching indicator equals 1 during the years after a state permits branching by merger and acquisition; the post-interstate banking indicator equals 1 during the years after a state permits interstate banking.

	Mean	Standard deviation
Growth in new incorporations	4.66%	12.55%
Annual business loan growth	8.67%	10.07%
Personal income growth	8.26%	3.96%
Post-branching indicator	0.564	0.496
Post-interstate banking indicator	0.420	0.494
Deposit HHI (average of MSAs in state)	0.191	0.067
Share of bank assets held by small banks (under $100 M)	0.203	0.173
Share of bank assets held by small banks (under $500 M)	0.426	0.283
Share of assets in banks with capital <2% of assets	0.005	0.020
Share of assets in banks with capital between 2% and 6% of assets	0.310	0.229
Share of population with college degree	0.255	0.048

less than $500 million in assets. The share of assets held by banks with very low capital (under 2 percent of assets) averaged 0.005, and the share of assets held by banks with low capital (between 2 percent and 6 percent) averaged 0.310. Finally, the share of workers with a college degree averaged 0.255.

Table 6 reports the reduced-form relationship between business lending and our demand-side and supply-side proxies. We estimate the relationship using the growth rate in total business lending in the state as the dependent variable.[17]

The results provide very clear evidence that the broad trends toward more competition and greater consolidation have increased the availability of bank loans to businesses. The first column of the table reports the results with just loan demand controls (state and time fixed effects and personal income growth), along with the two deregulation indicator variables. Here, we find that loan growth increased significantly after both branching deregulation and interstate banking deregulation.[18] Moreover, the estimated effects of deregulation are quantitatively, as well as statistically, significant. For example, the regression coefficients in column 1 for the intrastate branching indicator of 0.029 suggests that loan growth increased by 2.9 percentage points after

Table 6
Panel regression relating business lending to banking deregulation indicators and measures of banking structure. Regressions include state fixed effects as well as a set of annual fixed effects. In addition, the regressions include state-level personal income growth along with two lags. Standard errors are reported in parentheses. Asterisk denotes statistical significance at 10% level. The Herfindahl-Hirschmann Index is the sum of squared market shares based on deposits for all MSAs in the state. For states with more than one MSA, we average this across MSAs weighted by depositors. The post-branching indicator equals 1 during the years after a state permits branching by merger and acquisition; the post-interstate banking indicator equals 1 during the years after a state permits interstate banking.

Dependent variable	Growth in business loans		
Post-branching indicator	0.029*	0.027*	0.028*
	(0.009)	(0.009)	(0.009)
Post-interstate banking indicator	0.025*	0.021*	0.022*
	(0.011)	(0.011)	(0.010)
Deposit HHI (average of MSAs in state)	—	0.042	0.026
		(0.095)	(0.095)
Fraction of assets in small banks (<$100 M)	—	−0.131*	—
		(0.067)	
Fraction of assets in small banks (<$500 M)	—	—	−0.060
			(0.042)
Share of assets in banks with capital <2% of assets	—	−0.883*	−0.896*
		(0.136)	(0.137)
Share of assets in banks with capital between 2% and 6% of assets	—	0.046*	0.047*
		(0.018)	(0.018)
N	823	823	823
R^2 (within)	0.57	0.61	0.60

branching deregulation; the coefficient on the interstate banking indicator of 0.025 suggests an increase of 2.5 percentage points in loan growth after interstate banking deregulation. Loan growth averaged about 8.7 percent per year over the whole sample, so these are very large increases. It is important to recognize, however, that these effects may not persist indefinitely. They more likely reflect an increase in the size of the banking industry after removal of long-standing constraints on the expansion of better-run banks into new markets. In fact, Jayaratne and Strahan (1996) find that state economies grew faster after deregulation too, but that these effects tended to wane somewhat over time.

The second and third columns of table 6 introduce the structure and financial condition variables to the model. These specifications suggest that the declining trend in the importance of small banks has, if anything, *increased* overall bank lending rather than reduced it. We do

not find that local market concentration is correlated with lending, although this could occur because there has been no trend toward more (or less) concentrated markets over time (table 2). The size result is considerably stronger when we define a small bank as one with less than $100 million in assets. In this case, a decline of one standard deviation in the small banks' share of assets in a state—that is, a decline of 0.173 in the small banks' share—leads to an increase in loan growth of about two percentage points. The coefficient on the share of assets in banks with less than $500 million is negative as well, but not statistically significant. These size effects support the view that large and better diversified banks have a lower cost of lending than small banks (Diamond 1984). These results do not, of course, tell us whether loans to small, relationship borrowers increased or decreased, since the data on these kinds of loans are not available as distinct from total business lending.

The financial condition of the banking industry also appears to have important effects on lending. At very low levels of capital, lending appears to be inhibited. For example, our estimates suggest that when a large share of a state's assets are held by very weak banks—banks with less than 2 percent capital as a fraction of assets—lending declines. A concrete example may help illuminate this result. In 1989, about 10 percent of the assets held by banks in Massachusetts were held by very poorly capitalized banks. During the subsequent year, business lending fell by 14 percent. This is an admittedly extreme example. Overall, our results suggest that a one-standard-deviation increase in the share of assets by the weakest banks, an increase of two percentage points, is associated with a decline in loan growth of about 1.6 percentage points.

The coefficients on the other capital-asset ratio market share variable suggest that these under-capitalized (but not critically undercapitalized) banks, if anything, lend somewhat more aggressively than banks that are well capitalized. For example, a one-standard-deviation increase in the share of assets by the banks in the range of 2–6 percent (relative to the over-6-percent range) is associated with an increase in loan growth of about 1 percentage point.

The Effect of Bank Lending on Business Formation

In table 7, we estimate the relationship between bank lending and the growth rate of new incorporations using both an ordinary least squares

Table 7
Instrumental variables regression relating growth of new incorporations to bank lending growth. Regressions include state fixed effects as well as a set of annual fixed effects. Standard errors are reported in parentheses. Asterisk denotes statistical significance at 10% level. In the IV model, the identifying instruments are post-branching indicator, post-interstate banking indicator, share of assets in small banks (banks with assets under $100 or $500 million), deposit-market concentration, and capital market share variables. See table 6 for the reduced-form relationship between lending and these instruments.

		IV		
	OLS	All instruments: small banks <$100 million	All instruments: small banks <$500 million	Deregulation indicators and HHI only
Growth of business loans	0.214* (0.057)	0.436* (0.199)	0.393* (0.200)	0.785* (0.411)
Personal income growth	0.755* (0.176)	0.492* (0.288)	0.534* (0.290)	0.060 (0.526)
Personal income $growth_{(t-1)}$	−0.079 (0.154)	−0.231 (0.198)	−0.203 (0.199)	−0.455 (0.309)
Personal income $growth_{(t-2)}$	−0.385* (0.144)	−0.518* (0.184)	−0.493* (0.184)	−0.724* (0.285)
Share of workers with a college degree or more	0.037 (0.179)	0.048 (0.183)	0.042 (0.182)	0.102 (0.200)
N	823	823	823	823
R^2	0.33	0.31	0.32	0.24

(OLS) and an instrumental variables (IV) procedure. Since an increase in entrepreneurs' desire to start businesses will likely increase the demand for bank loans (and other kinds of credit), the simple OLS relationship between lending and new incorporations may not reveal the extent to which an increase in bank credit *supply* can help spur business formation. That is, the OLS approach does not allow us to separate loan demand effects from loan supply effects. Moreover, we use total business loan growth, rather than loan growth to small and new firms, as the explanatory variable in the model due to data constraints. Thus, there is a potential measurement error problem in the OLS approach.

Our IV approach uses the banking deregulation indicators, structure variables, and financial condition variables as instruments that shift the supply of bank lending in a state. The idea of the IV procedure is to estimate the coefficient on loan growth using *only* variation in loan growth that has to do with supply-side factors. This way, we avoid the

potential upward bias in the coefficient that could result from the fact that an increase in entrepreneurial activity in a state will increase the amount of bank credit demanded to start businesses.[19] Of course, the key assumption that allows IV to work is that the instruments are really related to supply-side factors only. In our case, there may be a concern that the small bank market share variables and the bank financial condition variables could, at least in part, reflect changes in loan demand. For example, there may be an increase in the amount of business available to smaller banks when entrepreneurial activity is particularly robust. Moreover, there may be declines in bank financial condition at the same time that the entrepreneurial sector is weak. To account for these potential biases, we report our findings first with all of the instruments, and then with only the branching and interstate banking indicators and the deposit HHI index as instruments.

The IV approach also allows us to correct the measurement error problem associated with our use of total business lending rather than lending to small and new businesses. Since correcting measurement error would generally lead to a larger coefficient, we can not predict, a priori, whether the IV estimate ought to be larger or smaller than the OLS estimate.

The results in table 7 clearly suggest that business formation increases when bank lending becomes more easily available. In both the OLS and IV approaches, we find a positive and statistically significant effect of lending growth on the growth of new incorporations. The positive effects are robust across all four models. Even when we include just the deregulation indicators in the IV procedure, we continue to find a positive effect of bank loan growth on incorporations. In this last model, both the coefficient on loan growth, as well as its standard error, rise somewhat. We can not, however, reject the hypothesis that the coefficients from the restricted model (that is, the model using only deregulation indicators and the deposit HHI as instruments) are equal to the coefficients from the unrestricted IV models (Hausman 1978).

To understand the magnitude of the effects of lending, consider again the experience of Massachusetts. In 1994, lending in Massachusetts began to grow robustly again after the recession of the early 1990s, rising from just 1 percent in 1993 to 5 percent in 1994. At the same time, the growth in new incorporations rose from 4 percent in 1993 to 9.5 percent in 1994. These effects in Massachusetts, of course, do not necessarily reflect the typical experience. Nor does this simple

comparison allow us the distinguish shifts in lending supply from lending demand. Overall, the regression coefficient suggest that a one-standard-deviation increase in loan growth (about 10 percent) would be associated with an increase in new incorporations of 2–4 percent.

Conclusions

The banking industry has undergone a profound change over the past 25 years as a consequence of technological and regulatory innovations. These changes have created a much more open and competitive banking system. At the same time, large and expansion-minded banks have been able to increase their market share, leading to a dramatically consolidated industry structure at the national level. The theoretical and empirical literature speaks with one voice about how finance generally will affect business formation: better finance leads to more entrepreneurship. But the literature remains divided on the expected effects of increased competition and consolidation within the banking sector for borrowers that must establish a relationship with their bank. Our results suggests that policies such as branching and interstate banking reform, which fostered competition and consolidation in the banking sector, as well as the associated decline in the importance of small banks, increased lending overall, and that this increase in lending helped entrepreneurs start businesses.

Acknowledgments

We thank Rebecca Demsetz, Bill Gentry, Beverly Hirtle, Doug Holtz-Eakin, George Kaufman, Marc Saidenberg, and Kevin Stiroh for comments and David Fiore for his excellent research assistance. The opinions expressed here are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of New York or the Federal Reserve System.

Notes

1. For evidence that more concentrated banking markets are less competitive, see Berger, Demsetz, and Strahan 1999.

2. More broadly, Fazzari, Hubbard, and Petersen (1988) argue that access to internally generated cash can raise investment even among large firms because external funds are costly to raise. This interpretation of the relationship between cash flow and investment, however, is controversial (Kaplan and Zingales 1997).

3. On relationship lending, see also Berger and Udell 1998 and Cole 1998.

4. One limitation of this evidence is that data from the National Survey of Small Business Finance are cross-sectional. Thus, Petersen and Rajan cannot follow a given lending relationship over time. Rather, they can only compare firms with short relationships against other firms with longer banking relationships.

5. In fact, Cole, Goldberg, and White (1999) find that large banks are more reliant on observable firm characteristics in making lending decisions than small banks.

6. See Keeton 1996; Peek and Rosengren 1996, 1998; Strahan and Weston 1996, 1998; Craig and dos Santos 1998; Kolari and Zardkoohi 1997; Zardkoohi and Kolari 1997, 2001; Walraven 1997; Berger et al. 1998; Sapienza 2001; Berger, Demsetz, and Strahan 1999; Cole, Goldberg, and White 1999; Jayaratne and Wolken 1999.

7. As evidence, Strahan and Weston show that lending to small businesses increases after small banks are acquired. Berger et al. (1998) also find increases in small-business lending after small banks are acquired.

8. Regulatory initiatives allowed banks to branch across state lines in a limited fashion beginning at this time, and some large bank holding companies consolidated operations across state lines. Wide-scale consolidation of banking operations across state lines began in 1997, after the Interstate Banking and Branching Efficiency Act was passed.

9. State-level deregulation of restrictions on branching often occurred incrementally. In the typical case, state would first permit banks to branch into new markets in the state by buying existing branches held by other banks in the state. Then, after a lag of a few years, the state would permit banks to enter new markets by opening new branches. Kroszner and Strahan (1999) show that the political economy of branching deregulation can be best understood as a battle between interests within the banking industry and between banking and other interests in the financial industry.

10. During the early 1980s, the moral-hazard problem associated with deposit insurance was particularly acute because declines in bank charter value increased their incentive to engage in high-risk activities (Keeley 1990; Demsetz, Saidenberg, and Strahan 1996). In addition, moral-hazard problems were exacerbated by the explicit "too big to fail" doctrine articulated in 1984 in the wake of the failure of a large bank (O'Hara and Shaw 1990).

11. In fact, the 1988 Basle Accord explicitly places a higher required capital ratio on risky lending than on other kinds of bank activities such as mortgage lending. Some have argued that the tighter regulatory environment exacerbated the recession of 1990–91 (Bernanke and Lown 1991; Peek and Rosengren 1995).

12. Davis, Haltiwanger, and Schuh (1996) show that job creation in new, small plants varies procyclically with the business cycle, although less so than job creation in older plants. Evans and Leighton (1989) find that entrepreneurial activity is higher among better educated people. Bates (1990) reports that highly educated people are more likely to start businesses that survive.

13. Differences in tax rates, for example, may affect the incentive for entrepreneurial activity; including the state effects should eliminate much of this variation (Gentry and Hubbard 2000b). The state fixed effects helps control for systematic differences in the tax environment across states.

14. While the data are available back into the 1960s, our analysis begins in 1976 due to limits to other variables in the model.

15. Most large corporations are incorporated in Delaware, so our use of business incorporations everywhere except Delaware (and South Dakota) means that the results are generated by incorporations of small businesses.

16. We control for inflation in our regression equation with annual fixed effects.

17. Ideally we would focus on small-business lending; however, data limitations require us to use total business lending as a proxy.

18. These increases in growth, it should be noted, occur after controlling for state-level economic growth. Jayaratne and Strahan (1996) find that state-level growth accelerates after passage of laws allowing statewide branching.

19. For a discussion of instrumental variables models, see Greene 1993.

References

Avery, Robert B., Raphael W. Bostic, and Katherine A. Samolyk. 1998. The role of personal wealth in small business finance. *Journal of Banking and Finance* 22: 1019–1061.

Bates, Timothy. 1990. Entrepreneur human capital inputs and small business longevity. *Review of Economics and Statistics* 72: 551–559.

Berger, Allen N., Rebecca S. Demsetz, and Philip E. Strahan. 1999. The consolidation of the financial services industry: Cause, consequences and implications for the future. *Journal of Banking and Finance* 23: 135–194.

Berger, Allen N., and Gregory F. Udell. 1995. Relationship lending and lines of credit in small firm finance. *Journal of Business* 68: 351–382.

Berger, Allen N., and Gregory F. Udell. 1996. Universal banking and the future of small business lending. In *Financial System Design*, ed. A. Saunders and I. Walter. Irwin.

Berger, Allen N., and Gregory F. Udell. 1998. The economics of small business finance: The roles of private equity and debt markets in the financial growth cycle. *Journal of Banking and Finance* 22: 613–673.

Berger, Allen N., Anthony A. Saunders, Joseph K. Scalise, and Gregory F. Udell. 1998. The effects of bank mergers and acquisitions on small business lending. *Journal of Financial Economics* 50: 187–229.

Bernanke, Benjamin S., and Cara S. Lown. 1991. The credit crunch. *Brookings Papers on Economic Activity* 2: 205–248.

Black, Sandra E., and Philip E. Strahan. 2002. Entrepreneurship and bank credit availability. *Journal of Finance* 57 (6) (December): 2807–2833.

Bonaccorsi di Patti, Emilia, and Giovanni Dell'Ariccia. 2001. Bank Competition and Firm Creation. IMF working paper 01/21.

Boot, Arnoud W. A., and Anjan V. Thakor. 2000. Can relationship banking survive competition? *Journal of Finance* 55 (2) (April): 679–713.

Cetorelli, Nicola, and Michele Gambera. Forthcoming. Bank structure, financial dependence and growth: International evidence from industrial data. *Journal of Finance* 56: 617–648.

Cole, Rebel A. 1998. The importance of relationships to the availability of credit. *Journal of Banking and Finance* 22: 959–977.

Cole, Rebel A., and John D. Wolken. 1995. Financial services used by small business: Evidence from the 1993 National Survey of Small Business Finances. *Federal Reserve Bulletin* 81 (7) (July): 629–667.

Cole, Rebel A., Lawrence G. Goldberg, and Lawrence J. White. 1999. Cookie-cutter versus character: The micro structure of small business lending by large and small banks. In Proceedings of Conference on Business Access to Capital and Credit, Federal Reserve Bank of Chicago.

Craig, Ben R., and João A. C. dos Santos. 1998. Study of the banking consolidation impact on small business lending. *Proceedings of the Federal Reserve Bank of Chicago* (May): 569–588.

Davis, Steven J., Haltiwanger, John C., and Scott Schuh. 1996. *Job Creation and Destruction*. MIT Press.

Demsetz, Rebecca, S., Marc R. Saidenberg, and Philip E. Strahan. 1996. Banks with something to lose: The disciplinary role of franchise value. *Federal Reserve Bank of New York Economic Policy Review* 2 (2) (October): 1–14.

Demsetz, Rebecca S., and Philip E. Strahan. 1995. Historical Patterns and Recent Changes in the Relationship between Bank Size and Risk. *Federal Reserve Bank of New York Economic Policy Review* 1 (2) (July): 13–26.

Diamond, Douglas. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51: 393–414.

Evans, David S., and Boyan Jovanovic. 1989. An estimated model of entrepreneurial choice under liquidity constraints. *Journal of Political Economy* 97: 808–827.

Evans, David S., and Linda S. Leighton. 1989. Some empirical aspects of entrepreneurship. *American Economic Review* 79: 519–535.

Fairlie, Robert W. 1999. The absence of the African-American owned business: An analysis of the dynamics of self-employment. *Journal of Labor Economics* 17 (1) (January): 80–108.

Fazzari, Steven M., R. Glenn Hubbard, and Bruce Petersen. 1988. Financing constraints and corporate investment. *Brookings Papers on Economic Activity* 1: 141–195.

Fisman, Raymond, and Mayank Raturi. 2000. Does competition encourage cooperation? Evidence from trade credit relationships. Unpublished manuscript.

Gentry, William, and R. Glenn Hubbard. 2000a. Entrepreneurship and Household Savings. NBER working paper 7894.

Gentry, William, and R. Glenn Hubbard. 2000b. Tax policy and entrepreneurial entry. *American Economic Review* 90: 283–287.

Greene, William H. 1997. *Econometric Analysis*. Prentice-Hall.

Hausman, Jerry. 1978. Specification tests in econometrics. *Econometrica* 46: 1251–1271.

Holtz-Eakin, Douglas, David Joulfaian, and Harvey S. Rosen. 1994a. Entrepreneurial decisions and liquidity constraints. *RAND Journal of Economics* 23: 334–347.

Holtz-Eakin, Douglas, David Joulfaian, and Harvey S. Rosen. 1994b. Sticking it out: Entrepreneurial survival and liquidity constraints. *Journal of Political Economy* 102: 53–75.

Holtz-Eakin, Douglas, and Harvey S. Rosen. 1999. Cash Constraints and Business Start-Ups: Deutschmarks versus Dollars. Working paper 62, Center for Economic Policy Studies, Princeton University.

Huck, Paul, Sherrie L. W. Rhine, Philip Bond, and Robert P. Townsend. 1999. A comparison of small business finance in two chicago minority neighborhoods. In Proceedings of Conference on Business Access to Capital and Credit, Chicago.

Jayaratne, Jith, and Philip E. Strahan. 1996. The finance-growth nexus: Evidence from bank branch deregulation. *Quarterly Journal of Economics* 101: 639–670.

Jayaratne, Jith, and Philip E. Strahan. 1998. Entry restrictions, industry evolution and dynamic efficiency: Evidence from commercial banking. *Journal of Law and Economcis* 41: 239–274.

Jayaratne, Jith, and John D. Wolken. 1999. How important are small banks to small business lending? New evidence from a survey of small firms. *Journal of Banking and Finance* 23: 427–458.

Kaplan, Steven, and Luigi Zingales. 1997. Do investment-cash flow sensitivities provide useful measures of financing constraints? *Quarterly Journal of Economics* 112: 169–215.

Keeley, Michael. 1990. Deposit insurance, risk and market power in banking. *American Economic Review* 80: 1183–1200.

Keeton, William R. 1996. Do bank mergers reduce lending to businesses and farmers? New evidence from tenth district states. *Economic Review* (Federal Reserve Bank of Kansas City) 81: 63–75.

Kolari, J., and A. Zardkoohi. 1997. The Impact of Structural Change in the Banking Industry on Small Business Lending. Report to U.S. Small Business Administration.

Kroszner, Randall S., and Philip E. Strahan. 1999. What drives deregulation? Economics and politics of the relaxation of bank branching restrictions. *Quarterly Journal of Economics* 114: 1437–1467.

O'Hara, Maureen, and Wayne Shaw. 1990. Deposit insurance and wealth effects: The value of being "too big to fail." *Journal of Finance* 45: 1587–1600.

Peek, Joe, and Eric S. Rosengren. 1995. Bank regulation and the credit crunch. *Journal of Banking and Finance* 19: 679–692.

Peek, Joe, and Eric S. Rosengren. 1996. Small business credit availability: How important is the size of the lender?" In *Universal Banking*, ed. A. Saunders and I. Walter. Irwin.

Peek, Joe, and Eric S. Rosengren. 1998. Bank consolidation and small business lending: It's not just bank size that matters. *Journal of Banking and Finance* 22: 799–819.

Petersen, Mitchell A., and Raghuram G. Rajan. 1994. The benefits of lending relationships: Evidence from small business data. *Journal of Finance* 49: 3–37.

Petersen, Mitchell A., and Raghuram G. Rajan. 1995. The effect of credit market competition on lending relationships. *Quarterly Journal of Economics* 110: 407–443.

Petersen, Mitchell A., and Raghuram G. Rajan. 2002. Does distance still matter? The information revolution in small business lending. *Journal of Finance* 57: 2533–2570.

Sapienza, Paola. 2002. The effects of banking mergers on loan contracts. *Journal of Finance* 57: 329–367.

Spong, Kenneth. 2000. Banking Deregulation: Its Purposes, Implementation, and Effects, fifth edition. Division of Supervision and Risk Management, Federal Reserve Bank of Kansas City.

Strahan, Philip E., and James P. Weston. 1996. Small business lending and bank consolidation: Is there cause for concern? *Current Issues in Economics and Finance* (Federal Reserve Bank of New York) 2 (3).

Strahan, Philip E., and James P. Weston. 1998. Small business lending and the changing structure of the banking industry. *Journal of Banking and Finance* 22: 821–845.

Sylla, Richard, John Legler, and John Wallis. 1987. Banks and state public finance in the new republic: The United States, 1790–1860. *Journal of Economic History* 48: 391–403.

Uzzi, Brian. 1999. Embeddedness in the marking of financial capital: How social relations and networks benefit firms seeking financing. *American Sociological Review* 64 (4) (August): 481–505.

Zardkoohi, A., and J. Kolari. 2001. The effect of bank mergers and acquisitions on the credit decision process in small business lending. *Finance India* 15 (March): 119–141.

4

Public Policy and Innovation in the U.S. Pharmaceutical Industry

Frank R. Lichtenberg

The pharmaceutical industry is one of the most R&D-intensive industries in the economy. According to the National Science Foundation, in 1997 company-funded R&D expenditure as a percentage of the net sales of R&D performing companies was 10.5 percent in the industry, more than three times as high as it was in manufacturing as a whole. Moreover, R&D intensity is increasing much more rapidly in pharmaceuticals than it is in other industries. (See figure 1.)

In several respects, the government plays a larger role as both customer and regulator of the pharmaceutical industry than it does of most other industries. As figure 2 indicates, since 1960 the share of industry output purchased by the public sector has increased steadily. In 1998, more than one-fifth of the industry's output was paid for by the government, primarily under the Medicaid program.[1] The prices paid by the government are regulated under the Medicaid Drug Rebate Program, which requires a drug manufacturer to enter into and have in effect a national rebate agreement with the Secretary of the Department of Health and Human Services for states to receive federal funding for outpatient drugs dispensed to Medicaid patients.[2]

But perhaps what most distinguishes the pharmaceutical industry from other industries is the extent of the government's direct control over innovation. According to section 505 of the Federal Food, Drug, and Cosmetic Act, "No person shall introduce or deliver for introduction into interstate commerce any new drug, unless an approval of an application ... is effective with respect to such drug.... Such person shall submit to the Secretary as a part of the application ... full reports of investigations which have been made to show whether or not such drug is safe for use and whether such drug is effective in use."[3]

For these reasons, the pharmaceutical industry appears to be a good industry in which to study the interaction between public policy and

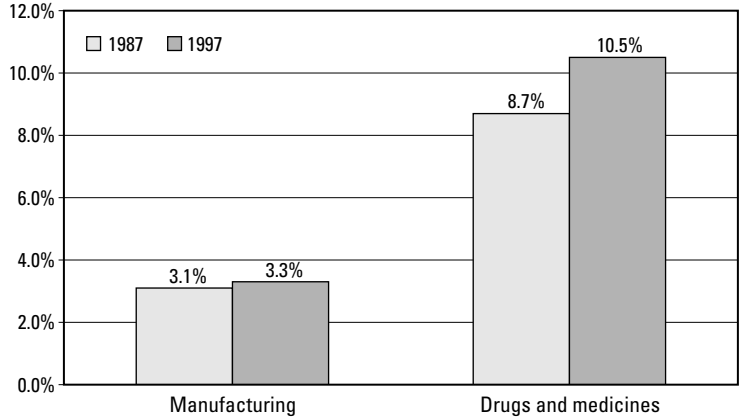


Figure 1
Company-funded R&D intensity, drugs vs. total manufacturing. Source: National Science Foundation.

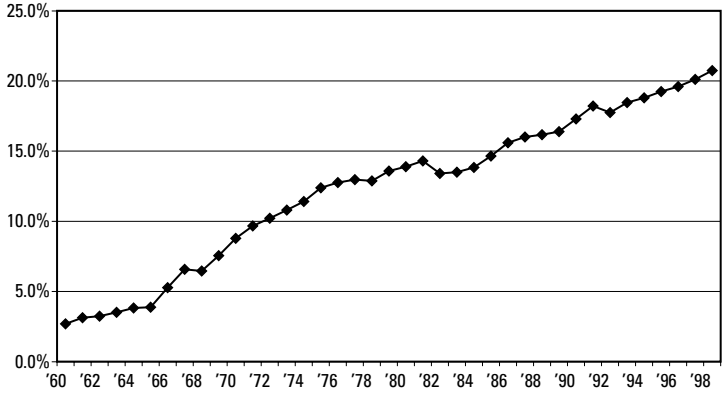


Figure 2
Public pharmaceutical expenditure as percentage of total pharmaceutical expenditure.

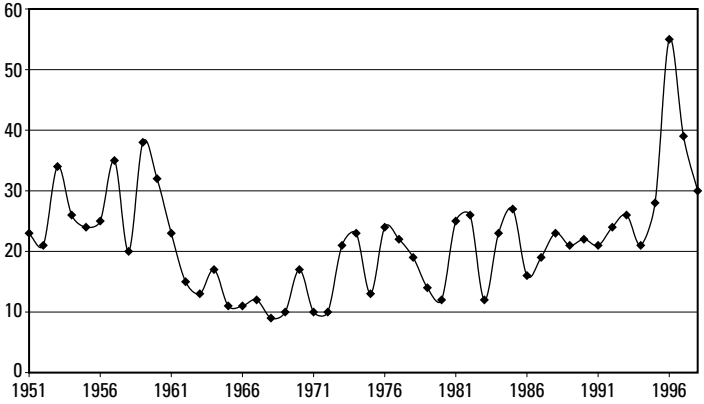


Figure 3
Number of new molecular entities approved by FDA, 1951–1998.

innovation. In this chapter, I will consider the effects of five public policies on innovation in the industry.

The 1962 Kefauver-Harris Amendment and the 1992 Prescription Drug User Fee Act

As noted above, introduction of a new drug requires approval by the FDA. New molecular entities are the most important new drugs.[4] Figure 3 presents annual data for the period 1950–1998 on the number of new molecular entities (NMEs) approved by the FDA. Between 1950 and 1961, the average annual number of NMEs approved was 27, and the minimum number was 20. Between 1962 and 1972, the average annual number of NMEs approved was 12, and the *maximum* number was 17. This precipitous drop in the number of NMEs approved was due to the passage of the 1962 Kefauver-Harris amendment. Congress passed this amendment, which required extensive animal pharmacological and toxicological testing before a drug could be tested in humans, in response to the thalidomide tragedy.[5] The data from these studies had to be submitted in the form of an IND ("Notice of Claimed Investigational Exemption for a New Drug") and approved by the FDA before clinical studies could begin. The amendment also required that manufacturers submit to the FDA "substantial evidence" of the unapproved (investigational) drug's efficacy, as well as safety, in the form of an NDA ("New Drug Application"). Therefore, in addition to

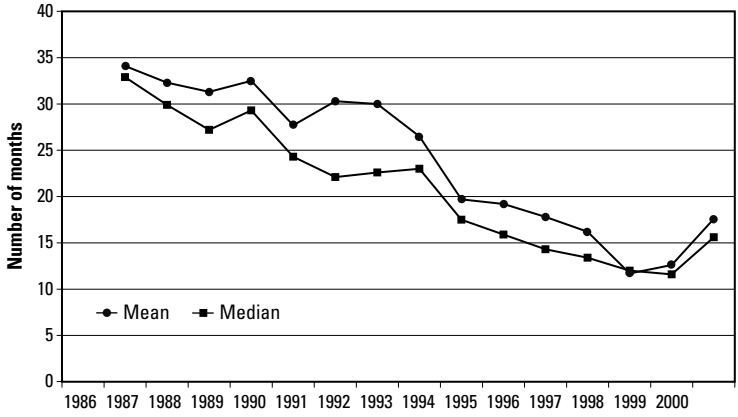


Figure 4
FDA approval times for new molecular entities.

safety, the manufacturer was now required to demonstrate efficacy (effectiveness), as well.[6] Passage of this amendment appears to have led to a significant, roughly 10-year decline in the number of new drugs approved.

Another notable feature of figure 3 is the huge, albeit transitory, increase in the number of NMEs approved in the mid 1990s, especially in 1995 and 1996. Much of this increase can be attributed to the decline in mean and median FDA approval times depicted in figure 4. Mean approval time decreased more than a third between 1992 and 1994, from 30 months to less than 20 months. This reduction can be traced to the passage of the 1992 Prescription Drug User Fee Act, which increased the budget for, and accelerated, the drug approval process.

Medicare and Medicaid

Titles XVIII and XIX (Medicare and Medicaid) of the 1965 Social Security Amendments have undoubtedly also had important effects on pharmaceutical innovation, via their effect on the demand for prescription drugs, although these effects are harder to identify. Prescription drugs are covered under Medicaid. In 1964, less than 4 percent of national expenditure on prescription drugs was publicly funded. In 1998, over 20 percent was publicly funded, and the Medicaid program accounted for over 80 percent of this funding.

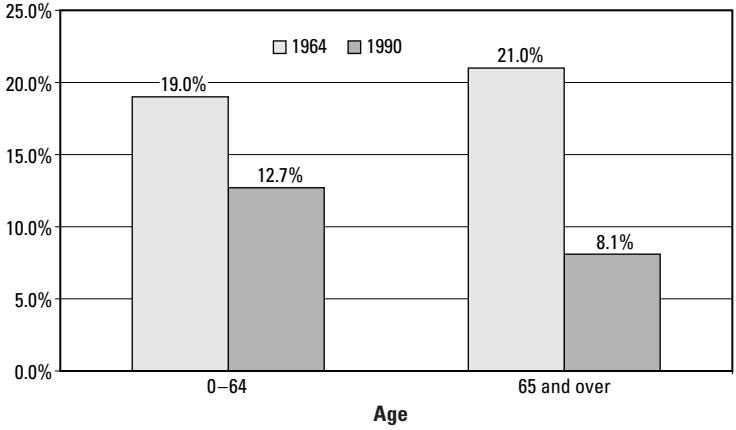


Figure 5
Probability of no physician contact within last 2 years, by age, 1964 vs. 1990.

Although Medicare has not, until the present, paid for most prescription drugs, Medicare Part B (Supplementary Medical insurance) pays part of the cost of a service that is complementary with (necessary to receive) prescription drugs: doctor visits.[7] The data shown in figure 5 suggest that Medicare had a significant effect on utilization of ambulatory care by the elderly. Between 1964—immediately before Medicare was established—and 1990, the probability that a person over 65 had not seen a doctor in the last 2 years declined from 21.0 percent to 8.0 percent. The corresponding probability for people under 65 (who are generally not covered by Medicare) also declined, but by much less.

In a recent paper (Lichtenberg 2000), I present evidence that increases in real per capita U.S. health expenditure during the period 1960–1997—including the increases attributable to Medicare and Medicaid—contributed to the increase in longevity during that period.[8] As figure 6 indicates, mean pharmaceutical consumption increases sharply with age, especially after age 55. By significantly expanding the size of the elderly population, Medicare and Medicaid significantly increased the demand for pharmaceuticals. In 1996, people over 65 accounted for about 11 percent of the population but for about a third of aggregate drug expenditure.

If enactment of Medicare caused an increase in the demand for drugs by the elderly, both absolutely and relative to the demand for drugs by

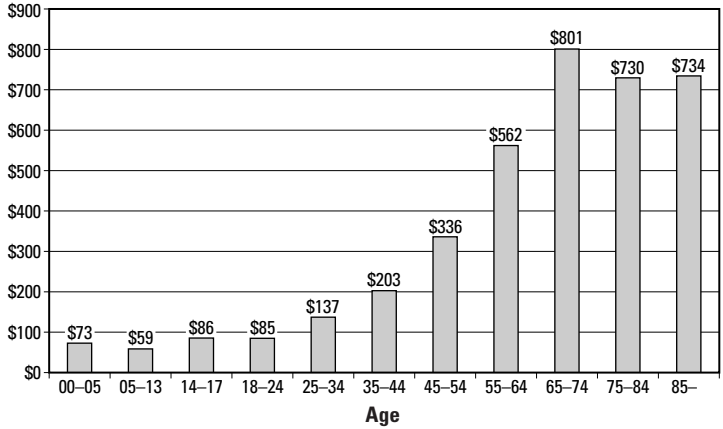


Figure 6
Mean 1996 expenditures on prescription drugs, by age.

younger people, then one would expect the ratio of drugs developed for illnesses experienced by the elderly to all drugs developed to be higher in the post-Medicare era than it was before Medicare. According to the Pharmaceutical Research and Manufacturers Association (PhRMA), "it takes an average of 12–15 years to discover and develop a new medicine."[9] If the orientation of drug development shifted toward the elderly after 1965, one would expect the drugs approved by the FDA after about 1980 (whose development presumably began in the mid to late 1960s) to be more targeted to the elderly than the drugs approved before 1980. As figure 7 (which is based on data from the 1996 Medical Expenditure Panel Survey) shows, there is evidence supporting this hypothesis. In 1996, 37 percent of the prescriptions for drugs approved after 1980 were consumed by the elderly, whereas the elderly consumed only 27 percent of the prescriptions for drugs approved before 1981. Public policy may have shifted both the rate and direction of private innovation.

The Hatch-Waxman Act

In passing the 1984 Hatch-Waxman Act, Congress attempted to balance the interests of the generic drug industry against those of manufacturers of innovator drugs. That act contained two sets of changes. First, it eliminated the duplicative testing requirements necessary to

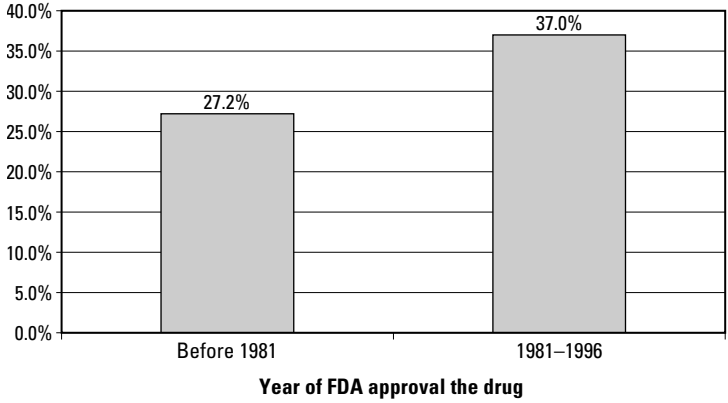


Figure 7
Percentage of 1996 prescriptions consumed by people 65 and over, by year of FDA approval.

obtain approval for a generic copy of a previously approved innovator drug. Specifically, it created an abbreviated approval process for generic copies of innovator drugs.[10] It allowed manufacturers of generic drugs to file an abbreviated new drug application and conduct clinical tests demonstrating bioequivalence with a brand-name drug before that drug's patent expires. As a result, the FDA can approve many of those applications immediately after patent expiration. That provision overturned a 1984 decision by the Court of Appeals for the Federal Circuit that clinical tests conducted by generic manufacturers before patent expiration constitute patent infringement. It also established a process to handle patent disputes between generic manufacturers and innovator firms. Those provisions helped to increase the availability of generic drugs after patent expiration.

Second, the act established patent-term extensions for innovator drugs. Because such drugs receive patents from the Patent and Trademark Office before they receive approval from the FDA, part of their time under patent is spent in the clinical trials necessary for FDA approval. The patent extensions were intended to offset part of the patent term used up during the approval process. Under the new procedures, manufacturers of a newly approved innovator drug that contains an active ingredient never before approved by the FDA can apply for a patent-term extension that equals the sum of all the time spent in the NDA review process plus half of the time spent in the clinical

testing phase. Two limitations exist. A patent-term extension cannot exceed 5 years, nor can it allow the period between product approval and patent expiration to exceed 14 years. The average length of patent-term extensions granted under this provision is 3 years.[11]

By extending patents on brand-name drugs while making it easier for generic drugs to enter the market after patents expire, the Hatch-Waxman Act aimed to benefit consumers by increasing the supply of generic drugs while preserving drug companies' incentive to invest in research and development.

A report issued by the Congressional Budget Office (CBO) in 1998 concluded that "the point in the life of an average drug at which generic entry occurs did not change much under the act, because the average length of a patent extension roughly offsets the average delay between patent expiration and generic entry that existed before 1984" (U. S. Congress, CBO, chapter 4, conclusions). According to the CBO, the two Hatch-Waxman Act changes that allowed generic manufacturers to obtain FDA approval more quickly once the patent on an innovator drug has expired shortened the average time between patent expiration and generic entry for top-selling drugs from 3 or 4 years to less than 3 months. This roughly offsets the average 2.8 year delay in generic entry provided by the patent-term extensions and exclusivity provisions in the Hatch-Waxman Act (chapter 4, p. 7).

I have performed an analysis of data on all new molecular entities approved by the FDA since 1940. My analysis indicates that imitation lags have been considerably shorter since 1984 than they were before 1984.

I obtained from the FDA (by filing a Freedom of Information Act request) lists of all 4,370 New Drug Applications (NDAs) and 6,024 Abbreviated New Drug Applications (ANDAs) approved by the FDA from 15 September 1938 to 28 January 1997. Both lists included the application number, applicant name, approval date, a list of up to 13 generic ingredients (e.g., butabarbital sodium 15 mg), the dosage form and route of administration of the drug, and whether the application was for a prescription or over-the-counter (OTC) drug. The NDA list also indicated the application type[12] and the review status (priority or standard) of the application. I concatenated these two lists and then sorted the file of 10,394 applications by the first three generic ingredients,[13] dosage, firm, route of administration, and application date.

For each drug (where a drug is defined by the first three generic ingredients, dosage form, and route of administration), I identified the first and (if present) second dates on which the FDA approved an NDA or ANDA for that drug. I selected only those drugs that the FDA designated as new molecular entities on the first approval date. There were 1,277 drugs in this sample. Approximately 25 percent of these drugs were imitated by at least one other firm before the period covered by the sample ended on January 28, 1997. For examples of drugs that were imitated and not imitated by the end of the sample period, see table 1.

My objective is to obtain good estimates of the imitation-probability profile—the relationship between the probability of imitation and the time elapsed since innovation—and to test whether this profile has shifted over time. For the majority of innovations that had not been imitated by the end of the sample period, the time until imitation is unknown. Such data cannot be analyzed by ignoring the censored observations since, among other considerations, the longer-lived observations are generally more likely to be censored. I computed nonparametric estimates of the "survival" distribution using both the censored and noncensored observations, where "survival" means survival as the exclusive producer of the drug.[14] (The probability of survival is the probability of not being imitated.) To test whether the imitation-probability profile has shifted over time, I stratified the sample into two groups: drugs first approved during the period 1940–1968 ($N = 320$), and drugs first approved during the period 1969–1996 ($N = 957$).

Estimates of the imitation-probability profile, by period of innovation, are reported in table 2 and graphed in figure 8. There has been a very sharp upward shift in the profile, particularly in the first 15 years of the new drug. As shown in rows 2 and 3 of table 2, the probability that a new molecular entity introduced during the period 1940–1968 was imitated was 2 percent after 10 years and 7 percent after 15 years; for new molecular entities introduced during the period 1969–1996, these probabilities were 17 percent and 33 percent, respectively. The hypothesis of equality of the two periods' imitation-probability profiles is decisively rejected.

Table 2 shows the "cumulative" probability of being imitated between year 0 and year t ($t = 5, 10, 15, \ldots$). The *annual* hazard rates, or probabilities of imitation between years t and $t + 1$, conditional on no previous imitation, are given at the top of p. 94.

Table 1
Examples of new molecular entities imitated and not imitated.

	Innovation date	Innovator	Imitation date	Imitator
Imitated before 1 Jan 1997				
Meperidine Hydrochloride 50 mg	10 Nov 42	Sanofi Winthrop	31 Jul 73	Wyeth Ayerst Labs
Tubocurarine Chloride 3 mg/ml	20 Feb 45	Apothecon	3 Jan 47	Abbott Labs
Procainamide Hydrochloride 500 mg	13 Jun 50	Apothecon	7 Jun 77	Danbury Pharma
Sulfasalazine 500 mg	20 Jun 50	Pharmacia, Upjohn	12 Nov 73	Lederle Labs
Trichlormethiazide 4 mg	9 Mar 60	Schering	16 May 77	Lannett
Sulfinpyrazone 100 mg	6 May 60	Ciba	26 May 82	Danbury Pharma
Flurazepam Hydrochloride 30 mg	7 Apr 70	Roche Prods	27 Nov 85	Mylan
Levodopa 250 mg	4 Jun 70	Roche	6 Jul 72	Roberts Labs
Clorazepate Dipotassium 7.5 mg	10 Mar 80	Abbott Labs	26 Jun 87	Able Labs Inc
Trifluridine 1%	10 Apr 80	Glaxo Wellcome	6 Oct 95	Steris Labs
Atenolol 25 mg	9 Apr 90	Zeneca Pharms Grp	28 Apr 92	Lederle Labs
Miconazole Nitrate 2%	15 Feb 91	Johnson Rw	30 Oct 92	Copley Pharm
Not imitated before 27 Jan 97				
Estrogens, Conjugated 1.25 mg	8 May 42	Wyeth Ayerst Labs		
Diphenhydramine Hydrochloride 50 mg	4 Mar 46	Parke Davis Div Wl		
Methantheline Bromide 50 mg	8 May 50	Roberts Labs		
Corticotropin 25 Units/Vial	3 Jul 50	Rhone Poulenc Rorer		
Triethanolamine Polypeptide Oleate Condensate 10%	16 Feb 60	Purdue Frederick		
Warfarin Sodium 7.5 mg	16 Feb 60	Dupont Merck		
Flavoxate Hydrochloride 100 mg	15 Jan 70	Smithkline Beecham		
Ketamine Hydrochloride Eq 10 mg base/ml	19 Feb 70	Parke Davis Div Wl		
Metronidazole 500 mg	27 Feb 80	Searle Pharms		
Meclocycline Sulfosalicylate 1%	30 May 80	J And J Cons Prods		
Fluconazole 50 mg	29 Jan 90	Pfizer Central Res		
Nafarelin Acetate Eq 0.2 mg base/inhalant	13 Feb 90	Searle		

Table 2
Estimates of imitation probabilities (standard errors in parentheses).

Years since innovation	Period of innovation	
	1940–1968	1969–1997
0	0 0	0 0
5	0.01 (0.005)	0.06 (0.008)
10	0.02 (0.007)	0.17 (0.015)
15	0.07 (0.014)	0.33 (0.022)
20	0.19 (0.022)	0.36 (0.024)
25	0.29 (0.025)	0.40 (0.028)

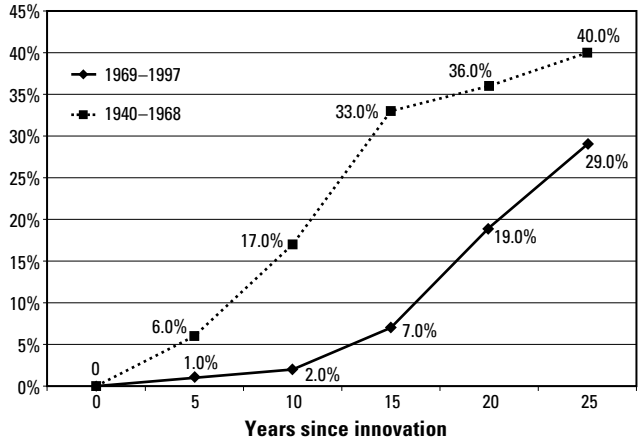


Figure 8
Probability of being imitated, by period of innovation.

Years since innovation	Period of innovation	
	1940–1968	1969–1997
0–5	0.0019	0.0120
5–10	0.0013	0.0245
10–15	0.0104	0.0426
15–20	0.0279	0.0110
20–25	0.0271	0.0132
25–30	0.0151	

While these estimates reveal a significant shift in the imitation-probability profile, they do not establish the precise timing of the shift. We attempted to do that by computing the fraction of NMEs that were imitated within 10 years, by year of innovation, grouped in 5-year intervals. These data are presented in table 3, which shows, for example, that 21 percent of the drugs introduced during the period 1975–1979 were imitated within 10 years. The table indicates that the drugs that were introduced during the period 1970–1974 were the first to experience the sharp increase in imitation probability. About a third of these were introduced in 1974, 10 years before Hatch-Waxman.

I used a different data set—the FDA's electronic Orange Book of approved drug products—to provide evidence about the effect of the Hatch-Waxman Act on the *average number of producers* of given drugs

Table 3
Percentage of innovations imitated within 10 years, by year of innovation.

Year of innovation	Number of innovations	% of innovations imitated within 10 years
1940–1944	4	0
1945–1949	21	5
1950–1954	66	3
1955–1959	81	2
1960–1964	97	2
1965–1969	69	1
1970–1974	97	15
1975–1979	143	21
1980–1984	170	19
1985–1989	180	24

approved before 1982. For each of these 1,009 drugs, I calculated the number of firms that the FDA had approved to market the drug by the end of year t ($t = 1982, 1983, \ldots, 2000$). I then estimated the regression

$$\ln(\text{N\_FIRM}_{it}) = \alpha_i + \delta_t + u_{it},$$

where N_FIRM_{it} = the number of firms approved to market drug i by the end of year t, α_i is a fixed effect for drug i, δ_t is a fixed effect for year t, and u_{it} is a disturbance. The difference $(\delta_t - \delta_{t-1})$ may be interpreted as the percentage change from the previous year in the mean number of firms approved to market a given drug. The estimates of these differences are shown in figure 9.[15] As one would expect, the largest increases in the average number of producers occurred soon after the Hatch-Waxman Act was passed. The mean number of producers increased at an average rate of 5.6 percent per year during the period 1985–1988. This is more than double the average annual rate at which the mean number of producers increased during the entire period 1982–2000 (2.6 percent).

The CBO report reached two apparently inconsistent conclusions: the Hatch-Waxman Act had no effect on effective drug life but reduced the expected net present value (NPV) of launching a new drug by about 12 percent. My findings of faster and more frequent generic entry after the act can account for this reduction in the NPV of innovation.

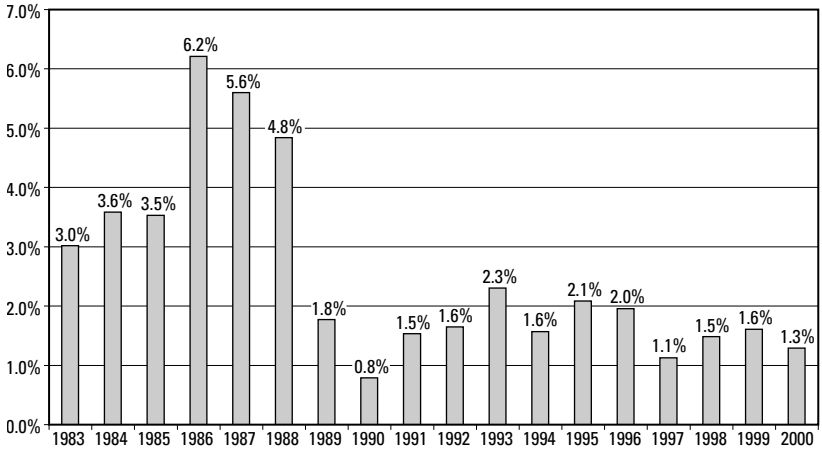


Figure 9
Percent increase from previous year in mean number of firms approved to market drugs.

Does an Increase in the Speed or Ease of Imitation Reduce the Rate of Innovation?

Some (very simple) theoretical models, such as the Cournot duopoly model of an innovator and an imitator presented in the appendix, imply that an increase in the speed or ease of imitation will reduce the rate of innovation. That model implies that the more benefit an imitator gets from a dollar of the innovator's R&D (holding constant the innovator's benefit), the fewer R&D dollars will be spent. Although an increase in the spillover rate unambiguously reduces the innovation rate, it may either increase or decrease social welfare (the sum of consumer surplus and innovator and imitator profits).[16] Also, it is possible for an increase in the spillover rate to reduce the imitator's profits. If theft (of intellectual or physical assets) becomes too easy, there is little incentive to accumulate wealth, and there is not much to steal. "A viable parasite does not kill his host" is a well-known maxim in sociobiology.

However, other (much more complicated) theoretical models imply that the effect of a policy-induced fall in the private cost of imitation on the steady-state innovation rate may be positive, zero, or negative. Grossman and Helpman (1991, chapters 11 and 12) explored the nature of the relationship between innovation and imitation in a context in which both of these activities are risky and both result from the investment decisions of farsighted entrepreneurs. Specifically, they investigated interdependencies between the learning processes in the industrialized North and the developing South. They assumed that most learning in the South takes the form of imitation of technologies previously developed in the North, rather than of invention of entirely new products and processes. Imitation gives rise to product-cycle trade, as goods initially are invented and produced in the North, and then copied and exported by the South.[17]

Grossman and Helpman considered the long-run effects of policies that governments might use to encourage local accumulation of knowledge. The government in the South might relax patent protection laws as applied to foreign intellectual property. This would encourage imitation by reducing the cost to a Southern entrepreneur of inventing around existing patents. They analyzed the effect of a policy-induced fall in the private cost of imitation on the steady-state innovation rate in a variety of settings. First, they assumed that innovation in a particular product line ceases at the moment that a new good is introduced.

They showed that if the world economy begins in a "wide-gap equilibrium," then a policy-induced fall in private cost of imitation causes an *increase* in both the rate of imitation and the rate of innovation. Entrepreneurs in the North expand their research efforts because the more rapid pace of copying implies greater expected profits for the typical new variety. If the world economy begins in a "narrow-gap equilibrium," the fall in private imitation costs has no lasting effect on research activities in either country.

As Grossman and Helpman observed, the implication that "an increase in the rate of imitation ... strengthens the incentive to innovate ... is a strong result. But it is one that relies heavily on the particulars of the specification" (p. 306).

Grossman and Helpman then assumed that innovation in a particular product line continues when a new good is introduced. Not only must entrepreneurs in the North look ahead to their ultimate displacement from the market by imitators in the South; so too must Southern firms foresee their own eventual demise in the wake of further technological advances in the North. They showed that when followers are "inefficient," then imitation subsidies increase the rate of innovation: there is a positive feedback between the processes of innovation and imitation. However, when followers are efficient, then imitation subsidies reduce the rate of innovation. Grossman and Helpman found that the effect of a policy-induced fall in the private cost of imitation on the steady-state innovation rate may be positive, zero, or negative.

To summarize, while some theoretical models imply that an increase in the speed or ease of imitation will reduce the rate of innovation, other models suggest that this is not necessarily the case. Perhaps this issue can only be resolved empirically—but good empirical evidence may be hard to find!

President Clinton's 1993 Proposals for Reforming Health Care

Economic theory suggests that government policy affects the incentives of private agents to undertake R&D investment. Indeed, Douglass North has argued that sustained world economic growth (the Industrial Revolution) did not begin until a few centuries ago because only then did governments establish and enforce intellectual property rights.[18]

Although one could try to examine the effect of policy on pharmaceutical innovation directly, we believe that a more fruitful approach

is to analyze the relationships between each of these variables and a third, "intervening" variable: the market value of pharmaceutical firms. The relationships among these three variables may be represented schematically as follows:

(1)	(2)	
Government policy →	Market value of pharmaceutical firms	→ R&D investment.

The first relationship can be estimated using high frequency (daily) data. The second relationship can be estimated using low frequency (annual) data (from Compustat and CRSP). Though estimated at different frequencies, parameter estimates from the two models can be combined to determine the effect of policy changes on R&D. Examination of these relationships within a unified framework should enable us to make precise inferences about the effects of changes in the environment on innovation in the pharmaceutical industry.[19]

Public Policy → Market Value of Pharmaceutical Firms

We hypothesize that some government policy events affect the expected future net cash flows of pharmaceutical firms. If the stock market is efficient, the value of the firm at time t is the expected present discounted value (PDV) of its future net cash flows, conditional on the information available at time t. Hence policy events affect the value of the firm.

There has been some debate about the relationship between market value and the expected PDV of future net cash flows (or between *changes* in market value and *changes* in the expected PDV of future net cash flows), and about "market efficiency" in general.

Several recent empirical studies have found a strong relationship between market value and subsequent cash flows (actual or forecast).

Cornett and Tehranian (1992) examined the performance of bank mergers using remerger and postmerger accounting data for 15 large interstate and 15 large intrastate bank acquisitions completed between 1982 and 1987. They found "significant correlations ... between stock market announcement-period abnormal returns and the cash flow and accounting performance measures. The results suggest that, for large bank mergers, expectations of improved bank performance underlie the equity revaluations of the merging banks."

Healy, Palepu, and Ruback (1992) analyzed accounting and stock return data for a sample of the largest 50 US mergers between 1979 and mid-1984 to examine postacquisition performance, as well as the sources of merger-induced changes in cash flow performance. They reported that a "strong positive relationship exists between postmerger increases in operating cash flows and abnormal stock returns at merger announcements. This indicates that expectations of economic improvements underlie the equity revaluations of the merging firms."

Kaplan and Ruback (1995) compared the market value of highly leveraged transactions (HLT) to the discounted value of their corresponding cash flow forecasts. For a sample of 51 HLTs completed over the period 1983–1989, the valuation of discounted cash flow forecasts are within 10 percent, on average, of the market values of the completed transactions. These valuations perform at least as well as valuation methods using comparable companies and transactions.

Malkiel (1990, pp. 187–188) reviews "all the recent research proclaiming the demise of the efficient-market theory"[20] and concludes that "while the stock market may not be perfect in its assimilation of knowledge, it does seem to do quite a creditable job." Malkiel also notes that "one has to be impressed with the substantial volume of evidence suggesting that stock prices display a remarkable degree of efficiency. Information contained in past prices or any publicly available fundamental information is rapidly assimilated into market prices."

Conroy et al. (1992) provided evidence that four major legislative and regulatory initiatives directed toward the pharmaceutical industry during the 1970s and the 1980s—the Maximum Allowable Cost Program (1975), the Prospective Payment Plan (1982–83), the Drug Price Competition and Patent Term Restoration Act (1984), and the Catastrophic Protection Act (1987–88)—had significant, often negative, effects on share returns (market value).

Market Value of Firms → R&D Investment

The second relationship to be examined is the effect of a firm's market value on its rate of R&D investment. John Maynard Keynes (1936) may have been the first economist to hypothesize that the incentive to invest depends on the market value of capital relative to its replacement cost. James Tobin (1969) provided a rigorous theoretical foundation for this

hypothesis. Below we show that under certain plausible assumptions, a value-maximizing firm's investment (relative to its capital stock) is a (linear) function of Tobin's q—the ratio of the market value of the firm to the replacement cost of its capital.[21]

In each period t, firm i's real net cash flow X is given by

$$X_{it} = \mathrm{F}(K_{i,t-1}, N_{it}) - w_t N_{it} - p_t I_{it} + \mathrm{C}(I_{it}, K_{i,t-1}),$$

where K is the capital stock, F( ) is the real revenue (production) function of the firm, N is employment, w is the wage rate, p is the real price of investment goods, and C( ) is the function determining the cost of adjusting the capital stock. The marginal cost of newly installed capital is therefore $p_t + \mathrm{C}_I(I_{it}, K_{i,t-1})$.

Under the assumption of value maximization, the firm maximizes the present value of its future net cash flows. Letting β_{is} be the discount factor appropriate for the ith firm at time s, the firm's value at time t is

$$V_{it} = \max \mathrm{E}_{it} \sum_{s=t}^{\infty} \left(\prod_{j=t}^{s} \right) \beta_{is} X_{is},$$

where E_{it} is the expectations operator for firm i conditional on information available at time t. The firm chooses the past of investment and employment, given the initial capital stock, to maximize firm value. The change in the capital stock—net investment—is given by $I_{it} - \text{‰} K_{i,t-1}$, where ‰ is the (assumed constant) proportional rate of depreciation.

For investment, the solution to the problem requires that the marginal value of an additional unit of investment (denoted by q_{it}) equal its marginal cost:

$$q_{it} = p_t + \mathrm{C}_I(I_{it}, K_{i,t-1}). \tag{1}$$

Assume that the adjustment cost function is quadratic,

$$\mathrm{C}(I_{it}, K_{i,t-1}) = (\omega/2)[(I_{it}/K_{i,t-1}) - m_i]^2 K_{i,t-1},$$

where μ is the steady-state rate of investment and ω is the adjustment cost parameter. Then equation 1 can be rewritten as an investment equation:

$$(I_{it}/K_{i,t-1}) = m_I + (1/\omega)[p_{it} - p_t].$$

This equation cannot be estimated directly, in general, because (marginal) q is unobservable. However, Hayashi (1982) showed that if the

firm is a price taker in input and output markets, and the production function exhibits constants returns to scale, marginal q equals average q (denoted Q), defined as

$$Q_{it} = (V_{it} + B_{it})/K^{R}_{i,t-1},$$

where V is the market value of the firm's equity, B is the market value of the firm's debt, and K^{R} is the replacement value of the firm's capital stock. This formulation stresses the relationship between investment and the net profitability of investing, as measured by the difference between the value of an incremental unit of capital and the cost of purchasing capital. The hypothesis that investment in general is positively related to Tobin's q—the ratio of the stock market value of the firm to replacement costs—is now widely accepted.[22] Dornbusch and Fischer (1994, pp. 341–355) argue that "when [q] is high, firms will want to produce more assets, so that investment will be rapid," and therefore that "a booming stock market is good for investment," and that "the managers of the company can ... be thought of as responding to the price of the stock by producing more new capital—that is, investing—when the price of shares is high and producing less capital—or not investing at all—when the price of shares is low." Similarly, Hall and Taylor (1991, p. 312) state that "investment should be positively related to q. Tobin's q provides a very useful way to formulate investment functions because it is relatively easy to measure."[23]

When the firm has some market power (as pharmaceutical firms do, at least on their patented products), average Q is no longer exactly equal to (a perfect indicator of) marginal q, but it is still highly positively correlated with (a good indicator of) marginal q. Under these conditions, the estimated coefficient on Q will be biased toward zero (the magnitude of the bias depends on the "noise-to-signal ratio" in measured Q), and tests of the hypothesis that market value affects investment are "strong tests." Moreover, many economists believe that there are many industries in which firms exercise some market power, and there is much evidence at both the macro and micro level that is consistent with the q theory of investment. Hall and Taylor note that, in 1983, q was quite high, and investment was booming in the United States, even though the real interest rate and the rental price of capital were also high. Eisner (p. 112) presented microeconometric evidence that supports the theory; he found that "even given past sales changes, the rate of investment tends to be positively related to the market's evaluation of the firm both for the current year and the past year."

Investment is positively related to Q under imperfect as well as under perfect competition.

This evidence relates to fixed investment in the business sector as a whole, not specifically to R&D investment in the pharmaceutical industry. But Griliches and others have argued that many of the tools and models developed to analyze conventional investment can also fruitfully be applied to R&D investment. For example, there is a stock of "knowledge capital" (resulting from past R&D investment) analogous to the stock of physical capital. Therefore one would expect to observe a strong positive relationship between the market value of pharmaceutical firms and their rate of R&D investment. If this is the case, then government policy events that significantly reduce market value also tend to reduce R&D investment.

We will estimate the relationship between both R&D and fixed investment and Tobin's q using annual panel data for 46 publicly traded pharmaceutical firms included in the Compustat Annual Industrial File.

Tobin's q theory implies that investment in general (and R&D investment in particular) should be high when market value is high, *holding constant the firm's assets*. In a low-tech industry (e.g., the lumber industry), most of the firm's assets are tangible assets (property, plant, and equipment). But in the pharmaceutical industry, as in other high-tech industries, a significant part of the firm's assets are intangible—not recorded on the firm's balance sheet.[24] Hence, in order to effectively account (and statistically control) for firms' assets in the R&D investment equation, we need to first construct measures of firms' intangible assets.

We constructed two different kinds of intangible asset measures: input and output.

The "input" measure is the cumulated stock of past R&D investment (under alternative assumptions about depreciation of R&D). The "output" measure is the stock of FDA drug approvals, by type.

First we will examine the relationship between market value and measures of the firm's tangible and intangible assets and of its competitive environment. The estimates are presented in table 4. The dependent variable in all equations is the logarithm of market value, and all equations include fixed year effects. The regression in column 1 includes the logs of tangible assets (property, plant and equipment), the stock of R&D, and an inverse competition indicator—the reciprocal of the average number of firms selling each drug sold by the firm[25]—

Table 4
Regressions of market value of pharmaceutical firms on measures of tangible and intangible assets (t-statistics in parentheses). The dependent variable is the logarithm of market value. All equations include year dummies. Estimates are based on an unbalanced panel of firms during the period 1953–1996.

	1	2	3	4	5
Firm effects?	No	Yes	Yes	Yes	Yes
Log(tangible assets)	0.566	0.873	0.736	0.747	0.809
	(28.5)	(26.1)	(17.3)	(18.3)	(25.7)
Log(stock of R&D)	0.237	0.114	0.198	0.227	0.131
	(13.0)	(3.58)	(5.36)	(5.39)	(4.17)
Inverse competition	0.938	0.380			
	(11.0)	(2.20)			
Log(no. of NDAs)			0.213		
			(2.80)		
Log(no. of NMEs)				0.287	0.099
				(4.25)	(2.65)
Log(no. of other NDAs)				−0.008	
				(0.12)	
Log(no. of ANDAs)			0.007	0.058	
			(0.17)	(1.34)	
R^2	0.931	0.971	0.97	0.968	0.973
No. of firms	38	38	21	20	25
No. of observations	723	723	416	395	546

but does not include fixed firm effects. The regression in column 2 also includes fixed firm effects. This regression indicates that increases in a firm's tangible and intangible assets both increase its market value and that increases in the extent of competition it faces reduce its market value.

The regressions in columns 3–5 include measures of the firm's (cumulative) innovative output—various kinds of FDA approvals—as well as its innovative input (stock of R&D). Column 3 includes the number of New Drug Approvals (NDAs) and Abbreviated New Drug Approvals (ANDAs). The latter are approvals of generic drugs. The estimates indicate that NDAs have a positive effect on market value, but that ANDAs do not. As indicated earlier, there are several kinds of NDAs: new molecular entities (NMEs), new combinations, new formulations, etc. Only about a third of all NDAs are NMEs, which are generally thought to be the most medically and economically significant innovations. NDAs are disaggregated into two components—NMEs and other NDAs—in column 4. Increases in the number of

NMEs, but not of other NDAs, are associated with increases in market value. Due to the insignificance of both other NDAs and ANDAs, in column 5 we include only the number of NMEs. The estimated elasticity of market value with respect to the number of NMEs, conditional on the stocks of tangible and intangible assets, is 0.099.

The following conclusions may be drawn from these estimates. The market value of pharmaceutical firms is strongly related to the magnitudes of both intangible and tangible assets. Moreover, past R&D input and R&D output both positively affect market value. In other words, the market values both R&D effort and R&D productivity. The market values some kinds of FDA approvals more than others. In particular, market value is strongly related to the number of previously approved new molecular entities but unrelated to the number of abbreviated NDAs ("imitations"). Also, market value depends on the firm's competitive environment, holding constant its stocks of assets. The smaller the average number of competitors a firm faces in its various product markets, the greater its market value.

These findings indicate that the market value of pharmaceutical companies is determined, to an important extent, by observable indicators of their tangible and intangible assets and the returns to those assets. Of course, these indicators don't explain all of the cross-sectional and time-series variation in market value: there is variation in the ratio of market value to an index of the firm's assets (Tobin's q).

The q theory of investment predicts that the current rate of (R&D or fixed) investment should be positively related to q: firms should invest more when their market value is high, relative to their stocks of tangible and intangible assets. Estimates of R&D investment equations based on panel data for pharmaceutical firms are presented in table 5. The dependent variable in all of the equations is the logarithm of R&D expenditure, and all equations include fixed year effects. The equation in column 1 includes the logarithm of market value but not fixed firm effects. The coefficient on market value is close to 1. This is not surprising since both R&D expenditure and market value are closely linked to firm size (e.g., sales or employment): if firm A is twice as large as firm B, its market value will tend to be twice as high and it will perform twice as much R&D. The equation in column 2 includes fixed firm effects. In this equation, the coefficient on market value indicates the effect of changes in a firm's market value on its R&D expenditure. The coefficient is smaller than it is in column 1, but is still highly significant. The equation in column 3 includes several (time-varying) covari-

Table 5
Estimates of models of pharmaceutical firms' R&D investment (t-statistics in parentheses). Dependent variable: logarithm of R&D expenditure. All equations include year dummies. Estimates are based on an unbalanced panel of firms during the period 1953–1996.

	1	2	3
Firm effects?	No	Yes	Yes
Log(market value)	0.942	0.327	0.225
	(72.86)	(13.60)	(5.68)
Log(tang. assets)			0.011
			(0.28)
Log(stock of R&D)			0.431
			(15.6)
Log(cash flow)			0.114
			(4.77)
R^2	0.853	0.960	0.980
No. of firms	64	64	55
No. of observations	1,001	1,001	872

ates—current cash flow and the stocks of R&D and tangible assets—as well as the fixed firm effects. The market value coefficient declines by about a third, to 0.225, but is still highly significant. This indicates that a 10 percent increase in market value is associated with a 2.25 percent increase in R&D expenditure, holding constant tangible assets, past R&D investment, and cash flow. The estimates are therefore highly consistent with the predictions of the q theory of investment. Indeed, Tobin's q does a much better job of explaining investment in these regressions than it does in many studies in the tax-investment literature.

As Ellison and Mullin (1997, p. 9) note, "in February and March 1993, rumors circulated that the Clinton Health Care Reform Task Force, which was operating in secrecy, was going to include regulation of drug prices in its plan. Such fears seem supported by statements by President Clinton and Hillary Rodham Clinton attacking the high prices of vaccines and other pharmaceuticals." Ellison and Mullin estimated that the threat of Clinton health-care reform reduced the market value of pharmaceutical firms by about 44 percent during the period September 1992–October 1993. My estimate of the elasticity of R&D investment with respect to market value is 0.225. My model implies that this would tend to reduce R&D investment by about 9.9 percent (0.225×0.44). During the period 1986–2000, the average annual number of new molecular entities approved by the FDA was 28.1. Hence,

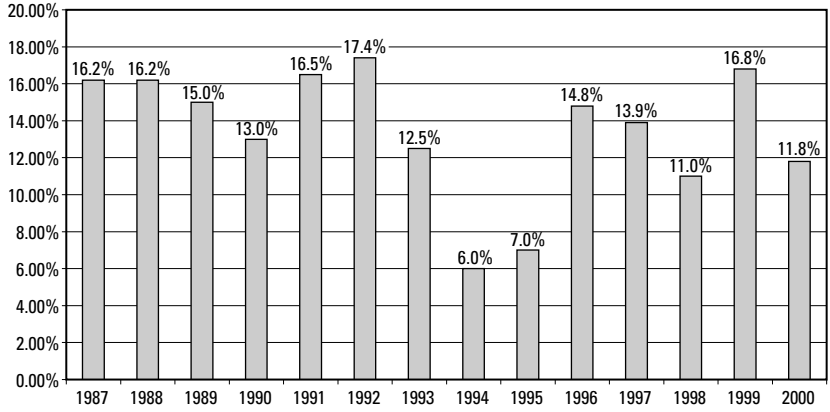


Figure 10
Annual percent change in pharmaceutical R&D, 1987–2000. Source: PhRMA Industry Survey.

the temporary reduction in R&D investment attributable to the threat of Clinton health-care reform may, with a lag of 12–15 years, temporarily reduce the number of new molecular entities approved by about 2.8 (9.9 percent $\times$ 28.1) per year.

Industry-level data are consistent with the hypothesis that the threat of Clinton health-care reform reduced the growth rate of R&D investment (with a 1- or 2-year lag). As figure 10 reveals, the annual growth of (nominal) R&D investment ranged from about 12 percent to 17 percent during the period 1987–1993. The 1993–94 and 1994–95 growth rates were less than half the growth rates of the previous 7 years. R&D growth in 1995–96 and later years was similar to the growth during the period 1987–1993.

The issue of pharmaceutical price controls has re-emerged within the context of the current debate about a Medicare outpatient drug benefit. Pharmaceutical manufacturers would benefit from Medicare drug coverage to the extent that it would lead to more purchases of their products. Industry leaders, however, have reacted cautiously to proposals for a broad-based benefit, since they fear that it might be tied to, or might eventually lead to, government measures to restrain drug prices. Since the potentially negative effect of Medicare drug coverage on drug prices would be offset by a positive effect on drug consumption, one would not expect Medicare drug benefit proposals—even those perceived as least favorable to the industry—to have as negative an effect

on market value (and R&D) as the 1993 Clinton proposals. The evidence seems to support this.

Summary and Conclusions

The pharmaceutical industry is one of the most R&D-intensive industries in the economy, and the government plays a larger role as both customer and regulator of it than it does of most other industries. This chapter offers brief analyses of the effects on innovation of three public policies—the 1962 Kefauver-Harris amendment, the 1992 Prescription Drug User Fee Act, and the 1965 Social Security Amendments (Titles XVIII and XIX: Medicare and Medicaid)—and detailed discussions of two policies: the 1984 Hatch-Waxman Amendments, and the first Clinton Administration's 1993 health-care-reform proposal, which was never implemented.

Introduction of a new drug requires approval by the FDA, and new molecular entities are the most important new drugs. Passage of the 1962 Kefauver-Harris amendment in response to the thalidomide tragedy appears to have led to a significant, roughly 10-year decline in the number of new drugs approved. Thirty years later, passage of the Prescription Drug User Fee Act, which increased the budget for, and accelerated, the drug approval process, may have led to a huge, albeit transitory, increase in the number of drugs approved in the mid 1990s.

The Medicaid and Medicare programs, established in 1965, have undoubtedly also had important effects on pharmaceutical innovation, via their effect on the demand for prescription drugs. In 1998, over 20 percent of national expenditure on prescription drugs was publicly funded, and the Medicaid program accounted for over 80 percent of this funding. The prices paid by the government are regulated under the Medicaid Drug Rebate Program.

Although Medicare has not, until the present, paid for most prescription drugs, it pays part of the cost of a service that is complementary with (necessary to receive) prescription drugs: doctor visits. There is evidence that Medicare and Medicaid contributed to the increase in longevity since 1965, and mean pharmaceutical consumption increases sharply with age. Public policy may have shifted both the rate and direction of private innovation: the orientation of drug development appears to have shifted toward the elderly after 1965.

In its study to evaluate the effects of the 1984 Hatch-Waxman Act on prices and returns in the pharmaceutical industry, the CBO concluded

that "the point in the life of an average drug at which generic entry occurs did not change much under the act, because the average length of a patent extension roughly offsets the average delay between patent expiration and generic entry that existed before 1984." My analysis of data on all new molecular entities approved by the FDA since 1940 does not support this conclusion: I find that imitation lags have been considerably shorter since 1984 than they were before 1984. For example, the probability that a new molecular entity introduced during the period 1940–1968 was imitated was 2 percent after 10 years and 7 percent after 15 years; for new molecular entities introduced during the period 1969–1996, these probabilities were 17 percent and 33 percent, respectively. Some theoretical models imply that an increase in the speed or ease of imitation will unambiguously reduce the rate of innovation, but others imply that the effect of a policy-induced fall in the private cost of imitation on the steady-state innovation rate may be positive, zero, or negative.

If firms base R&D investment decisions on their expectations about the present discounted value of future net cash flows, policies that affect these expectations will affect R&D investment. Under the market efficiency hypothesis, the value of the firm at time t is the expected present discounted value of its future net cash flows, conditional on the information available at time t. Hence policies that reduce market value might also be expected to reduce R&D investment. Estimates of R&D investment equations based on firm-level panel data are consistent with this hypothesis: firms invest more when their market value is high, holding constant tangible assets, past R&D investment, and cash flow. It has been estimated that the threat of Clinton health-care reform reduced the market value of pharmaceutical firms by about 44 percent during the period September 1992–October 1993. My model implies that this would tend to reduce R&D investment by about 8.8 percent (0.20×0.44). Industry-level data are consistent with the hypothesis that the threat of Clinton health-care reform reduced the growth rate of pharmaceutical R&D expenditure.

Appendix: A Cournot Duopoly Model of an Innovator and an Imitator

Consider a (homogeneous product) industry consisting of two firms: an innovator (firm 1) and an imitator (firm 2). The industry demand curve is

$$P = a - bQ \qquad (a > 1, b > 0),$$

where P is price and Q is total output. The innovator's marginal production cost is

$$m_1 = \exp(-\alpha X) \qquad (0 \leq \alpha < 1, X \geq 0),$$

where X is the innovator's R&D expenditure. If the innovator does no R&D, his marginal cost (MC) is 1. As R&D expenditure increases, his marginal cost declines, at a decreasing rate. The parameter α reflects the "effectiveness" or productivity of R&D.

The imitator does not have the opportunity to perform R&D but may receive spillovers from firm 1. The imitator's marginal production cost is

$$m_2 = \lambda m_1 + (1 - \lambda) \qquad (0 \leq \lambda \geq 1).$$

The parameter λ reflects the strength of R&D spillovers. If there are no spillovers ($\lambda = 0$), the imitator's marginal cost is 1 (regardless of the amount of R&D performed by firm 1). If there are complete spillovers ($\lambda = 1$), the imitator's MC is the same as the innovator's MC.

The innovator chooses the level of R&D investment that maximizes his profits (π_1). Assuming that the two firms behave as Cournot duopolists, the innovator's profit function is

$$\pi_1 = \frac{(a - 2m_1 + m_2)^2}{9b} - X.$$

Note that the innovator's profit is positively related to the imitator's MC (which is inversely related to the innovator's R&D expenditure). In the presence of spillovers, investing in R&D reduces the imitator's as well as the innovator's cost, which makes the imitator a more effective competitor.

Solution of the model implies that $dX/d\lambda < 0$: the equilibrium (innovator-profit-maximizing) level of R&D investment is inversely related to the R&D spillover rate.

Notes

1. Enactment of a Medicare drug benefit, currently under consideration by Congress, would probably result in a sharp increase in the share of industry output paid for with public funds.

2. This law was amended by the Veterans Health Care Act of 1992, which also requires a drug manufacturer to enter into discount pricing agreements with the Department of

Veterans Affairs and with covered entities funded by the Public Health Service in order to have its drugs covered by Medicaid. Approximately 500 pharmaceutical companies participate in this program. All 50 States and the District of Columbia cover drugs under the Medicaid program. As of January 1, 1996, the rebates for covered outpatient drugs were as follows: for Innovator Drugs, the larger of 15.1% of the Average Manufacturer Price (AMP) per unit or the difference between the AMP and the best price per unit and adjusted by the CPI-U based on launch date and current quarter AMP; for Non-innovator Drugs, 11% of the AMP per unit.

3. Source: http://www.fda.gov.

4. Other kinds of new drugs include new combinations and new formulations.

5. Thalidomide is a drug that was marketed outside of the United States in the late 1950s and the early 1960s. It was used as a sleeping pill, and to treat morning sickness during pregnancy. However, its use by pregnant women resulted in the birth of thousands of deformed babies.

6. "Substantial evidence" is defined by section 505(d) of the FD&C Act as "evidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could be fairly and responsibly concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling thereof."

7. At least one prescription drug is prescribed in more than 60% of doctor visits; this percentage is even higher for the Medicare population (Woodwell 1999, table 19).

8. Life expectancy at birth of Americans increased approximately 10% (from 69.7 to 76.5 years) between 1960 and 1997.

9. "Why Do Prescription Drugs Cost So Much, and Other Questions About Your Medicines," p. 2.

10. A similar abbreviated process already existed under FDA regulations for generic copies of antibiotics and of innovator drugs approved before 1962.

11. If an innovator drug is not protected by a patent, it may still benefit from certain exclusivity provisions that delay the approval or filing of an abbreviated new drug application in some cases.

12. There are essentially six application types: (1) New molecular entity, or NME: An active ingredient that has never been marketed in this country; (2) New derivative: A chemical derived from an active ingredient already marketed (a "parent" drug); (3) New formulation: A new dosage form or new formulation of an active ingredient already on the market; (4) New combination: A drug that contains two or more compounds, the combination of which has not been marketed together in a product; (5) Already marketed drug product but a new manufacturer: A product that duplicates another firm's already marketed drug product: same active ingredient, formulation, or combination; (6) Already marketed drug product, but a new use: A new use for a drug product already marketed by a different firm.

13. The fields for ingredients 2–13 were usually blank; most NDAs and ANDAs are for single-ingredient drugs.

14. For an excellent discussion of survival data analysis, see Kalbfleisch and Prentice 1980.

15. The estimated standard errors of the δ_ts are approximately 0.021.

16. This model is completely static; it does not incorporate imitation lags. However it seems reasonable to view the spillover rate as inversely related to the imitation lag: shorter imitation lags imply greater R&D spillovers, and therefore less R&D investment.

17. Drawing an analogy between branded and generic pharmaceutical firms and Northern and Southern firms in the Grossman-Helpman model seems reasonable.

18. For a discussion of this point, see Jones 1998.

19. Jaffe and Palmer (1996) examined the effect of *environmental* regulation on innovation.

20. Malkiel distinguishes between three forms of the Efficient-market Hypothesis: the *weak form* (random-walk hypothesis), which states that investment returns are serially independent; the *semi-strong form*, which states that all public information about a company is already reflected in the stock's price; and the strong form which states that no technique of selecting a portfolio can consistently outperform a strategy of simply buying and holding a diversified group of securities that make up the popular market averages.

21. This derivation is adapted from the one provided by Hassett and Hubbard (1998). For simplicity, we ignore taxes.

22. According to the theory, Tobin's q is equal to the ratio of the marginal benefit of capital to the marginal (user) cost of capital. When Tobin's is high (greater than 1), the firm should invest. We hypothesize that changes in the environment affect the (marginal) expected benefits and costs of "knowledge capital," hence the incentives to invest in R&D.

23. Hassett and Hubbard (1998, p. 31) argue that determining "the response of investment to [q]" is "easiest during periods in which large exogenous changes in the distribution of structural determinants occur, as during tax reforms"—or, in our context, changes in the regulatory or legal environment.

24. For a discussion of accounting for intangible assets in the pharmaceutical and other industries, see Lev and Sougiannis 1996. Their methodology has the implausible implication that the firm realizes most of the economic benefits of pharmaceutical R&D investment within 5 years. The lag from investment to FDA approval and marketing is generally recognized to be much longer.

25. For example, if the firm sells two drugs, is the sole producer of one, and has four competitors for the other, the value of its inverse competition measure is $1/\text{mean}(1, 5) = 1/3$.

References

Chirinko, Robert S. 1992. Do tax incentives work? The real effects of the Tax Reform Act: "Comment." *National Tax Journal* 45 (3) (September): 291–297.

Conroy, Robert, Robert Harris, and Thomas Massaro. 1992. Assessment of Risk and Capital Costs in the Pharmaceutical Industry. Unpublished paper, University of Virginia Darden School.

Cornett, Marcia, and Hassan Tehranian. 1992. Changes in corporate performance associated with bank acquisitions. *Journal of Financial Economics* 31, April: 211–234.

Dornbusch, Rudiger, and Stanley Fischer. 1994. *Macroeconomics*, sixth edition. McGraw-Hill.

Eisner, Robert. 1978. *Factors in Business Investment*. HarperBusiness.

Ellison, Sara Fisher, and Wallace P. Mullin. 1997. Gradual Incorporation of Information into Stock Prices: Empirical Strategies. Working paper W6218, National Bureau of Economic Research.

F&S Index United States Annual. Various years. Information Access Co.

Grabowski, Henry. 1968. The determinants of industrial research and development: A study of the chemical, drug, and petroleum industries. *Journal of Political Economy* 76: 292–306.

Grabowski, Henry, and John Vernon. 1981. The determinants of research and development: Expenditures in the pharmaceutical industry. In *Drugs and Health*, ed. R. Helms. American Enterprise Institute.

Grossman, Gene M., and Elhanan Helpman. 1991. *Innovation and Growth in the Global Economy*. MIT Press.

Hall, Robert, and John Taylor. 1991. *Macroeconomics*, third edition. Norton.

Hassett, Kevin, and R. Glenn Hubbard. 1998. Tax Policy and Investment. Working paper 5683, National Bureau of Economic Research.

Hayashi, Fumio. 1982. Tobin's marginal q and average q: A neoclassical interpretation. *Econometrica* 50: 213–224.

Healy, Paul, Krishna Palepu, and Richard Ruback. 1992. Does corporate performance improve after mergers? *Journal of Financial Economics* 31, April: 135–175.

Jaffe, Adam, and Karen Palmer. 1996. Environmental regulation and innovation: a panel data study. Working paper 5545, National Bureau of Economic Research.

Jensen, Elizabeth. 1987. Research expenditures and the discovery of new drugs. *Journal of Industrial Economics* 36: 83–95.

Jensen, Elizabeth. 1988. The determinants of research and development expenditures in the ethical pharmaceutical industry. Discussion and working paper 88/7, Department of Economics, Hamilton College.

Jones, Charles. 1998. *Introduction to Economic Growth*. Norton.

Kalbfleisch, J., and R. Prentice. 1980. *The Statistical Analysis of Failure Time Data*. Wiley.

Kaplan, Steven, and Richard Ruback. 1995. The valuation of cash flow forecasts: An empirical analysis. *Journal of Finance* 50, September: 1059–1093.

Lev, Baruch, and Theodore Sougiannis. 1996. The capitalization, amortization, and value-relevance of R&D. *Journal of Accounting and Economics* 21, no. 1: 107–138.

Lichtenberg, Frank. 1988. The private R&D investment response to federal design and technical competitions. *American Economic Review* 78, no. 3: 550–559.

Lichtenberg, Frank. 1992. *Corporate Takeovers and Productivity*. MIT Press.

Lichtenberg, Frank. 1996. Do (more and better) drugs keep people out of hospitals? *American Economic Review* 86, May: 384–388.

Lichtenberg, Frank. 2000. The effect of pharmaceutical utilization and innovation on hospitalization and mortality. In *Productivity, Technology, and Economic Growth*, ed. B. Van Ark et al. Kluwer.

Lichtenberg, Frank. 2002. Sources of U.S. Longevity Growth, 1960–1997. Working paper 8755, National Bureau of Economic Research.

Malkiel, Burton. 1990. *A Random Walk down Wall Street*, fifth edition. Norton.

Rogerson, William. 1989. Profit regulation of defense contractors and prizes for innovation. *Journal of Political Economy* 97: 1284–1305.

Sakakibara, Mariko, and Lee Branstetter. 1999. Do Stronger Patents Induce More Innovation? Evidence from the 1988 Japanese Patent Law Reforms. Working paper W7066, National Bureau of Economic Research.

Summers, Lawrence H. 1981. Taxation and corporate investment: A *q*-theory approach. *Brookings Papers on Economic Activity* 1: 67–127.

Tobin, James. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1, February: 15–29.

Tobin, James, and William Brainard. 1977. Asset markets and the cost of capital. In B. Balassa and R. Nelson, eds., *Economic Progress, Private Values, and Public Policy*. North-Holland.

US Food and Drug Administration. 2003. *Electronic Orange Book.* 23rd edition. Government Printing Office.

Woodwell, D. A. 1999. *National Ambulatory Medical Care Survey: 1997 Summary, Advance data from vital and health statistics, no. 305*. National Center for Health Statistics.

5 Dimensions of Nonprofit Entrepreneurship: An Exploratory Essay

Joseph J. Cordes, C. Eugene Steuerle, and Eric Twombly

Because entrepreneurship is typically associated with the creation of new business ventures and innovation in the for-profit sector of the economy, "nonprofit entrepreneurship" may seem to be a contradiction in terms. Yet many large and successful nonprofit organizations that exist today can trace their lineage back to the enterprise and vision of a founder, such as the International Red Cross (Jean-Henri Dunant), Mothers Against Drunk Driving (Candace Lightner), Girl Scouts (Juliet Lowe), and Tax Analysts and Advocates (Thomas Field). More locally, it is also not unusual to find a press account of a recently founded nonprofit that appears to be meeting a particular need in a new and creative way.

Thus, the growth and evolution of organizations in the nonprofit sector of the economy, which by some estimates accounts for roughly 7 percent of the U.S. GDP, is thus clearly shaped by individuals who fit Webster's definition of an entrepreneur as "one who *organizes*, *manages*, and *assumes the risks* of a business or *enterprise*."[1] There is also suggestive, though still largely anecdotal, evidence that some new socially oriented businesses have been established by entrepreneurs who seek to combine for-profit ventures with an explicit charitable purpose.

Scholars have paid some attention to what can be described as entrepreneurial behavior by managers of existing nonprofit organizations.[2] But, aside from Bowen et al. (1994), less attention has been given to studying the individual and environmental factors that affect the creation of *new* nonprofit enterprises,[3] and still less to examining why some for-profit entrepreneurs may be motivated to harness their talents in the pursuit of social or charitable purposes.

With this in mind, our chapter focuses on several questions pertaining to the formation of new enterprises with a charitable or social mission.

- What are recent patterns and trends in the formation of new traditional nonprofit organizations, and of new "socially oriented" for-profit enterprises?
- Why might rational economic actors invest their time, talents, and even financial resources to create new nonprofits and/or socially oriented for-profits?
- How do external factors, such as demand for charitable outputs, access to financing for new ventures, and the blurring of the boundaries between for-profit and not-for-profit activities affect the creation of new nonprofit and socially oriented for-profit enterprises?
- How does public policy shape the incentives for individuals to become nonprofit entrepreneurs, and the external environment in which new organizations come into being?

Births and Deaths among Organizations with Charitable Purposes

The volume of startups among both traditional nonprofits and socially oriented for-profit ventures is a measure of the scope of nonprofit entrepreneurship that is analogous to the number of new business formations that is often used to gauge for-profit entrepreneurial activity.[4] In this section we present tabulations of the number of new traditional nonprofit organizations drawing on data from the National Center for Charitable Statistics (NCCS), which is the national repository of data on the nonprofit sector.[5] We also summarize some anecdotal evidence about the creation of for-profit ventures with explicitly charitable or social missions.

Growth and Change in the Number of Operating Public Charities

Although the tax code recognizes several different forms of nonprofit tax-exempt enterprises, within this broad group we focus attention on the formation of new "charitable" nonprofits, or 501(c)(3) organizations, that are eligible to receive tax deductible contributions from individuals and businesses.[6] Within the general category of charities, we focus further on "operating charities," or those which are eligible to receive tax deductible contributions *and* are classified as providing a tangible service, as distinguished from nonprofits whose purpose is to support other operating charities. Our period of analysis is 1992 to 1996, during which time there were more than 300,000 501(c)(3)

operating charities in the United States. We use the date on which a charity is officially recognized as a 501(c)(3) organization by the Internal Revenue Service as indicating the organization's date of entry or formation.[7]

Table 1 presents data on the change in the number of operating charities between 1992 and 1996, and on the components of change. Comparison of columns 1 and 5 shows that the number of operating public charities increased significantly from 1992 to 1996. While roughly 190,000 operating charities were in existence in the United States at the beginning of 1992, the number had increased by nearly 75 percent to approximately 245,000 groups by the close of 1996.[8] The annual growth rate of operating charities was just over 5.0 percent during this period, which was considerably higher than the growth in the number of for-profit businesses, which according to the U.S. Department of Commerce expanded at an annual rate of roughly 1.4 percent between 1992 and 1997.[9]

The higher rate of growth in the number of nonprofit enterprises between 1992 and 1996 is not just a phenomenon of the 1990s. It is broadly consistent with comparative trends that have been observed in nonprofit and for-profit sectors in previous years. For example, Hodgkinson et al. (1996), report that between 1977 and 1992 the number of operating nonprofits grew at an annual growth rate of 4.7 percent from 1977 to 1992 compared with an annual growth rate of 3.0 percent in for-profit businesses. Similarly, Bowen et al. (1994) note that, among public charities, between 1981 and 1991, the number of entrants grew at an annual rate of 6.5 percent, compared with an annual increase in the rate of business incorporations over the same period of 5 percent per year.

Entry and Exit Rates

Columns 2 and 4 in table 1 also show how organizational entry and exit affects the overall change in the number of operating nonprofits. Since new organizations must generally be founded by someone, these data provide a rough statistical gauge of the importance of entrepreneurship to institutional growth and change.[10]

Table 1 shows that nearly 130,000 new operating charities officially came into being between 1992 and 1997.[11] Table 2 presents entry rates and further breakdowns of exits among both startups and existing nonprofits. Subtracting startup exits shown in column 4 from total

Table 1
Entry and exit of nonprofit organizations, 1992–1996. Source: National Centery for Charitable Statistics, Center on Nonprofits and Philanthropy, Urban Institute.

Type of organization	(1) Number, 1992	(2) Entrants, 1992–1996	(3) Startup exits, 1992–1996	(4) Existing exits, 1992–1996	(5) Number, 1996
Arts, culture, humanities	20,847	13,906	5,345	3,336	26,072
Education (not K–12)	29,232	17,018	6,116	4,914	35,220
K–12 education	3,205	1,279	459	170	3,855
Environment	3,238	4,280	2,200	570	4,748
Animals	2,382	1,669	529	280	3,242
Health, general	18,547	5,600	1,440	2,211	20,496
Health, mental	6,347	2,256	675	784	7,144
Disease, disease disorders	4,271	866	178	604	4,355
Medical research	1,553	790	242	258	1,843
Crime, legal related	3,320	2,334	662	499	4,493
Employment, job related	3,046	1,026	273	338	3,461
Food, agriculture, nutrition	1,944	673	176	214	2,227
Housing, shelter	8,697	5,108	1,298	943	11,564
Public safety, disaster relief	2,023	1,610	395	290	2,948
Recreation, sports & leisure	11,813	6,712	2,008	2,092	14,425
Youth development	5,364	3,938	1,957	754	6,591
Human services, multipurpose	28,189	12,703	3,683	3,189	34,020
Int. & fgn. affairs	1,821	1,250	354	362	2,355
Civil rights & advocacy	1,386	1,363	656	244	1,849
Community improvement	7,636	6,056	2,134	1,449	10,109
Philanthropy & grantmaking	9,152	9,794	1,597	1,608	15,741
Science & tech. rsch. inst.	1,452	673	242	229	1,654
Social sci. rsch. inst.	621	279	49	84	767
Public & societal benefit	1,545	913	345	256	1,857
Religion related	9,037	21,943	10,699	2,113	18,168
Mutual/membership benefit	608	200	84	87	637
Unknown/unclassified	2,931	4,401	2,079	763	4,490
All nonprofits	190,207	128,640	45,875	28,641	244,331

startups in column 2 shows that almost 80,000, or about three out of five new nonprofits that were formed between 1992 and 1996 were still in existence in 1996.

Table 2 also shows that entry and survival (or exit) rates differ among different nonprofit activities, and also between new and established organizations. For example, column 5 in table 2 indicates that new nonprofits that provided either health or social services were less likely to exit (more likely to survive) than, for example, new nonprofits providing arts and cultural services, environmental, or education services. A comparison of columns 5 and 7 shows further, as might be expected, that new entrants as a group were also considerably more likely to exit than their more "established counterparts."

Table 3 shows the result of entry and exit on the composition of "new" and "old" organizations as of 1996. In 1996 one out of three operating nonprofits had been founded within the preceding 5 years; in some sectors, over half of operating charities were new entrants.

The general statistical portrait painted in tables 1–3 seems clear. The population of operating public charities has experienced considerable growth and change that is fostered by the creation of new organizations.

Data Limitations and Caveats

The files maintained by NCCS comprise the most comprehensive time-series data on nonprofits. Nonetheless, the use of these data to track the formation of new charitable nonprofits is subject to some caveats.

First, as has already been noted, we follow Bowen et al. (1994) and Twombly (2000), in using the IRS ruling date as indicating the date of nonprofit entry. But a nonprofit organization may already be in existence in some form before receiving formal recognition from the IRS as a 501(c)(3) organization.

Second, small nonprofits with less than $25,000 in annual gross revenue, and most religious congregations are not required to seek formal tax-exempt status from the IRS. These charities are not included in the NCCS data, which include only organizations that are legally required to file the IRS Form 990 information return. Smith (1997) argues that focusing on organizations that file the IRS 990 return excludes many grassroots nonprofit organizations function that do not need to see formal IRS recognition.

A third caution stems from the manner in which organizations are classified in the data files. NCCS applies a code from the National

Table 2
Entry and exit rates, 1992–1996. Source: National Center for Charitable Statistics, Center on Nonprofits and Philanthropy, Urban Institute.

	Existing orgs.	Entrants		Startup exits		Existing exits	
Type of organization	(1) Number, 1992	(2) Number	(3) Rate: (2)/(1)	(4) Number	(5) Rate: (4)/(2)	(6) Number	(7) Rate: (6)/(1)
Arts, culture, humanities	20,847	13,906	0.67	5,345	0.38	3,336	0.16
Education (not K–12)	29,232	17,018	0.58	6,116	0.36	4,914	0.17
K–12 education	3,205	1,279	0.40	459	0.36	170	0.05
Environment	3,238	4,280	1.32	2,200	0.51	570	0.18
Animals	2,382	1,669	0.70	529	0.32	280	0.12
Health, general	18,547	5,600	0.30	1,440	0.26	2,211	0.12
Health, mental	6,347	2,256	0.36	675	0.30	784	0.12
Disease, disease disorders	4,271	866	0.20	178	0.21	604	0.14
Medical research	1,553	790	0.51	242	0.31	258	0.17
Crime, legal related	3,320	2,334	0.70	662	0.28	499	0.15
Employment, job related	3,046	1,026	0.34	273	0.27	338	0.11
Food, agriculture, nutrition	1,944	673	0.35	176	0.26	214	0.11
Housing, shelter	8,697	5,108	0.59	1,298	0.25	943	0.11
Public safety, disaster relief	2,023	1,610	0.80	395	0.25	290	0.14
Recreation, sports & leisure	11,813	6,712	0.57	2,008	0.30	2,092	0.18
Youth development	5,364	3,938	0.73	1,957	0.50	754	0.14
Human services, multipurpose	28,189	12,703	0.45	3,683	0.29	3,189	0.11
Int. & fgn. affairs	1,821	1,250	0.69	354	0.28	362	0.20
Civil rights & advocacy	1,386	1,363	0.98	656	0.48	244	0.18

Community improvement	7,636	6,056	0.79	2,134	0.35	1,449	0.19
Philanthropy & grantmaking	9,152	9,794	1.07	1,597	0.16	1,608	0.18
Science & tech. rsch. inst.	1,452	673	0.46	242	0.36	229	0.16
Social sci. rsch, inst.	621	279	0.45	49	0.18	84	0.14
Public & societal benefit	1,545	913	0.59	345	0.38	256	0.17
Religion related	9,037	21,943	2.43	10,699	0.49	2,113	0.23
Mutual/membership benefit	608	200	0.33	84	0.42	87	0.14
Unknown/unclassified	2,931	4,401	1.50	2,079	0.47	763	0.26
All nonprofits	190,207	128,640	0.68	45,875	0.36	28,641	0.15

Table 3
New entrants as a share of all nonprofits, 1996. Source: National Centery for Charitable Statistics, Center on Nonprofits and Philanthropy, Urban Institute.

Type of organization	Total	Surviving new entrants	Percent
Arts, culture, humanities	26,072	5,345	20.5
Education (not K–12)	35,220	10,902	31.0
K–12 education	3,855	820	21.3
Environment	4,748	2,080	43.8
Animals	3,242	529	16.3
Health, general	20,496	4,160	20.3
Health, mental	7,144	1,581	22.1
Disease, disease disorders	4,355	688	15.8
Medical research	1,843	548	29.7
Crime, legal related	4,493	1,672	37.2
Employment, job related	3,461	753	21.8
Food, agriculture, nutrition	2,227	497	22.3
Housing, shelter	11,564	3,810	32.9
Public safety, disaster relief	2,948	1,215	41.2
Recreation, sports & leisure	14,425	4,704	32.6
Youth development	6,591	1,981	30.1
Human services, multipurpose	34,020	9,020	26.5
Int. & fgn. affairs	2,355	896	38.0
Civil rights & advocacy	1,849	707	38.2
Community improvement	10,109	3,922	38.8
Philanthropy & grantmaking	15,741	8,197	52.1
Science & tech. rsch. inst.	1,654	431	26.1
Social sci. rsch inst.	767	230	30.0
Public & societal benefit	1,857	568	30.6
Religion related	18,168	11,244	61.9
Mutual/membership benefit	637	116	18.2
Unknown/unclassified	4,490	2,322	51.7
All nonprofits	244,331	78,938	32.3

Taxonomy of Exempt Entities (NTEE) classification system to categorize the primary organizational activity of each nonprofit in the NCCS data files.[12] Although the NTEE is widely utilized in nonprofit sector research, some have raised concerns about its reliability and validity (Salamon and Anheier 1992; Gronbjerg 1994). Indeed, while the NTEE system is quite useful for analyzing broad sets of similar organizations, such as human service nonprofits, it becomes problematic when identifying nonprofits that provide more specialized services, such job training or respite care to AIDS patients.

The implications of these limitations is that using IRS data to track the entry (and exit) of nonprofit organizations will result in treating some organizations as "new" that are already in existence, and miss altogether the formation of some (very) small nonprofits. In addition, the vagaries of the NTEE classification system mean that the NCCS data can result in misclassifying the true outputs and activities of some nonprofits.

Nonetheless, we believe the NCCS data provide a reasonable picture of nonprofit entry and exit for several reasons. First, although in individual cases the IRS data may capture the entry of some nonprofits with a lag, organizations that seek formal recognition from the IRS are apt to do so fairly soon after their initial "informal" formation because formal IRS recognition confers a number of legal and tax advantages. Indeed, IRS regulations allow organizations that file within fifteen months of their "informal founding" to "back-date" their "official founding" as a 501(c)(3) charity to the actual founding date, instead of the time at which a ruling is requested.[13] Thus, using the IRS ruling date seems to be a reasonable proxy for formation of a new organization. Second, although the data based on IRS 990 returns do not include the full array of not-for-profits, a majority of nonprofit organizations (aside from religious congregations) are formally registered with the IRS, and these organizations account for a substantial share of the financial resources that flow through the nonprofit sector (Weisbrod 1988, p. 82). Thus, focusing on nonprofits that file the IRS 990 return captures the majority of enterprises that are eligible to benefit from tax deductible donations, to participate in federated campaigns, and to receive government contracts and foundation grants. Last, the potential for misclassifying individual organizations that by using the NTEE system can be dealt with to some extent by using "higher" rather than "lower" levels of aggregation. Thus, the tabulations presented in tables

1–3 generally focus on broad groupings of nonprofits in order to increase the utility of the NTEE system.

Social Venturing and Formation of Charitable For-Profit Enterprises

The data presented in tables 1–3 provide a picture of the potential importance of the "traditional" mode of nonprofit entrepreneurship in which a new 501(c)(3) nonprofit organization is formed to meet a charitable need. In recent years, however, increasing attention has been paid to a different form of entrepreneurship that appears to combine the creation of new for-profit enterprises with an explicit charitable intent.

In some cases these new "charitable for-profits" represent businesses that are founded to teach market skills to needy individuals such as drug addicts, runaway youth, and youthful offenders. In others, the new business is a for-profit venture founded by someone who may recognize a social need, but who chooses to found a for-profit businesses as a means of creating new wealth to help meet these needs, instead of founding a new nonprofit organization and then seeking funding from other sources.

The exact scope of "for-profit social venturing" is unknown because there are no empirical data either about the number of new for-profit enterprises that have been founded by social entrepreneurs, or the amount of financial support that such enterprises provide to charity. At present, it appears that the scope of these activities is modest; and it would be hard to demonstrate that at least as of yet for-profit entrepreneurs who are "charitable entrepreneurs in disguise" have displaced more traditional for-profit or not-for-profit entrepreneurs.

At the same time, anecdotal evidence indicates that such enterprises exist. For example, the Roberts Enterprise Development Fund of San Francisco supports 10 nonprofit organizations that together have founded and operate more than 20 for-profit businesses whose mission is both to earn profits and to provide job training (Streisand 2001). Pioneer Humans Services of Seattle integrates self-supporting commercial businesses with a range of services for its clients who include former offenders and substance abusers.[14] Commercial enterprises established by the Los Angeles Venture Fund Initiative sell goods and services including salad dressing (Food from the Hood), janitorial services (Pueblo Nuevo Development), and computer support (Breakaway Technologies) (Buttenheim 1998).

These types of organizations are also seen to be potentially important as evidenced by creation of groups of individuals and organizations committed to expanding the scope of social entrepreneurship; and the creation of courses at well-regarded graduate schools of businesses to teach social entrepreneurial skills.[15] Indeed, as observed by Gregory Dees, who has written extensively on social venturing, one should not be surprised that organizations with social missions may choose the for-profit form if that choice better serves the mission than does organization as a nonprofit.[16]

Motivations of Nonprofit Entrepreneurs

Whether nonprofit entrepreneurship takes the "traditional" form of creating a new nonprofit enterprise, or follows the social venturing model of founding a new for-profit business with a charitable mission, it is plausible to assume that in most cases the impetus for creating these new organizations comes from an individual, or a group of individuals. These persons become nonprofit entrepreneurs when they identify a need or a "demand" for some type of charitable good or service, and then spend time and energy assembling the productive inputs that are needed to satisfy that demand, using either the nonprofit or the for-profit form of organization.

Utility Maximization and Charitable Impulses

What factors might motivate individuals to become traditional nonprofit entrepreneurs and found new 501(c)(3) charities? As Jerald Schiff (1986) and Schiff and Weisbrod (1993) have noted, an attribute of nonprofit entrepreneurs that distinguishes them from their for-profit counterparts is that they are "utility maximizers" rather than "profit maximizers." That is, they derive satisfaction from providing some charitable output or fulfilling a social mission rather than simply from pursuing profit as a means of increasing their income.[17]

Schiff assumes that each actual or potential nonprofit manager or entrepreneur has some reservation utility, U_R, equal to the level of well-being that could be attained in some alternative activity. Then, if U^* is the maximum level of well-being that can be attained, the equilibrium entry/exit condition corresponding to the zero-profit condition in the for-profit entry/exit model is that $U^* = U_R$. Entry thus occurs whenever $U^* > U_R$, and exit whenever $U^* < U_R$.

As will be seen below, treating nonprofit and socially minded for-profit entrepreneurs as utility maximizers rather than profit maximizers is a useful heuristic device for examining factors in the external environment that might encourage or discourage nonprofit entrepreneurship through their effects on U^* and U_R. But to argue that someone becomes a charitable entrepreneur to attain a higher level of utility begs the question of why such a person would choose to channel his or her entrepreneurial abilities through the traditional nonprofit form of organization—for example, by founding a new 501(c)(3) organization, or why, as has been suggested, a more effective means in some cases may be to create what might better be described as a for-profit business that is a "charity in disguise"—for example, found a new social venture.[18]

Nonprofit Organizational Form and the Nondistribution Constraint

We consider first what factors might affect the propensity of individuals to become traditional nonprofit entrepreneurs. To this end, a brief digression is in order on what economists take to be the main difference between for-profit and not-for-profit enterprises. In popular parlance, the term "not-for-profit" is often taken to mean that an enterprise organized as a nonprofit does not earn a profit on its activities. This common perception, however, is economically misleading. Profit, after all, is simply a positive difference between revenues realized from providing a service and costs of providing the service, and there is no reason in principle why an organization that provides even a traditional charitable service, such as helping the homeless, could not "earn" a profit. Indeed, when one examines the financial statistics of many nonprofit organizations, more than half of such organizations report surpluses. Moreover, even an organization that reported little or no accounting surplus might nonetheless still earn a "hidden surplus" that was paid out in the form of wages and salaries, or other perks.

If earning a profit does not distinguish a nonprofit from a for-profit organization, what is the defining difference? According to Henry Hansmann (1987) it is the imposition of the so-called nondistribution constraint, which prevents any surplus that is garnered from nonprofit economic activity from being directly distributed to owners in the form of equity shares as in the case of for-profit business. (As noted above, however, such a nondistribution constraint might still allow a portion

of the surplus to be distributed to members of the organization in the form of salary and perks.).

With this distinction in mind, the question of what might prompt individuals to found traditional nonprofit can be rephrased. It is understandable why traditional economic self-interest would prompt individual entrepreneurs to organize for-profit enterprises to meet new demands for goods and services. But, what would motivate economically rational individuals to invest time and even financial resources to create enterprises whose organizational form expressly limits the disposition of any economic surplus that might result from such activities?

Nonprofit Organizational Form and the Provision of Public Goods

One answer to this question lies in the "consumption technologies" of certain goods and services. It has long been recognized that nonprofit organizations often provide public, or collective, consumption goods. Weisbrod (1977, 1988) develops a formal model in which nonprofit organizations come into existence in order to meet demands for such goods that are not satisfied through traditional tax-financed channels. The model implicitly assumes that the nonprofit form is the preferred vehicle for organizing the provision of such goods, but it does not explicitly consider the question of why this particular organizational form would be preferred.

This question is taken up by Bilodeau and Slivinski (1998), who model the interaction between people who desire to support the provision of collective consumption goods and services and those who found the organizations that satisfy these demands. As Bilodeau and Slivinski note, if there is no institution to provide the good or service (e.g., feeding the homeless in the 1970s), it will not be produced, even if there are enough people who would contribute to finance some positive quantity, unless someone is prepared to collect these contributions and actually organize production and distribution of the charitable good.

In this setting, the nonprofit entrepreneur becomes the member of the group of "latent" or potential demanders who assumes the responsibility for creating the new organization. But in so doing, the entrepreneur has a choice. Should the new organization be organized as a for-profit enterprise or as a nonprofit-organization (thereby imposing the nondistribution constraint)?

In the Bilodeau-Slivinksi analysis, choice of institutional form is modeled as the outcome of a game between the group of latent demanders and the entrepreneur/founder of the new enterprise. All players in the game are assumed to be economically rational in the sense that each player seeks to maximize the utility that can be obtained from consuming the public good. The game itself has several stages:

(1) A member of the group of latent demanders must choose to step forward and become an entrepreneur.

(2) The person who becomes the entrepreneur must choose between the for-profit and not-for profit form of organization.

(3) Conditional on the chosen form of organization, the nonprofit entrepreneur must decide how much of her own income or wealth to invest in the new enterprise.

(4) Conditional on the chosen form of organization, and her own contribution, the entrepreneur collects contributions from the other latent demanders.

(5) Conditional on total contributions collected, the entrepreneur decides how much to produce of the public good and how to produce it.

In the last stage of the game, when all contributions have been collected, the entrepreneur ultimately has the final say on how much of the public good is to be produced. At this point, if there is no nondistribution constraint—that is if the enterprise is organized as a for-profit firm—nothing would prevent the entrepreneur from distributing some or all of the other players' contributions to herself. In contrast, imposing a credible, binding nondistribution constraint by organizing as a nonprofit organization prevents such an outcome from happening.

In this set-up the other players in the game know that, in the final stage, if there is no nondistribution constraint, the utility-maximizing entrepreneur may have not only an incentive but also an institutional opportunity to appropriate their contributions for her (the entrepreneur's) private use. The other players' rational response in this case would be to withhold making contributions to a for-profit firm. In this outcome, no contributions are collected and none of the public good is provided.

In contrast, organizing as a nonprofit organization and imposing the nondistribution constraint provides a means whereby the nonprofit

entrepreneur can credibly signal to the other players that their contributions will, in fact, "go 100 percent to charity" as opposed to being distributed to the entrepreneur; and this signal gives the other players a sufficient incentive to make at least some positive contributions to provide the public good. This result has two important implications. First, the entrepreneur needs to contribute less of her own resources toward provision of the public good if provision is organized through a nonprofit organization because other players are likely to give more. Second, the total amount of contributions that are raised from all players will be greater if provision is organized through not-for-profit than through for-profit form. Thus, the rational strategy of a utility-maximizing entrepreneur who cares about meeting a particular social need (providing a public good) is to establish a nonprofit organization.

It should, however, be emphasized that the above outcome depends critically on the willingness of the latent demanders to believe that the nondistribution constraint is actually binding. If it is not, simply labeling an organization as a nonprofit would not, in the eyes of the other players, prevent the entrepreneur from appropriating their contributions, and they would then have no more economic incentive to contribute than if the new enterprise is labeled as a nonprofit.

This point has an obvious but important policy implication. Namely, government regulations, such as the non-inurement rules of the Internal Revenue Code that prevent employees of nonprofit organizations from receiving "excessive" compensation, have a broader social purpose than merely preventing individual cases of fraud and/or abuses of the tax-exempt status of nonprofits. Such regulations also provide a legal framework that helps make the nondistribution constraint binding; and this in turn provides the incentives that enable entrepreneurs to successfully organize the voluntary finance of public goods through nonprofit organizations.

Nonprofit Form and the Provision of Goods Where Quality Matters

Weisbrod (1988) and Hansmann (1987) have also suggested that the nonprofit form of organization may be well suited to cases where the good or service has important yet difficult-to-verify dimensions of quality. Glaeser and Shleifer take this insight as a starting point for modeling the choice of nonprofit vs. for-profit form by an entrepreneur who must decide how to organize the production of a good where quality matters.

In the set-up of this model, the entrepreneur plays the role of the agent in a standard principal-agent model, and potential demanders for the good that is to be produced are the principals. The principals have a latent demand for a good whose "true" quality can be observed by both them and the agent. But, true quality may differ from *externally observed* quality that can be legally *verified* for purposes of determining, for example, whether contractual obligations have been fulfilled. A specific example of this type of good would be nursing home care. Two different nursing homes could, on one hand, be in compliance with all federal, state, and local standards of care, and hence offer "equal" externally observed quality while at the same time differing widely in how residents were actually treated on a daily basis by staff.

In this case, unlike that of the public good examined by Bilodeau and Slivinksi, what we shall term the "quality-defined good" is much more likely to be provided on a fee-for-service basis in the marketplace. In addition, the entrepreneur who is assumed to organize the production of the quality-defined good in the Glaeser-Shleifer model is "less altruistic" than the entrepreneur in the Bilodeau-Slivinski model. Although the potential nonprofit entrepreneurs in both models are assumed to be utility maximizers, in the Glaeser-Shleifer set-up, utility depends on the entrepreneur's income rather than on the amount of the good that is provided (or even its quality).

In the Glaeser-Shleifer model, once the new enterprise is set up, the entrepreneur's income equals the difference between revenue from producing and selling the new good less the costs of producing it. If the enterprise is organized as a for-profit firm, there is no nondistribution constraint, and the income realized by the entrepreneur is simply the profit from producing the good. If instead the enterprise is organized as a nonprofit organization, the nondistribution constraint prevents the entrepreneur from directly capturing this surplus. But, in this case a portion of the surplus (less than or equal to 100 percent) can be captured in the form of perks, such as compensation, staff size and quality, and so forth.

The question posed by Glaeser and Shleifer is why, when faced with a choice between organizing as a for-profit or nonprofit, an entrepreneur would ever choose the nonprofit form since the nondistribution constraint limits the portion of the surplus that could be captured? The answer is that *if* quality of the good matters, in the sense that potential demanders would be willing to pay a higher price for a good of verifiable higher quality, *and* if imposing the nondistribution constraint pro-

vides a means of signaling that the agent (entrepreneur) has less of an incentive to trade-off cost for quality, then organizing as a nonprofit has both advantages and disadvantages that must be weighed against the pluses and minuses of the for-profit form of organization.

If the enterprise is organized as a nonprofit organization, α_{NP} (the portion of the surplus S_{NP} that can be captured by the entrepreneur in the form of perquisites) is less than unity (100 percent). If instead the enterprise is organized as a for-profit organization, the entrepreneur can capture all of the surplus, S_{FP}. On the other hand, if potential demanders for the good care enough about quality so that they are willing to pay more for higher quality, and if organizing as a nonprofit organization offers a credible signal that the goods offered will be of higher quality, then the surplus generated by the nonprofit organization, S_{NP} could be greater than the surplus generated by the for-profit enterprise. Thus, choice of nonprofit form turns on whether $\alpha_{NP}S_{NP} \geq S_{FP}$, where $\alpha_{NP} \leq 1$—e.g., the nondistribution constraint limits the portion of the surplus that can be captured—but where $S_{NP} \geq S_{FP}$—potential demanders may be willing to pay more for the good provided by the nonprofit.

The Glaeser-Shleifer model offers important insights about determinants of nonprofit entrepreneurship that complement those of the Bilodeau-Slivinski model. First the model suggests that entrepreneurs will be more likely to devote their time and energies to organizing nonprofit rather than for-profit enterprises in areas where the quality of the goods or services matter *and* quality is hard to verify.[19] Second, the model reinforces and amplifies the previous point made about the link between government regulatory policies and the credibility of the nondistribution constraint. Just as the Bilodeau-Slivinski model shows that imposing the nondistribution constraint facilitates organizing the voluntary finance of public goods only if the constraint is binding and credible, so too the Glaeser-Shleifer model implies that imposing the constraint facilitates the provision of "higher quality" goods by nonprofit organizations only if it is credible. The potential entrepreneur will choose the nonprofit form only if $S_{NP} \geq S_{FP}$, which actually requires that potential demanders believe that $\alpha_{NP} < 1$, because it is only when $\alpha_{NP} < 1$ that the entrepreneur/agent has the incentive to produce goods and services of verifiable higher quality. Thus, regulations, such as those limiting inurement of nonprofit managers, that make the nondistribution constraint "hard" rather than "soft" not only make sense from the standpoint of regulatory and tax policy, they are

also necessary to encourage use of the nonprofit form as a means of producing a range of quality-dependent social goods.

Social Ventures

A variation on the general theme of utility maximization also offers insights about what might prompt some persons who combine entrepreneurial skills with a charitable impulse to choose "to do good" by founding a for-profit business rather than organizing a more traditional not-for-profit organization. Consider the case of someone who has the skills and instincts needed to be a successful for-profit entrepreneur, but who, like a more traditional nonprofit entrepreneur, is motivated by utility maximization rather than profit maximization. Assume further that in such cases the potential entrepreneur's utility depends both on the satisfaction derived from creating something new—which some students of for-profit entrepreneurship have argued is an important motivator for many for-profit entrepreneurs—*and* from a desire to do good works through charity.

Because time is a scarce commodity, such an entrepreneur would need to choose among several alternative courses of action.

(1) She could use her entrepreneurial skills to establish a new nonprofit organization and seek funding for the organization in the usual manner from grants or contributions.

(2) She could first establish a nonprofit organization, and then use her entrepreneurial talents to found a profit-making business subsidiary of the nonprofit organization whose income would support the nonprofit's activities.

(3) She could allocate her time to found a for-profit enterprise with the aim of using the expected profits to support one or more charitable activities.

(4) She could allocate her time to founding a for-profit enterprise, but then impose an additional constraint that some percentage of the enterprise's profits be devoted to charity.

One straightforward implication is that a rational entrepreneur would prefer either option 3 or 4 to option 1 or 2 if she was able to garner more financial resources to support charitable causes *by using the same skills and abilities* than would be case if she used these talents to establish a nonprofit organization that was funded from grants and

fees and charges. This insight is interesting because it offers an explanation of why many business concepts that have been identified as attractive prospects for social venturing often come from the information technology sector: software and computers, Web-based information, and so forth. Although it remains to be seen whether this sector will have the same profit-making potential in the future, clearly in the recent past, the opportunity to make substantial profits from new information technology ventures would have made founding a "socially oriented" information-technology or dotcom enterprise a potentially very attractive means of generating new resources to finance new charitable activities compared with founding a new nonprofit organization and then seeking to fund such an organization by applying for grants and contracts, and charging fees for services.

The simple utility-maximization framework also sheds light on why a socially minded entrepreneur, having chosen to use their talents to found a for-profit enterprise, might choose option 4 rather than option 3. In a spirit similar to the signaling role played by the nondistribution constraint in the case of a nonprofit organization, one might expect socially minded for-profit entrepreneurs to signal to potential buyers that a portion of the profits earned from the sale of their product is to be donated to charity.

Indeed, if the entrepreneur planned to donate to charity in any event, it would be economically irrational not to provide such a signal. At a minimum, buyers might ignore the signal (or treat it as not being credible), in which case the owner-manager of the enterprise would be no worse off for having promised to donate X percent to charity (assuming that she would have been willing to donate at least X percent of profits to charity in any event). But, in more favorable circumstances, at least some buyers might either be willing to pay somewhat more for the product (because it was seen as being of "higher quality"), or be willing to substitute the "socially responsible" product in place of other goods, thereby allowing the socially responsible enterprise to earn somewhat higher profits. Under either of favorable scenarios, the "profit contribution signal" would raise the socially responsible entrepreneur's utility, either by allowing more of the charitable good(s) to be provided for the same contribution made by the socially responsible owner-manager, or by allowing a given amount of the charitable good(s) to be provided at a lower level of the owner-manager's contribution.

Charitable Entrepreneurs and the External Environment

Whether entrepreneurial impulses of *individuals* are ultimately translated into *entry* of new nonprofit and/or socially minded *enterprises* will be affected by the external environment. In this section, we discuss how entry of traditional nonprofits is affected by changing demands for charitable goods and services, and the availability of financing for new nonprofit startups. We also describe how changes in the economy at large may be blurring the boundaries between nonprofit and for-profit organizations in ways that create new opportunities for the creation of socially minded for-profit businesses.

External Factors and Entry of Traditional Nonprofit Organizations

Although we have argued that founders of traditional nonprofit organizations are apt to be motivated more by maximization of utility than by profit, factors in the external environment that encourage or discourage entry of new nonprofit organizations are broadly similar to those that prompt entry of for-profit enterprises.

As has been noted by Kwoka and Snyder (2000), the industrial organization literature identifies several factors that seem to affect entry of for-profit enterprises. One is increased demand for a particular good or service. Empirical evidence fairly consistently shows that growth in demand spurs entry both because it increases the expected reward from entry, and because growth in demand increases the "room" in an existing market for new entrants.

Entry also depends on hurdles that new entrants face when entering a new market. These include high fixed capital costs, financing constraints, advertising, and R&D. The likely behavioral response of incumbents can be another possible entry barrier. If, for example, incumbents have the ability and/or capacity to expand or modify their operations in response to changing or growing demand, there may be less of a niche for new entrants to meet market demands.

In the case of for-profit enterprises, each of the above factors affects entry through its effect on the profit (π^*) that a potential entrant expects to earn compared with what we might term the "reservation profit" (π_R) that could be earned in alternative activity. Yet it is easy to see how the same factors as those listed above might also affect the expected utility from forming a new nonprofit organization.

Demand for Nonprofit Goods and Services and Entry

Like for-profit businesses, nonprofits often face shifting demands for their goods and services. In some cases, these demand shifts may be fairly short-lived, if relatively intense. For example, a natural disaster, such as a hurricane or earthquake, may cause immediate need for emergency services such as food, water, and shelter. Spikes in natural gas and electricity prices may necessitate the need for short-term cash assistance to allow low-income families and the elderly to pay their bills. In such cases, one might expect demand to be met by existing organizations, rather than the entry of new ones. But one would expect more permanent *increased demand* for particular charitable goods or services to encourage formation of new nonprofit organizations by raising the expected utility of nonprofit entrepreneurs above the reservation utility level. One would also expect certain *changes in the mix of demands* for charitable goods and services to encourage entry if existing organizations lacked the ability and/or the capacity to respond to these changing demands.

Nonprofit Entry in Response to Changes in Government Policy

Demand changes for nonprofit services are often likely to be prompted by broad shifts in government policy. One interesting policy question is whether nonprofits are "entrepreneurial enough" to respond to changes in demand caused, for example, by policy reforms.[20] This question becomes particularly important as interest grows in privatizing the supply of a range of public services by having such services financed by tax dollars, but provided by private organizations, which are made up largely of nonprofit organizations.

Nonprofit Response to AFDC Waivers: A Case Study

The response of nonprofits to the adoption of state AFDC waivers that presaged the enactment of federal welfare reform in 1996 presents an interesting case study. The AFDC waiver program changed the environment in which human service nonprofits operated in two ways. First, AFDC waivers signaled the reorientation of public welfare goals—from income support to economic self-sufficiency and personal responsibility—that changed the expectations of social service systems and the organizations that operate within them. Indeed, the

growing policy emphasis on economic independence of welfare recipients raised the need for local social service organizations to concentrate on job skills, child care, and transportation. Second, the broad trend toward privatization, of which the AFDC waiver program was a key manifestation, altered the manner in which many human services were provided to clients. Welfare reform involved not only a growing use of market-based policy tools such as vouchers and managed care, but also promised less bureaucratic interference in decentralized service systems. Fiscal and programmatic accountability shifted from higher levels of government to lower ones, and, in some cases under certain policy instruments, from government to the consumers.

Twombly (2000) examines how the adoption of these waivers affected the entry of nonprofit human service providers in metropolitan areas between 1992 and 1996, which was the period of widespread implementation of waivers in the states. He finds that entry of new nonprofit human service providers tended to be higher in areas in which AFDC waivers were adopted than in areas without waivers, after controlling for other factors.[21]

Twombly's analysis also shows that entry is affected by factors in addition to changes in demand, such as the extent to which the philanthropic community—donors, foundations, trusts—supports human service nonprofits. In particular, nonprofit entry is apt to be greater in metropolitan areas that have a cultural predisposition to rely heavily on organized charities, as distinguished from family or noninstitutional arrangements, to supply social services.

Twombly also finds that organizational density—or the ratio of nonprofit providers to people in poverty in a metropolitan area—is a key predictor of entry by nonprofit human service groups. The density of organizations can have a significant effect on nonprofit entrepreneurial activity because the degree of competition among groups can affect access to funding, clients, capital, and the perceived legitimacy of new entrants into existing social service networks. Holding other factors statistically constant, urban areas with dense (or crowded) nonprofit human service sectors were found to experience significantly slower rates of nonprofit entry than other metro regions.[22]

These results offer several insights about how nonprofit entrepreneurial activity can be shaped by the external environment. One is that a public policy change that shifts demand for nonprofit goods and services can have an important effect on the formation of nonprofit organizations. Adopting AFDC waivers appears to have encouraged

entry of urban nonprofit providers by shifting the thrust of welfare policy from entitlements to labor market entry.

The analysis also demonstrates that factors other than demand are also important. For example, metropolitan regions that had most heavily embraced social service provision by nonprofit organizations with respect to other methods proved highly hospitable to new entrepreneurial ventures, irrespective of the level of waiver experimentation.[23] Conversely, the presence of strong and active existing organizations in an urban area served as a brake on nonprofit entry.

Financing of New Charitable Organizations

Entry is affected not only by opportunities arising from shifts in the pattern of demand, but also by the ease with which potential entrepreneurs can obtain the financing needed to found new enterprises. Just as those who study the formation of new businesses have expressed concern about whether such enterprises have access to capital and financing, so too those who would study the formation of new nonprofit organizations need to pay heed to how such startup nonprofits are financed.

In the case of traditional nonprofits, Hansmann (1987) notes that a direct consequence of imposing the nondistribution constraint is that nonprofit organizations thereby forego access to equity capital as a source of finance. This leaves bank or other debt, grants and contributions, and revenue from providing charitable goods and services as the main venues for financing new nonprofits. Since organized lenders will be less apt to make loans to startup nonprofits, in practice, the seed-money for new nonprofit ventures seems most likely to come from private contributions and grants, and from fees for services.[24]

Grants, Contributions, and Foundations

Figure 1 shows the percentage of revenue derived from different sources by new and established nonprofits; figure 2 shows the revenue mix among new nonprofits that survive and those that fail. Figure 1 shows that although all nonprofit organizations depend on both public contributions and fees for services as their main source of revenue, startup nonprofits depend more heavily on grants and contributions as a revenue source than do their more established counterparts.

Broadly speaking, one might think of grants and contributions as playing a similar role in providing "seed money" for emerging

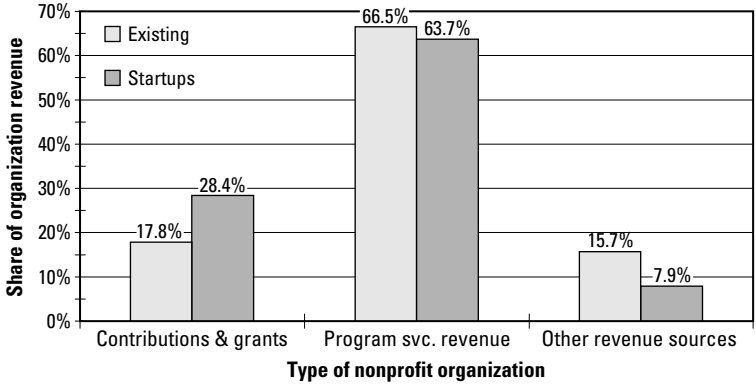


Figure 1
Revenue sources of new entrants and existing nonprofits.

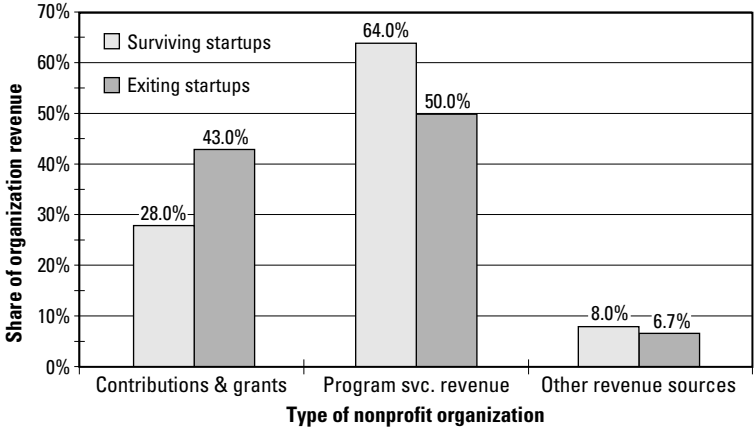


Figure 2
Revenue sources of surviving and deceased entrants.

nonprofit organizations as venture capital in providing a portion of the initial capitalization of new for-profit businesses. Surprisingly little, however, is known about the extent to which traditional sources of philanthropy, such as foundations, are willing to fund "risky" nonprofit ventures in the same way that venture capitalists are willing to fund risky for-profit startups.

The limited evidence that is available suggests that foundations do not play the same systematic role in helping to "grow" new nonprofits that venture capitalists do in the case of for-profit startups. Gronbjerg

et al. (2000) report that many family foundations use "review structures that create few and, in some cases, no opportunities for new agencies to come to the founder's attention." Letts et al. (1997), suggest that foundations that support new nonprofits could draw many lessons from venture capitalists.[25]

Fees and Charges

Figure 2 shows that new nonprofits that are "successful" tend to garner more of their revenue from fees and charges for services provided and less from grants and contributions than do nonprofit startups that fail. These results provide one clue for some of the sectoral variation in exit rates among new nonprofits. As may seen from table 3, exit rates among new nonprofits tend to be lower among those engaged in health and social service activities, broadly defined, which often depend on fees charged to clients as well as on payments received from government social service programs.

The fact that new entrants are more likely to survive when they derive more of their revenue from fees and charges than from grants and contributions has some interesting implications. One is that successful entry of new nonprofits may be more likely to occur in cases where the charitable good or service has more of the characteristics of a "private" rather than a "public" good. Like new for-profit enterprises, the ability of new nonprofits to survive will depend in part on their ability to generate revenue from "sales" of their output. As a result, entry of new organizations is more likely to be a potential source of change in nonprofit sectors where it is easier for charities to finance their activities from fees and charges instead of grants and contributions.

But, as Smith and Lipsky (1993) note, the ability to finance activities from fees and charges can be a two-edge sword. Fees and charges may offer relatively secure sources of finance. Yet dependence on these forms of revenue can also temper innovation. For example, receipt of government contracts may place a nonprofit in a dependency relationship on the agency that becomes its "most secure" customer.[26]

Financing For-Profit Social Ventures

To date, there is no systematic information on sources of finance for new, charitably minded social ventures. On the one hand, to the extent that such ventures have actual owners—e.g., do not impose the nondistribution constraint—for-profit startups with a charitable mission

may enjoy access to private capital markets not enjoyed by traditional nonprofits. Indeed, some established nonprofits have been encouraged in recent years to create for-profit subsidiaries as a means of gaining access to equity capital markets (see Steuerle 2001).

On the other hand, a for-profit charitable venture may forego access to more traditional sources of philanthropic support, which tend to be reserved for 501(c)(3) organizations. It is also an open question whether a for-profit startup that commits to a social mission as well as to the traditional bottom line can compete as effectively in traditional capital markets as a solely profit-oriented startup.

Blurring the Lines: The "New" Economy and Social Venturing

We conclude our discussion of external factors by examining how the changing shape of the general economy may facilitate a shift in the mix of social venturing and "traditional" nonprofit entrepreneurship. The basic observation is that wealthier societies increasingly demand more services, such as health care, that have traditionally been provided by nonprofit organizations.[27] A consequence is that the current economy produces more products that could fit either under a charitable or 501c(3) definition, or in some cases could be produced and sold for profit. For example, health care used to be 1/30 of the economy, now it is 1/6. "Information" firms often sell material that is produced using a production function that may be quite similar to that used to provide nonprofit services that can be sold for profit or not. The data in tables 4 and 5 provide statistical evidence of some of these trends; table 4 shows that industries that produce "potentially charitable outputs"—which we define as goods or services that could be provided by traditional nonprofit organizations—have grown at faster rates than other sectors, and are projected to continue doing so. These trends are mirrored by growth rates in occupation, shown in table 5, where occupations with significant nonprofit penetration have grown by 67 percent between 1988 and 1999 compared with growth rates of 25 percent in other occupations.

Thus the "new economy" may be somewhat more conducive to creating a pool of potential entrepreneurs whose skills could be used to found either for-profit businesses or nonprofit organizations. As has already been noted, many of the skills that are needed to successfully found and manage new information enterprises may be similar to

those that are needed to organize and manage a variety of activities in the nonprofit sector.

Compare a world where the typical industry is based in manufacturing, such as steel, to one where it is based on information goods, such as computer services. In the former environment, the owner or manager of a steel company lives a divided life, making steel during the day and then going home and putting on his civic (e.g., nonprofit) hat. The inputs needed to make steel (including labor input) and to provide charitable services and the outputs of the manufacturing and the nonprofit activity are also apt to be quite different.

This situation may be contrasted with the case of the firm producing information goods and services. Because forms of intellectual capital figure heavily as both inputs and outputs for such enterprises, the inputs and outputs in the for-profit sector may be quite similar to those in the not-for-profit activity. Marketing and computer skills are also more likely to cut across sectors. Thus, changes in the economy at large—many of which are driven by conventional forms of for-profit entrepreneurship—may be creating an environment in which one may increasingly see nonprofit entrepreneurs choosing to found for-profit ventures as either the entire, or an integral part of their overall strategy for meeting social needs.

Conclusions

In the 1960s there was a saying that "military justice is to civilian justice, as military music is to civilian music." Our exploratory essay on the dimensions of nonprofit entrepreneurship suggests that this sobriquet clearly does not apply to the relationship between nonprofit and for-profit entrepreneurship.

Data on entry and exit of organizations in the nonprofit sector paint a picture of organizational change that is at least as rich and varied as that found in the for-profit sector of the economy. Moreover, although decisions about nonprofit entry and exit are more usefully modeled as being motivated by utility rather than by profit maximization, many factors, such as changes in demand, the presence of incumbent providers, and the presence of financing constraints, play similar roles in explaining broad patterns of entry and exit in the case of nonprofit organizations as they do in models of entry and exit by for-profit businesses.

Table 4
Relative growth of industries with potentially "charitable" output. Adapted from Thomson 1999. N.B.: Subcategories of industries sum to grand total because of chain weighting. (See Thomson 1999.)

		Jobs (1,000)			Projected change		Avg. annual growth rate,	Projected,
	SIC	1988	1998	2008	1988–1998	1998–2008	1988–1998	1998–2008
Charitable industries								
Computer and data processing services	737	673	1,599	3,472	926	1873	13.8	11.7
Management and public relations services	874	508	1,034	1,500	526	466	10.4	4.5
Social services	83	1,552	2,644	3,678	1,092	1,034	7	3.9
Museums, botanical and zoological gardens	84	58	93	131	35	38	6	4.1
Producers, orchestras, and entertainers	792	122	176	225	54	49	4.4	2.8
Educational services	82	1,567	2,177	2,690	610	513	3.9	2.4
Health services	80	7,106	9,846	12,667	2,740	2,821	3.9	2.9
Commercial sports	794	91	127	160	36	33	4	2.6
Membership organizations	86	1,740	2,361	2,600	621	239	3.6	1
Research and testing services	873	492	614	861	122	247	2.5	4
Advertising	731	229	268	323	39	55	1.7	2.1
Communications, including telephone, television, radio	48	1,280	1,470	1,768	190	298	1.5	2
Legal services	81	845	943	1,200	98	257	1.2	2.7
Printing and publishing	27	1,543	1,565	1,545	22	−20	0.1	−0.1

Manufacturing of medical equipment and technology-related instruments	38	1,031	868	887	−163	19	−1.6	0.2
Manufacturing of computers and computer systems	357	459	379	369	−80	−10	−1.7	−0.3
Subtotal		19,296	26,164	34,076	6,868	7,912	3.6	3
Other industries								
Other services	70–79, 84–87, 89; not 731, 737, 792, 794	9,883	14,704	19,036	4,821	4,332	4.9	2.9
Transportation and utilities	40–42, 44–47, 49	4,232	5,130	5,773	898	643	2.1	1.3
Construction	15, 16, 17	5,098	5,985	6,535	887	550	1.7	0.9
Retail trade	52–59	19,023	22,296	25,363	3,273	3,067	1.7	1.4
Government	—	17,386	19,819	21,688	2,433	1,869	1.4	0.9
Wholesale trade	50, 51	6,030	6,831	7,330	801	499	1.3	0.7
Finance, insurance and real estate	60–67	6,629	7,408	8,367	779	959	1.2	1.3
Agriculture	01, 02, 07, 08, 09	3,355	3,576	3,526	221	−50	0.7	−0.1
Nonagricultural self-employed and unpaid family	—	8,731	9,029	9,925	298	896	0.3	1
Other manufacturing	20–39; not 27, 247, 38	16,281	15,960	15,883	−321	−77	−0.2	0
Secondary job as a self-employed or unpaid family worker	—	1,990	1,897	1,901	−93	4	−0.5	0
Mining	10–14	713	590	475	−123	−115	−1.7	−1.9
Secondary wage and salary jobs in agriculture (except agricultural services); forestry, fishing, hunting, and trapping; and private households	—	211	163	158	−48	−5	−2.3	−0.3
Subtotal		99,562	113,388	125,960	13,826	12,572	1.4	1.1
Grand total		120,010	140,514	160,795	20,503	20,281	1.7	1.4

Table 5
Change and growth in "charitable" and "noncharitable" occupations. Sources: 1983 and 1993 figures adopted from Statistical Abstract of the United States, 1994, table 637; 1999 figures adopted from Statistical Abstract of the United States, 2000, table 669.

	Category	1983 (1,000)	1999 (1,000)	Change 1983–1999	% change 1983–1999
Occupations with Significant Nonprofit Penetration					
Managers: medicine and health	Health	91	716	625	686.8
Computer scientists	Information/Technology	276	1,549	1,273	461.2
Teachers' aides and early childhood education	Education	348	1,198	850	244.3
Recreation and related workers	Social Service	131	270	139	106.1
Social workers	Social Service	407	813	406	99.8
Adminstrators: education and related fields	Education	415	821	406	97.8
Athletes	Entertainment	58	110	52	89.7
Child care workers	Social Service	408	764	356	87.3
Managers: marketing and public relations	Financial	396	739	343	86.6
Social scientists	Professional—Other	261	460	199	76.2
Health assessment and treatment (therapists, pharmacists)	Health	528	891	363	68.8
Natural scientists	Science	357	578	221	61.9
Teachers: higher education	Education	606	978	372	61.4
Mathematics scientists (not computer)	Information/Technology	187	298	111	59.4
Writers, authors and entertainers	Entertainment	1,486	2,344	858	57.7
Teachers: pre-k through 12	Education	3,365	5,277	1,912	56.8
Registered nurses	Health	1,372	2,128	756	55.1
Health technologies	Health	1,111	1,701	590	53.1
Lawyers and judges	Law	651	964	313	48.1

Protective services (police, fire)	Protective Services (police, fire)	1,672	2,440	768	45.9
Health diagnosing occupations (physicians, dentists)	Health	735	1,071	336	45.7
Health services (not doctors or nurses)	Health	1,739	2,521	782	45
Librarians	Education	193	264	71	36.8
Welfare aids	Social Service	77	97	20	26
Clergy	Social Service	293	352	59	20.1
Education counselors	Education	213	247	34	16
Computer equipment operators	Administrative Support	605	356	−249	−41.2
Subtotal		17,981	29,947	11,966	66.5
Other Occupations					
Adjustors and investigators	Administrative Support	675	1,802	1,127	167
Data entry	Administrative Support	311	746	435	139.9
Financial managers	Financial	357	753	396	110.9
Architects	Professional—Other	103	194	91	88.3
Information clerks, including receptionists	Administrative Support	1,174	2,143	969	82.5
Executive, administrative and managerial (other)	Professional—Other	8,408	14,897	6,489	77.2
Accountants and auditors	Financial	1,105	1,658	553	50
Sales—other including supervisors and proprietors	Sales	6,331	9,267	2,936	46.4
Other technicians (engineering, science)	Other	1,942	2,654	712	36.7
Construction	Production	4,289	5,801	1,512	35.3
Engineers	Professional—Other	1,572	2,081	509	32.4
Transportation and material moving occupations	Operators, Fabricators and Laborers	4,201	5,516	1,315	31.3
Other administrative support	Administrative Support	1,642	2,087	445	27.1
Handlers and laborers	Operators, Fabricators and Laborers	4,147	5,265	1,118	27
Material recording, scheduling (dispatchers)	Administrative Support	1,562	1,959	397	25.4
Food workers	Personal Services—Other	4,860	6,091	1,231	25.3

Table 5
(continued)

	Category	1983 (1,000)	1999 (1,000)	Change 1983–1999	% change 1983–1999
Sales workers—retail and personal services	Sales	5,511	6,866	1,355	24.6
Mail and message distribution occupations	Administrative Support	799	990	191	23.9
Records processing	Administrative Support	866	1,047	181	20.9
Mechanics	Production	4,158	4,868	710	17.1
Cleaning services	Personal Services—Other	2,736	3,021	285	10.4
Other personal services	Personal Services—Other	1,793	1,936	143	8
Precision production occupations	Production	3,685	3,793	108	2.9
Supervisors—administrative support	Administrative Support	676	675	−1	−0.1
Statistical clerks	Administrative Support	96	94	−2	−2.1
Machine operators	Operators, Fabricators and Laborers	7,744	7,386	−358	−4.6
Other private household care workers	Personal Services—Other	572	536	−36	−6.3
Mail clerks	Administrative Support	68	63	−5	−7.4
Farming	Farming	3,700	3,426	−274	−7.4
Financial recording keeping (book keeping, record clerks)	Administrative Support	2,457	2,181	−276	−11.2
Secretaries	Administrative Support	4,861	3,457	−1,404	−28.9
Extractive occupations	Production	196	130	−66	−33.7
Communications operators (i.e., telephone operators)	Administrative Support	256	158	−98	−38.3
Subtotal		82,853	103,541	20,688	25
Total		100,834	133,488	32,654	32.4

Our analysis also suggests that public policy can affect nonprofit entrepreneurship in several ways. Government regulation of nonprofit organizations is important not simply as a means of preventing fraud and/or abuse of the nonprofit tax exemption, but also because it maintains the credibility of the nondistribution constraint, which figures prominently in the decision of entrepreneurs to choose the nonprofit form as a means of organizing the provision of charitable goods and services. Government policies that affect the demand for the goods and services provided by nonprofit organizations are also likely to encourage entry of new providers when these changes either increase or significantly alter the demand for a variety of charitable goods and services. Although limited, there is also empirical evidence that nonprofit entry decisions are responsive to such changes in public demands for social service, which is good news to those who seek to give the nonprofit sector a larger role in the provision of social services. Lastly, just as the behavior of venture capitalists and policies that affect venture capitalists are likely to affect the pool of seed money available to for-profit startups, so too the behavior of both private donors and public and private grants-making bodies are likely to be of somewhat greater importance to new nonprofits that depend on public contributions.

Notes

1. The estimate of the economic importance of nonprofit organizations is from p. 77 of Steuerle and Hodgkinson 1999. The definition of an entrepreneur is from the 1983 edition of *Webster's New Collegiate Dictionary*.

2. Much of this literature focuses on issues such as the increased reliance of nonprofit organizations on commercial sources of income. See e.g. Brody 1996; Brody and Cordes 2001; Hammack and Young 1993; James 1983; Steuerle 2001; Tuckman 1998; Weisbrod 1998.

3. In what is widely regarded as the definitive treatise on the economics of the nonprofit sector, Weisbrod (1988, p. 81) observes that whether one should "care whether growth in the nonprofit sector comes through growth of existing organizations rather than through entry of new organizations" is a "dimension of public policy that has escaped attention."

4. Throughout the chapter we use the term "traditional nonprofit" to refer to the classic form of public charity organized as a 501(c)(3) organization, and the term "socially oriented for-profit venture" to refer to for-profit businesses that have an explicit charitable purpose.

5. For a detailed description of these data, see http://nccs.urban.org/. See also National Center on Charitable Statistics 1999.

6. See Brody and Cordes 1999, p. 142.

7. For a description of the procedures that are required in order to be recognized as an 501(c)(3) organization, see Internal Revenue Service Package 1998.

8. Operating charities refer to nonprofit organizations that are eligible to receive tax deductible contributions and are classified as actually providing a nonprofit service, as distinguished from nonprofit organizations whose purpose is to support other operating charities. Unless otherwise noted, references to charities in this chapter refer to operating charities.

9. See Economic and Statistics Administration 1999. It is important to note that the 1997 totals do not include financial, insurance and real estate industries and auxiliary businesses. Railroad transportation and US Postal Service industries were also outside the scope of the 1997 Economic Census.

10. In some cases, a new organization may be founded by an existing organization instead of an individual or a group of individuals.

11. The entry data presented in table 1 define entry as the date at which a charity is officially recognized as a 501(c)(3) organization by the IRS through a letter ruling. Many of these organizations were already operating in some form before receiving such formal recognition.

12. The NTEE is a mixed notation, organizational classification system of 26 major groups, collapsible into ten major categories, and divisible by 645 subgroups (Stevenson et al. 1997).

13. See IRS Package 1023 1998.

14. For a full description of the range of services provided by Pioneer Human Services, see http://www.pioneerhumanserv.com.

15. See e.g. Internet Nonprofit Center 2001.

16. Comments from Gregory Dees as reported in Young 1999. More generally, see Dees et al. 1999.

17. As several discussants at the conference noted, the distinction between utility-maximization and profit-maximization is somewhat blurred because pursuit of profit can rightly be seen as a form of utility maximization in which utility depends only on the amount of income earned. Thus, more precisely, one could think of nonprofit entrepreneurs as utility maximizers whose utility functions include broader elements than income or profit, while for-profit entrepreneurs are utility-maximizers whose utility depends primarily on profit.

18. This insight was suggested by Gregory Dees (Young 1999).

19. Tables 4 and 5 document the growth of industries that produce these kinds of goods relative to the rest of the economy.

20. See Chambre 1995 for one analysis.

21. More specifically, when waivers were classified as having a minimal, moderate, or extensive degree of impact on social service systems, limited and moderate policy reforms under AFDC waivers were found to spur the entry of nonprofit providers in metropolitan areas, after controlling for other factors. Rates of entry of nonprofit human

service providers were also somewhat higher in urban regions that initiated extensive reform efforts than in non-reform regions, but these differences were not statistically significant.

22. Other environmental factors, however, had only marginal impact on the entry of new human service nonprofits during the period of waiver implementation. For example, when holding constant other factors, social welfare expenditures by local governments, total population, and the concentration of resources in the largest nonprofit providers in metropolitan areas are not significantly related to the entry of nonprofit providers. Even regional economic need, as measured by the rate of poverty, does not significantly affect the formation of human service organizations.

23. That rates of entry of nonprofit human service providers were not significantly different in urban regions that initiated extensive reform efforts than in non-reform regions also points to the important role of the local political and social culture on nonprofit entrepreneurship. Nearly half of the urban areas that introduced extensive welfare reforms already had charities that tend to work more independently of government than in other cultural settings. Over time, social service regimes in these metropolitan regions became highly institutionalized with relatively small cadres of larger and older groups providing the bulk of human services. Thus, despite the cues for entrepreneurial activity provided by welfare reform initiatives, nonprofit entrepreneurial activity may have been limited in these regions because there was less need.

24. For a good summary of elements in the "nonprofit capital market," see pp. 254–260 of Roberts Foundation 2000.

25. For a general discussion of the behavior of foundations, see Boris 1987.

26. On the relationship between nonprofits and governments that is less critical than Smith and Lipsky, see Salamon 1995.

27. As was noted above, these goods are also likely to have a significant and often unverifiable quality dimension.

References

Bilodeau, M., and A. Slivinski. 1998. Rational nonprofit entrepreneurship. *Journal of Economics and Management Strategy* 7, no. 4: 551–571.

Boris, E. T. 1987. Creation and growth: A survey of private foundations. In E. Boris and T. Odendahl, eds., *America's Wealthy and the Future of Foundations*. Foundation Center.

Boris, E. T., and Steuerle, C. E., eds. 1999. *Nonprofits and Government: Collaboration and Conflict.* Urban Institute Press.

Bowen, W. G., T. I. Nygren, S. E. Turner, E. A. Duffy, and J. K. Smith Jr. 1994. *The Charitable Nonprofits: An Analysis of Institutional Dynamics and Characteristics*. Jossey-Bass.

Brody, E. 1996. Agents without principals: The economic convergence of the nonprofit and for-profit organizational forms. *New York Law School Law Review* 40, no. 3: 457–536.

Brody, E., and J. J. Cordes. 1999. Tax treatment of nonprofit organizations: A two-edged sword? In E. Boris and C. Steuerle, eds., *Nonprofits and Government*. Urban Institute Press.

Brody, E., and J. J. Cordes. 2001. The Unrelated Business Income Tax: All Bark and No Bite? Policy Brief 4, Emerging Issues in Philanthropy, Urban Institute.

Buttenheim, A. 1998. Social enterprise meets venture philanthropy: A powerful combination. *Los Angeles Business Journal* 20, no. 46: 72–76.

Chambre, S. 1995. Creating new nonprofit organizations as response to social change: HIV/AIDS organizations in New York City. *Policy Studies Review* 14, no. 1–2: 117–126.

Dees, G., J. Emerson, and P. Economy. 2001. *Enterprising Nonprofits: A Toolkit for Social Entrepreneurs*. Wiley.

Emerson, J., and F. Twerksy, eds. 1996. *New Social Entrepreneurs: The Success, Challenges and Lessons of Nonprofit Enterprise Creation*. Roberts Enterprise Development Foundation.

Glaeser, E., and A. Shleifer. 1998. Not for Profit Entrepreneurs. Working paper W6810, National Bureau of Economic Research.

Gronbjerg, K. A. 1994. Using NTEE to classify non-profit organizations: an assessment of human service and regional applications. *Voluntas* 5, no. 3: 301–328.

Hammack, D., and D. Young, eds. 1993. *Nonprofit Organizations in a Market Economy: Understanding New Roles, Issues and Trends*. Jossey-Bass.

Hansmann, H. 1987. Economic theories of nonprofit organization. In W. Powell, ed., *The Nonprofit Sector*. Yale University Press.

Hodgkinson, V. A., M. S. Weitzman, J. A. Abrahams, E. A. Crutchfield, and D. R. Stevenson. 1996. *Nonprofit Almanac 1996–1997: Dimensions of the Independent Sector*. Jossey-Bass.

Internal Revenue Service. 1998. Application for Recognition of Exemption under Section 501(c)(3) of the Internal Revenue Code, Package 1023.

Internet Nonprofit Center. 2003. What Is Nonprofit Social Entrepreneurship?

James, E. 1983. How Nonprofits Grow: A Model. *Journal of Policy Analysis and Management* 2, no. 3: 350–366.

Kwoka, J., and C. Snyder. 1999. Entry, Growth, and Exit in the Higher Education Industry: An Exploratory Analysis. Unpublished manuscript, George Washington University.

Letts, C., W. Ryan, and A. Grossman. 1997. Virtuous capital: What foundations can learn from venture capitalists. *Harvard Business Review*, March-April: 36–43.

National Center for Charitable Statistics. 1999. *Guide to Using NCCS Data*. Urban Institute.

Salamon, L. 1995. *Partners in Public Service: Government Nonprofit Relations in the Modern Welfare State*. Johns Hopkins University Press.

Salamon, L. M., and H. K. Anheier. 1992. In search of the nonprofit sector II: The problem of classification. *Voluntas* 3, no. 3: 267–309.

Schiff, J. 1986. *Expansion, Entry and Exit in the Nonprofit Sector: The Long and Short Run of It*. Working paper 111, Program on Nonprofit Organizations, Yale University.

Schiff, J., and B. Weisbrod. 1993. Competition between for-profit and nonprofit organizations in commercial markets. In A. Ben-Ner and B. Gui, eds., *The Nonprofit Sector in the Mixed Economy*. University of Michigan Press.

Smith, David Horton. 1997. The rest of the non-profit sector: Grassroots organizations as the dark matter ignored in prevailing 'flat Earth' maps of the sector. *Nonprofit and Voluntary Sector Quarterly* 26, no. 2: 114–132.

Smith, S. R., and M. Lipsky. 1993. *Nonprofits for Hire: The Welfare State in an Age of Contracting*. Harvard University Press.

Steuerle, C. E. 2001. When Nonprofits Conduct Exempt Activities as Taxable Enterprises. Policy Brief 4, Emerging Issues in Philanthropy, Urban Institute.

Steuerle, C. E., and V. Hodgkinson. 1999. Meeting social needs: Comparing the resources of the independent sector and government. In E. Boris and C. Steuerle, eds., *Nonprofits and Government*. Urban Institute Press.

Streisand, B. 2001. The new philanthropy. *U.S. News and World Report*, June 11: 40–42.

Tuckman, H. P. 1998. Competition, commercialization, and the evolution of nonprofit organizational structures. In B. Weisbrod, ed., *To Profit or Not to Profit*. Cambridge University Press.

Twombly, Eric C. 1990. Organizational Response in an Era of Welfare Reform: Exit and Entry Patterns of Nonprofit Human Service Providers. Ph.D. dissertation, George Washington University.

Weisbrod, B. A. 1977. *The Voluntary Nonprofit Sector: An Economic Analysis*. Lexington.

Weisbrod, B. A. 1988. *The Nonprofit Economy*. Harvard University Press.

Weisbrod, B. A., ed. 1998. *To Profit or Not to Profit: The Commercial Transformation of the Nonprofit Sector*. Cambridge University Press.

Young, D. 1999. Economic Decisionmaking by Nonprofit Organizations in a Market Economy: Tensions between Mission and Market. National Center on Nonprofit Enterprise.

6

Does Business Ownership Provide a Source of Upward Mobility for Blacks and Hispanics?

Robert W. Fairlie

The differences between African-American and Hispanic self-employment rates and the white self-employment rate are striking. Approximately 11.6 percent of white workers are self-employed, whereas only 3.8 percent of black workers and 6.8 percent of Hispanic workers are self-employed (U.S. Bureau of the Census 1993). Furthermore, the 3:1 ratio of white to black self-employment rates has remained roughly constant over the past 80 years (Fairlie and Meyer 2000). Of the blacks and Hispanics who are self-employed, their businesses have lower revenues and profits, hire fewer employees, and are more likely to fail than white-owned businesses (U.S. Bureau of the Census 1997).

The relative absence of black- and Hispanic-owned businesses in the United States is a major concern among policy makers. It is particularly troubling because it has been argued that self-employment provides a route out of poverty and an alternative to unemployment or discrimination in the labor market.[1] For example, Glazer and Moynihan (1970, p. 36) argue that "business is in America the most effective form of social mobility for those who meet prejudice." Proponents also note that many disadvantaged groups facing discrimination or blocked opportunities in the wage/salary sector have used business ownership as a source of economic advancement. It has been argued, for example, that the economic success of earlier immigrant groups in the United States, such as the Chinese, Japanese, Jews, Italians, and Greeks, is due in part to their ownership of small businesses. (See Loewen 1971; Light 1972; Baron et al. 1975; Bonacich and Modell 1980.) More recently, Koreans have purportedly used business ownership for economic mobility (Min 1989, 1993).

Although a rapidly growing literature documents and examines the causes of ethnic and racial differences in rates of business ownership

in the United States, there is very little empirical evidence from longitudinal data on the relationship between business ownership and economic mobility for disadvantaged minorities.[2] An important question is whether business ownership provides a route for economic advancement for at least the relatively few blacks and Hispanics who are self-employed. To my knowledge, only two previous studies provide evidence from long-term panel data on this question. First, in previous research I use data for 1982–1996 from the National Longitudinal Survey of Youth (NLSY) to examine the earnings patterns of young less-educated business owners and make comparisons to young less-educated wage/salary workers (Fairlie 2000). Using fixed-effects earnings regressions, I find that the self-employed experience faster earnings growth on average than wage/salary workers after a few initial years of slower growth suggesting that, for at least some less-educated youths, business ownership provides a route for economic advancement. Second, Holtz-Eakin, Rosen, and Weathers (2000) examine one-year and five-year mobility rates in the income distribution for prime-age self-employed and wage/salary workers using data from the 1968–1990 waves of the Panel Study of Income Dynamics. They find that low-income self-employed workers experienced more upward mobility in the income distribution than did low-income wage/salary workers. Furthermore, they find some evidence that self-employment was more successful for blacks than non-blacks.

In this chapter, I examine the earnings patterns of young black and Hispanic business owners. Using data from the National Longitudinal Survey of Youth (NLSY), this study is the first to examine the long-term earnings patterns (1979–1998) of young self-employed blacks and Hispanics. To place these earnings patterns into context, I make comparisons to young black and Hispanic wage/salary workers and to young white self-employed and wage/salary workers. The key question is whether black and Hispanic youths who are self-employed early in their careers experience faster earnings growth than their counterparts employed in the wage/salary sector. I do not specifically model the selection process into self-employment, and thus cannot infer from these results whether self-employment is a "better" option for the randomly chosen black or Hispanic. Although this is an important question, no credible identifying instruments exist.[3] Nevertheless, the following analysis of earnings patterns may shed light on the potential for self-employment to provide a source of economic mobility and self-sufficiency for at least some blacks and Hispanics.

Data

I use data from the National Longitudinal Survey of Youth (NLSY), a nationally representative sample of 12,686 men and women who were between the ages of 14 and 22 when they were first interviewed in 1979.[4] Survey members were interviewed annually from 1979 to 1994, then in 1996, then in 1998. I exclude the sample of 1,280 youths designed to represent the population who were enlisted in the four branches of the military as of September 30, 1978, and the supplemental sample of 1,643 economically disadvantaged non-black, non-Hispanic youths. The resulting sample contains random samples of black, Hispanic, and non-black, non-Hispanic youths (referred to as whites).

Self-employed workers are defined as those individuals who identify themselves as self-employed in own business, professional practice, or farm on the class of worker question for the current or most recent job.[5] I remove individuals who report being enrolled in school and workers who report working fewer than 300 hours in the previous calendar year. The hours restriction rules out very small-scale business activities. In all annual earnings comparisons, I focus on workers reporting at least 1,400 hours in the past calendar year.

Total annual earnings are calculated by summing the responses to questions on military income, wage and salary income, and business or farm income (after expenses) in the past calendar year. I add the income from all three sources because 56.9 percent of the self-employed with positive earnings in my sample report wage and salary income but do not report business income. This is only partly due to incorporated business owners reporting their income as wage and salary income—55.3 percent of unincorporated business owners with positive total earnings report zero business income. As suggested by Jay Zagorsky at the Center for Human Resource Research, Ohio State University, it may partly be due to the ordering of questions on the questionnaire. Respondents were asked the following questions: (1) How much money did you get from the military? (2) Excluding military pay, how much money did you get from wages, salary, commissions or tips? (3) Excluding anything you already mentioned did you receive any business income? Thus, some of the self-employed may have reported their income in the second question and did not correct their mistake. Another possibility is that the self-employed report only their labor income from the business under wage/salary income. I explore this issue further below.

Table 1
Self-employment rates by race (NLSY 1979–1998). Sample consists of youths who worked at least 300 hours in the survey year. "Whites" includes all non-black, non-Hispanic individuals.

	Men		Women	
	SE Rate	*N*	SE Rate	*N*
Blacks	4.8%	12,682	2.6%	11,623
Hispanics	6.9%	8,957	4.6%	7,282
Whites	9.6%	24,207	6.6%	21,602

Earnings observations in all years are inflated to 1998 dollars. The responses for each of these three sources of income are top coded at $75,000 from 1979 to 1984, $100,000 from 1985 to 1994, and the top 2 percent for 1996 and 1998. Instead of using these top codes, I impose the 1994 top code in 1998 dollars for all years, which equals $109,987. I set all top-coded values to $150,000.[6]

Self-Employment Rates and Earnings Comparisons

Before examining self-employment and wage/salary earnings patterns, I present some descriptive results on self-employment rates and total earnings comparisons by race. Table 1 reports self-employment rates by sex and race. The self-employment rate is defined as the fraction of workers who are self-employed. The reported estimates indicate that self-employment rates differ substantially by race.[7] Similar to estimates reported in previous studies, blacks and Hispanics are much less likely to be self-employed than are whites. Only 4.8 percent of black men are self-employed compared to 9.6 percent of white men. The Hispanic male rate of 6.9 percent is also lower than the white rate, but higher than the black rate. Among women, the black/white and Hispanic/white self-employment rate ratios are similar to those for men. The main difference, however, is that for all racial groups female self-employment rates are lower.

These estimates from the NLSY are comparable to those from 1990 Census microdata using a similar age group (reported in appendix table 1). I generally find slightly lower rates using the Census, but the relative differences between the races are similar. Blacks and Hispanics are substantially less likely to own businesses than are whites.

Although relatively few blacks and Hispanics are self-employed it is important to determine whether these minority business owners are

Table 2
Self-employment and wage/salary earnings, full-time workers (NLSY 1979–1998). Sample consists of youths who worked at least 1,400 hours in survey year. "White" includes all non-black, non-Hispanic individuals. SD: standard deviation.

	Men		Women	
	Self-employed	Wage/salary	Self-employed	Wage/salary
Blacks				
Mean	$31,280	$24,461	$20,584	$20,168
Median	$22,261	$21,523	$14,916	$18,002
SD	$29,486	$16,268	$25,557	$11,998
Sample size	410	9,476	178	8,179
Hispanics				
Mean	$38,678	$27,697	$24,702	$21,660
Median	$26,344	$24,801	$17,899	$19,693
SD	$41,167	$17,225	$33,819	$12,674
Sample size	470	7,001	158	5,121
Whites				
Mean	$46,952	$33,663	$24,509	$24,088
Median	$33,002	$29,534	$17,912	$20,912
SD	$46,102	$22,290	$26,534	$15,625
Sample size	2,028	19,141	835	14,898

successful. In table 2, I report the mean, median, and standard deviation of total annual earnings for black and Hispanic self-employed and wage/salary youths. I only include full-time workers, defined here as working at least 1,400 hours in the past calendar year, to control for differences in hours worked. I first discuss the results for men. For both black and Hispanic men, the self-employed earn substantially more on average than do wage/salary workers. Self-employed blacks and Hispanics earn $6,819 and $10,981 more than their wage/salary counterparts, respectively.[8] These differences are large, representing 30–40 percent of average wage/salary earnings. A comparison of means can create a distorted picture, however, if a few business owners are extremely successful.[9] Comparing median income levels removes these concerns. For both blacks and Hispanics, median self-employment earnings are still higher than median wage/salary earnings; however, the differences are much smaller.

Although average and median earnings are higher for self-employed blacks and Hispanics, it is important to also compare the variance of earnings in the two sectors. For both races, the standard deviation of self-employment income is substantially higher than that of wage/

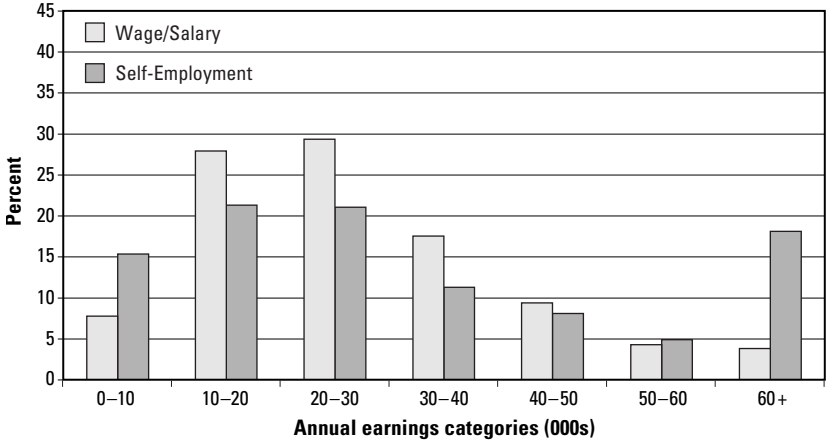


Figure 1
Earnings distributions for black male full-time workers (NLSY 1979–1998).

salary income. This dissimilarity is also apparent when examining earnings distributions for self-employed and wage/salary workers. In figures 1 and 2, I display earnings distributions for black and Hispanic men. For both groups, a much larger percentage of the self-employed have very high or very low earnings than wage/salary workers. For example, 18 percent of self-employed blacks earn more than $60,000 whereas only 4 percent of blacks in the wage/salary sector have earnings at this level. At the other end of the distribution, 15 percent of self-employed blacks earn less than $10,000 compared to 8 percent of wage/salary blacks.

I also report characteristics of the earnings distribution for white men in table 2 and figure 3. The most notable difference is that white men earn substantially more than either black or Hispanic men in both the self-employment and wage/salary sectors.[10] Of interest to this analysis, however, is the difference between the two sectors. Using mean or median earnings, self-employed white men earn substantially more than their wage/salary counterparts. The differences are also similar in magnitude when measured as a percentage of wage/salary earnings. Finally, the comparison of self-employment and wage/salary earnings distributions for white men reveals similar patterns as those for black and Hispanic men.

In table 2, I also report estimates of the mean, median and standard deviation for self-employment and wage/salary earnings for black

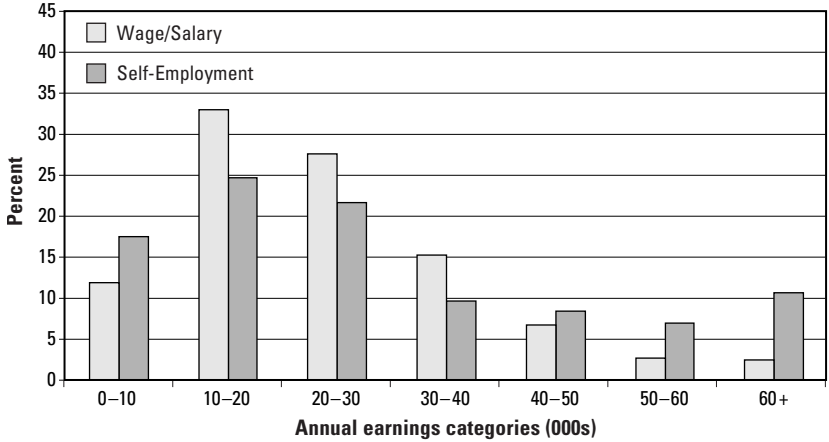


Figure 2
Earnings distributions for Hispanic male full-time workers (NLSY 1979–1998).

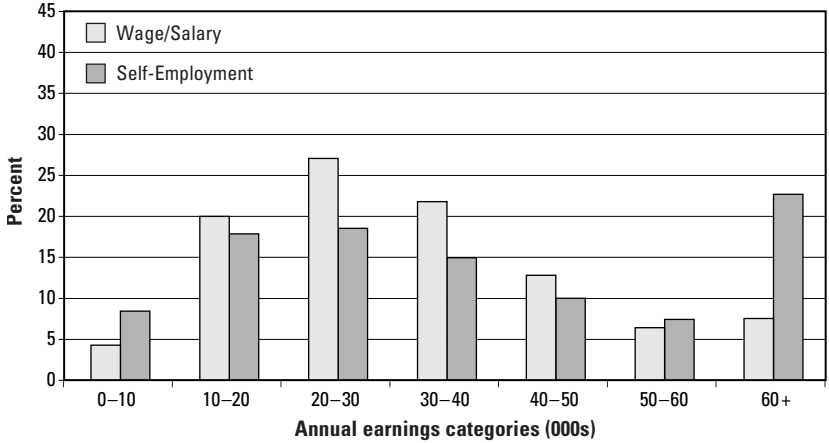


Figure 3
Earnings distributions for white male full-time workers (NLSY 1979–1998).

and Hispanic women. I should note, however, that some caution is warranted in interpreting these estimates for self-employment earnings as sample sizes are small. Similar to the results for men, I find that self-employed black and Hispanic women earn more than black and Hispanic women working in the wage/salary sector, although the difference is small for black women.[11] A major difference, however, is that median self-employment earnings are lower than median wage/salary earnings for black and Hispanic women. Median self-employment earnings are roughly $2,000–$3,000 less than median wage/salary earnings. Therefore, an evaluation regarding whether self-employed minority women earn more or less than minority women working in the wage/salary sector depends on the measure chosen.

The estimates reported in table 2 also indicate that self-employment earnings have a higher variance than wage/salary earnings for black and Hispanic women. The earnings distributions presented in figures 4 and 5 support this finding. Higher percentages of black and Hispanic women who are self-employed are found in the tails of the earnings distribution. Finally, I find that black and Hispanic women who are self-employed or work in the wage/salary sector earn less than white women. The one surprising exception is that mean earnings among self-employed Hispanic women is slightly higher than mean earnings among white women.

Returns to Capital

One issue that arises in comparing self-employment earnings to wage/salary earnings from survey data is the treatment of returns to capital. In the NLSY, the question regarding self-employment income asks "How much did you receive after expenses?" from your farm or business in the past calendar year. Although there is some uncertainty, respondents are likely to interpret this question to include both the returns to labor and the returns to capital. As noted above, however, the majority of the self-employed report their earnings as wage/salary income and not as business income. In the case of the respondent reporting income as business income it would be preferable to remove the returns to capital before making comparisons to the earnings of wage/salary workers.[12] This may not pose a substantial problem, however, because many business owners do not invest large amounts of capital. Data from the 1992 Characteristics of Business Owners sur-

vey indicate that 57 percent of small businesses require less than $5,000 of startup capital (U.S. Bureau of the Census 1997).[13] The percentages of black- and Hispanic-owned businesses started with less than $5,000 of capital are even greater (67 and 59 percent, respectively).

The NLSY contains two variables that may shed some light on the issue. It contains the market value of the individual's farm, business and/or other real estate and the total amount of debt owed on this farm, business and/or other real estate.[14] These two variables, however, suffer from three major problems. First, they are only for 1985–1990 and 1992–1998. Second, both measures include other real estate. There is a separate question asking whether the individual owns other real estate; however, a question on the value of the other real estate does not exist. Third, I do not have information on the percentage of the business owned by the respondent. The 1997 Survey of Minority Owned Businesses indicates that 90 percent of black firms and 86 percent of Hispanic firms, respectively, are individual proprietorships (U.S. Bureau of the Census 2001). With these reservations in mind, I proceed.

To remove the returns to capital from total self-employment income, I first need to calculate an opportunity cost for this capital. I calculate the owner's equity in the business, farm, and other real estate and multiply this by the rate of return on an alternative asset. I calculate estimates using both a less risky alternative (30-year Treasury Bond) and a more risky alternative (the S&P 500).[15] I then subtract this opportunity cost of capital from reported business income.[16] I do not subtract the opportunity cost of capital from reported wage/salary income for business owners. I assume that this income measure only captures the returns to labor.

Estimates of adjusted self-employment and wage/salary income are reported in table 3. I also report the average market value, debt, and equity in business, farm, and other real estate. Self-employed blacks and Hispanics have substantially lower levels of equity than do whites. Furthermore, within each racial group self-employed women have lower levels of equity than do self-employed men.[17]

In table 3, I also report unadjusted earnings for the self-employed and wage/salary workers for 1985–1990 and for 1992–1998. As expected, mean earnings are larger than reported in table 2. The differences between self-employment and wage/salary are generally similar. For all groups, the self-employed have higher earnings than wage/salary workers. As expected, the removal of the opportunity cost of

Table 3
Self-employment and wage/salary earnings, full-time workers (NLSY 1985–1998). Sample consists of youths who worked at least 1,400 hours in survey year. "White" includes all non-black, non-Hispanic individuals. Adjusted earnings remove the opportunity cost of equity in business, farm, and other real estate. See text for more details.

	Men			Women		
	Self-employed	Wage/salary	Difference	Self-employed	Wage/salary	Difference
Blacks						
Market value of business, farm, and other real estate	$33,590	$2,017	$31,573	$13,797	$1,429	$12,368
Debt owed on business, farm, and other real estate	$11,045	$983	$10,063	$974	$577	$397
Equity in business, farm, and other real estate	$22,544	$1,034	$21,510	$12,823	$852	$11,970
Unadjusted earnings	$31,550	$25,093	$6,457	$21,199	$20,703	$496
Adjusted earnings (30-year treasury bond)	$31,014	$25,084	$5,930	$21,056	$20,701	$355
Adjusted earnings (S&P 500)	$30,489	$25,076	$5,413	$20,998	$20,699	$300
Sample size	339	7,681		157	6,612	
Hispanics						
Market value of business, farm, and other real estate	$36,646	$5,866	$30,780	$25,270	$4,529	$20,741
Debt owed on business, farm, and other real estate	$13,289	$2,859	$10,430	$13,008	$1,909	$11,099
Equity in business, farm, and other real estate	$23,356	$3,006	$20,350	$12,262	$2,620	$9,642
Unadjusted earnings	$39,688	$28,486	$11,201	$23,828	$22,327	$1,501
Adjusted earnings (30-year treasury bond)	$39,017	$28,460	$10,557	$23,529	$22,322	$1,206
Adjusted earnings (S&P 500)	$38,317	$28,439	$9,879	$23,283	$22,317	$966
Sample size	391	5557		130	4039	

Whites						
Market value of business, farm, and other real estate	$81,287	$7,896	$73,390	$47,622	$8,870	$38,753
Debt owed on business, farm, and other real estate	$32,644	$3,186	$29,458	$17,478	$3,887	$13,591
Equity in business, farm, and other real estate	$48,642	$4,710	$43,932	$30,144	$4,983	$25,162
Unadjusted earnings	$48,943	$34,796	$14,146	$25,654	$24,858	$797
Adjusted earnings (30-year treasury bond)	$47,661	$34,753	$12,909	$25,054	$24,828	$226
Adjusted earnings (S&P 500)	$46,417	$34,718	$11,699	$24,637	$24,807	−$170
Sample size	1,617	15,379		673	11,665	

business, farm, and other real estate equity decreases relative self-employment earnings. For black and Hispanic men, however, the difference between mean self-employment earnings and wage/salary earnings remains large even when using the S&P 500 as the alternative investment. Self-employed blacks earn $5,413 more on average than wage/salary workers, and self-employed Hispanics earn $9,879 more. The earnings differences also decrease, but they remain positive for black and Hispanic women after adjusting for the opportunity cost of business equity. To conclude, the simple method used here to remove the returns to capital indicates that average self-employment earnings remain higher than average wage/salary earnings for blacks and Hispanics. The adjustment for the opportunity cost of capital does not substantially affect earnings comparisons. In view of these results and the uncertainty over how respondents interpret the income questions, I use total earnings in the remainder of the analysis.[18]

Estimates of Earnings Patterns

Overall, the results presented in table 2, table 3, and figures 1–6 provide evidence that self-employed black and Hispanic men earn more than black and Hispanic wage/salary workers. The evidence is less clear, however, for women. These estimates, which do not fully exploit the longitudinal nature of the data, provide some suggestive evidence

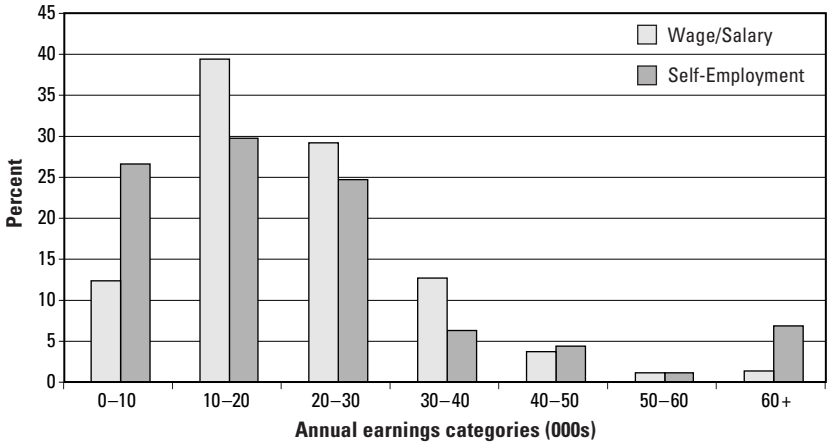


Figure 4
Earnings distributions for black female full-time workers (NLSY 1979–1998).

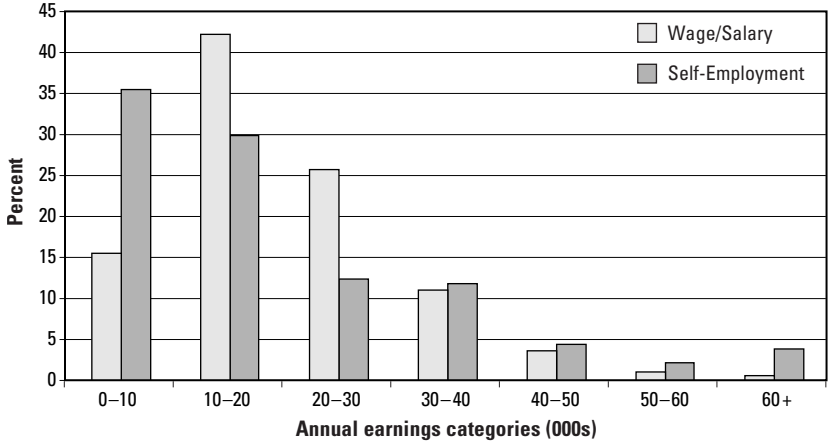


Figure 5
Earnings distributions for Hispanic female full-time workers (NLSY 1979–1998).

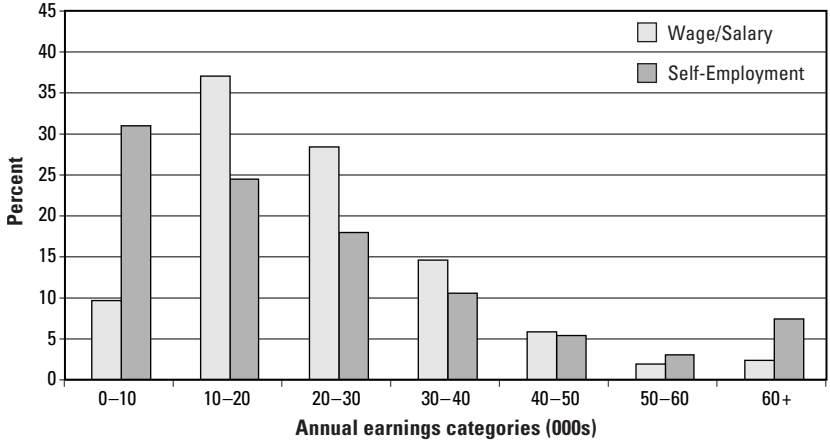


Figure 6
Earnings distributions for white female full-time workers (NLSY 1979–1998).

that self-employed blacks and Hispanics experience faster earnings growth than their wage/salary counterparts. Of course, this inference relies on the assumption that the two groups have the same initial earnings levels at entry into the labor market and have the same age distribution. In fact, pervious studies find that the self-employed are older on average than are wage/salary workers and differ in other important ways.[19] Another problem with the interpretation is that workers may select into the sector that provides the highest expected earnings. Therefore, even after controlling for differences in observable characteristics, self-employed and wage/salary workers may differ in unobservable characteristics.

Do black and Hispanic business owners experience faster earnings growth than black and Hispanic wage/salary workers? To explore this question, I compare the earnings patterns of black and Hispanic youths who were self-employed early in their careers to the earnings patterns of those who were wage/salary workers. The sample includes observations for young men and women who report working at least 1,400 annual hours in the survey year.

I estimate separate log earnings regressions for each race and sex. I control for current self-employment and wage/salary status and for differences in observable and unobservable characteristics. Specifically, I estimate the following reduced-form equation for annual earnings:

$$\ln y_{it} = \alpha_i + X'_{it}\beta + \gamma_1 t + \gamma_2 t^2 + \pi S_{it} + \gamma_1^S t S_{it} + \gamma_2^S t^2 S_{it} + \varepsilon_{it}, \quad (1)$$

where y_{it} is individual i's annual earnings in year t, α_i is an individual-level fixed effect, X_{it} is a vector of time-varying independent variables, t is a time trend which equals 0 at the completion of formal schooling, S_{it} is a dummy variable indicating whether the individual is self-employed in year t, and ε_{it} is the error term.[20] The individual-level fixed effects control for all observable and unobservable characteristics that do not change over time. The dummy variable for current self-employment status and its interactions with the time trend variables allow the earnings growth patterns to differ between self-employed and wage/salary workers. The difference between self-employment and wage/salary earnings at time t is equal to $\pi + \gamma_1^S t + \gamma_2^S t^2$. Because individuals make transitions between self-employment and wage/salary over time, comparisons of self-employment and wage/salary earnings for the same individual in different years contribute to identifying these coefficients.

Although estimates from equation 1 are useful in determining whether minorities who choose self-employment experience faster earnings growth on average than their wage/salary counterparts, it is impossible to infer from these estimates whether self-employment is a "better" option for the randomly chosen black or Hispanic. The standard economic model of the self-employment decision posits that workers choose the sector that provides the highest expected income or utility (see Evans and Jovanovic 1989, Rees and Shah 1986, and Reardon 1997 for examples). The fixed effects included in equation 1 control for the part of this selection that remains constant over time; however, they do not control for the possibility of a selection bias associated with workers choosing the sector that provides the fastest growth in earnings. Because of a lack of credible identifying instruments and the likely sensitivity of estimates to distributional assumptions, I do not address this issue. The difficulty lies in identifying a variable that theoretically affects the decision to become self-employed but does not affect self-employment and wage/salary earnings patterns.

I now turn to the results for black men. It is difficult to interpret the separate shift, linear growth, and quadratic growth coefficients for relative self-employment earnings in equation 1. Instead of simply reporting these coefficients, I simulate earnings patterns for the self-employed relative to wage/salary workers. These simulations are displayed in figure 7 (the actual coefficient estimates are reported in

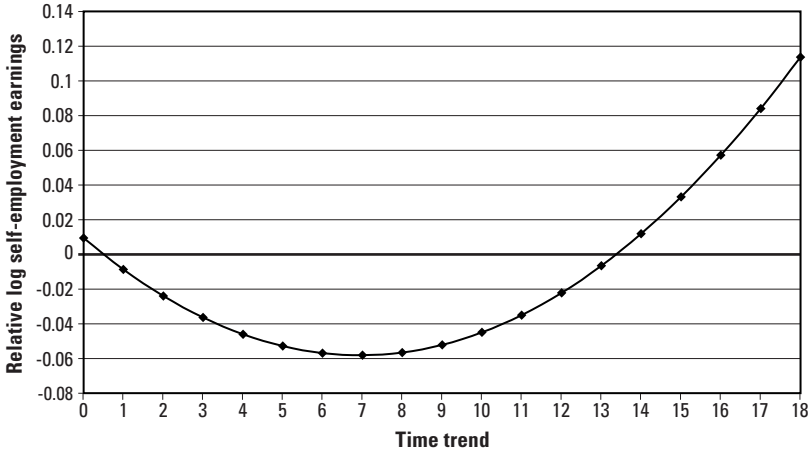


Figure 7
Combined effects of relative log self-employment earnings coefficients for black men (NLSY 1979–1998).

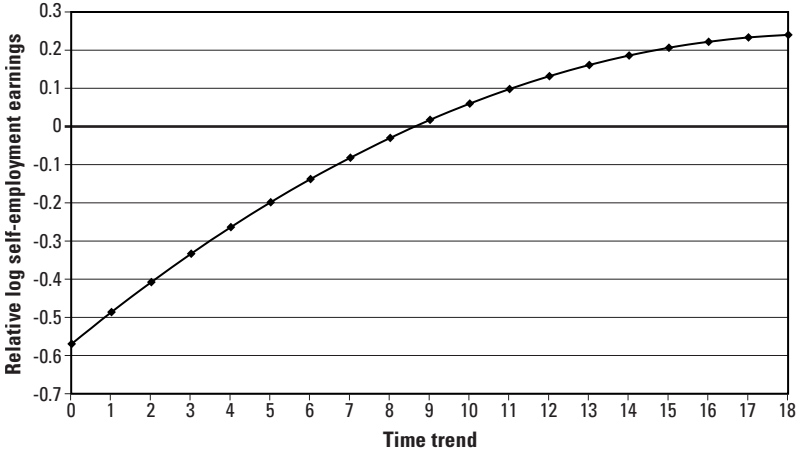


Figure 8
Combined effects of relative log self-employment earnings coefficients for Hispanic men (NLSY 1979–1998).

appendix table 2). The point estimates indicate that black men who are self-employed initially experience slower earnings growth than wage/salary workers; then after several years this reverses and they experience faster earnings growth and higher earnings. The two growth interaction coefficients, however, are not jointly statistically significant. I cannot reject the null hypothesis that the time trend interactions are different between the self-employed and wage/salary workers at conventional levels of significance. After removing these interactions, I find a positive and statistically significant coefficient on the self-employment dummy variable. This result confirms the previous findings that the self-employed earn more on average than wage/salary workers among black men.

Figure 8 displays the results for the sample of Hispanic men. The pattern suggests that Hispanic men who are self-employed start at much lower earnings levels than do wage/salary workers; however, they experience faster growth rates. In fact, the self-employed earn slightly more than wage/salary workers after 9 years. The hypothesis that the self-employment and wage/salary time trend coefficients are the same is easily rejected for Hispanic men. The time pattern suggests that on average self-employed Hispanic men may struggle in the first few years of owning a business relative to wage/salary workers, but they ultimately experience higher earnings. This pattern may also

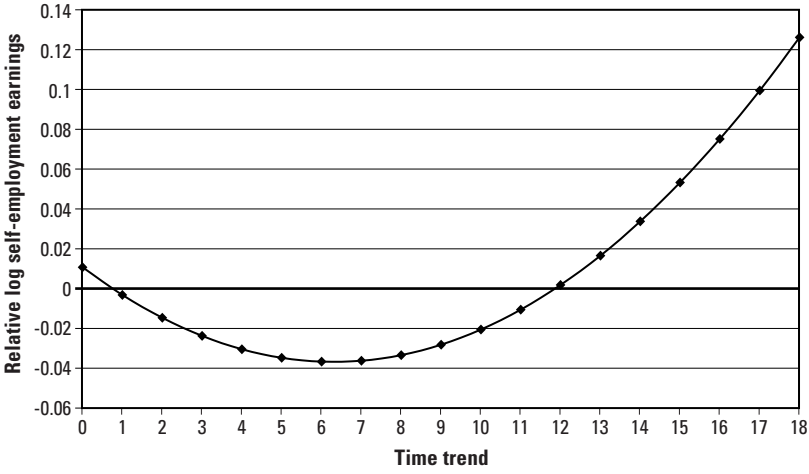


Figure 9
Combined effects of relative log self-employment earnings coefficients for white men (NLSY 1979–1998).

explain why Hispanic men have relatively low rates of self-employment as many may not be able to survive the initial years of low earnings.

To place the relative self-employment earnings patterns among blacks and Hispanics into context, it is useful to compare them to the patterns for white men. Figure 9 displays the results. The time pattern is strikingly similar to that for black men. The self-employed initially have lower earnings and slower growth than wage/salary workers. After several years, however, they experience faster growth and eventually higher earnings. The two growth coefficients are jointly significant at the $\alpha = 0.05$ level. The similarity of results suggests the possibility that the lack of statistical significance for the results among blacks may be due to small sample sizes. However, if the black and white male patterns are truly similar then it raises the question of why black self-employment rates are so much lower than white rates. It may have to do with blacks having difficulty obtaining credit (see Fairlie 1999 and Blanchflower, Levine, and Zimmerman 1998).

Figures 10 and 11 display the results for black and Hispanic women, respectively. The coefficient estimates imply similar patterns for the two groups. In both cases, the self-employed initially earn considerably less than wage/salary workers but essentially catch up after 10 years. For both groups, however, the pair of growth interactions is not jointly

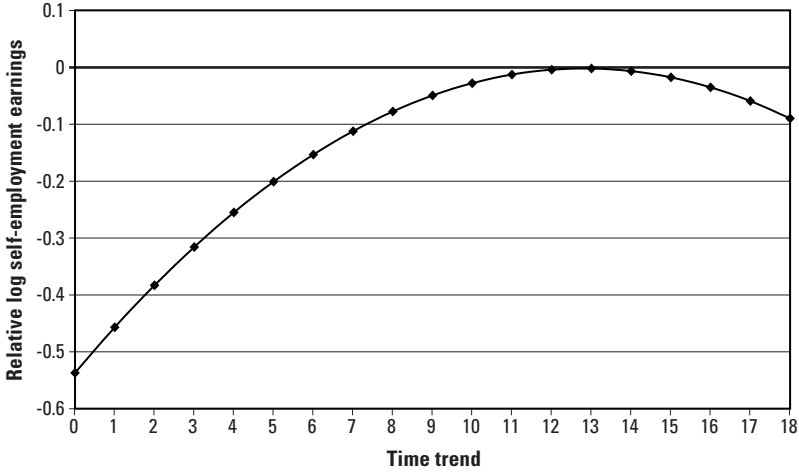


Figure 10
Combined effects of relative log self-employment earnings coefficients for black women (NLSY 1979–1998).

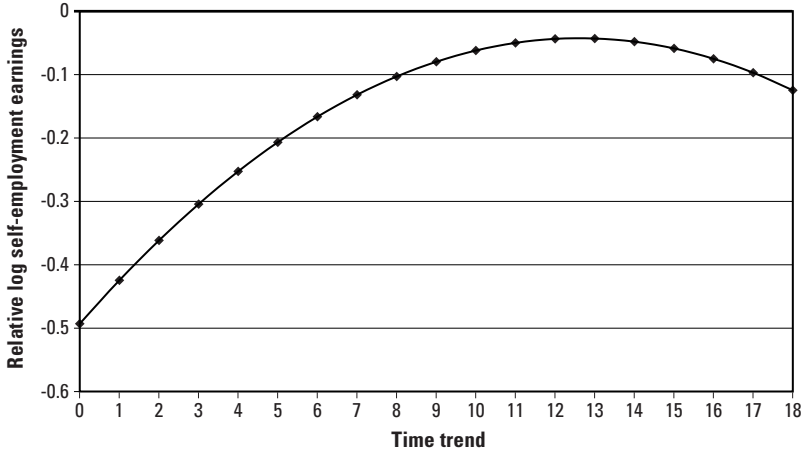


Figure 11
Combined effects of relative log self-employment earnings coefficients for Hispanic women (NLSY 1979–1998).

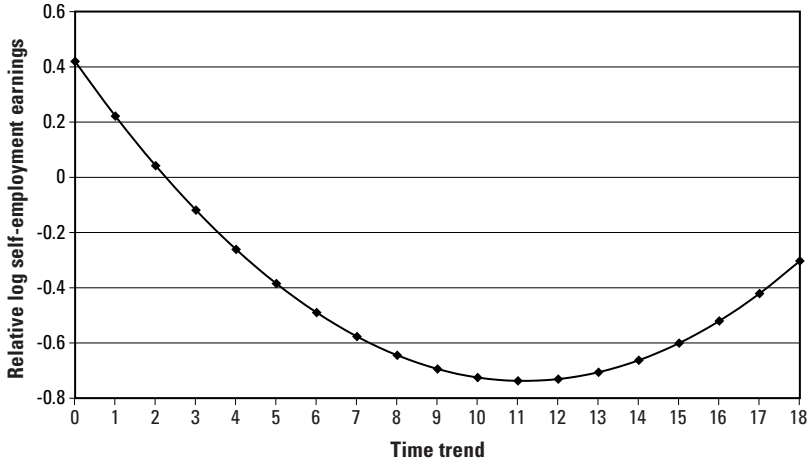


Figure 12
Combined effects of relative log self-employment earnings coefficients for white women (NLSY 1979–1998).

significant. Thus, it is difficult to infer much from these results. In contrast, the results for white women are statistically significant and indicate a different pattern (see figure 12). Relative self-employment earnings start out positive then become negative. After several years relative growth becomes positive and the gap narrows. The effects seem implausibly large, however. At 11 years, the self-employed earn more than 70 percent less than wage/salary workers. It is unclear what causes these patterns.

Additional Estimates

In all of the regressions discussed above, I enforce a consistent top code of $109,987 for each income question and assign these observations a value of $150,000. To determine whether my estimates of earnings growth are sensitive to these observations, I estimate equation 1 assigning $109,987 to all top-coded values of the income questions. This will further limit the influence that these high earnings observations have on the time trends. As was noted in the previous section, the self-employed are more likely to experience high earnings than are wage/salary workers. For all groups, the estimated relative self-employment earnings patterns are very similar to those displayed in figures 7–12. Evidently, the faster rates of earnings growth among the

self-employed are not simply due to the original assignment of values to top-coded observations.

Figures 1–6 also indicate that the self-employed are more likely to experience very low earnings observations than are wage/salary workers. Using a log specification for the earnings regression may allow these low earnings observations to overly influence the coefficient estimates. A simple method of checking the sensitivity of results to this concern is to censor all very low earnings observations. Specifically, I assign all earnings observations below $500 to equal $500.[21] In the sample, 2.3 percent of the self-employed and 0.6 percent of wage/salary workers are censored at $500. For black men, censoring results in a similar pattern for relative log self-employment earnings, with the curve shifting up slightly. The curve for Hispanic men shifts up more, moving the "break-even" point to the left. Relative self-employment earnings now become positive at 4 years instead of 9 years. For white men, the curve shifts upward to the point where relative self-employment earnings are always positive. The curve for black women shifts upward slightly, whereas the curve for Hispanic women shifts downward. Finally, the curve becomes more compressed for white women but has a very similar shape. Overall, these results suggest that the shape of the relative log self-employment earnings patterns are not sensitive to censoring at $500; however, for most groups it appears as though relative self-employment earnings would be higher.

I also examine whether previous self-employment has an independent effect on current earnings. For example, past self-employment may have a negative effect on wage/salary earnings if business failures often result in the owners being forced to take inferior wage/salary jobs. On the other hand, the experience gained from running a small business, even if it was unsuccessful, may be valuable to some employers. I include a vector of lag values of self-employment status for the previous 5 years in equation 1. Most of the coefficient estimates on the lag values of self-employment are statistically insignificant across samples. Furthermore, among the few statistically significant coefficients many are implausibly large. For example, I find a coefficient of 0.4214 on self-employment lagged 4 years for Hispanic women. I also experimented with including fewer lags and found roughly the same results. Overall, the findings from these regressions do not provide clear evidence that lagged self-employment has an independent effect on current earnings.

Conclusions

I use data from the National Longitudinal Survey (NLSY) to examine the earnings patterns of young black and Hispanic business owners and make comparisons to young black and Hispanic wage/salary workers. I find that self-employed black and Hispanic men have higher mean and median earnings than their wage/salary counterparts. The results for black and Hispanic women, however, are mixed.

I also compare the earnings growth of self-employed minorities to the earnings growth of minorities employed in the wage/salary sector. In particular, I estimate fixed-effects earnings regressions that control for differences in time-invariant observable and unobservable characteristics and time-varying observable characteristics. For black men, the point estimates from these earnings regressions indicate that the self-employed initially experience slower earnings growth than wage/salary workers. After several years this reverses and they experience faster earnings growth and higher earnings. The relative growth coefficients, however, are not statistically significant. For Hispanic men, the relative self-employment earnings coefficients suggest that the self-employed start at much lower earnings levels than do wage/salary workers, but they experience faster growth rates. In fact, the self-employed earn slightly more than wage/salary workers after 9 years. The relative growth coefficients are statistically significant. These patterns suggest that on average self-employed Hispanic men may struggle in the first few years of owning a business relative to wage/salary workers, but ultimately they experience higher earnings. Finally, the relative self-employment earnings coefficients are not statistically significant for either black or Hispanic women, possibly due to small sample sizes.

The results presented here provide some evidence that business ownership may provide a route for economic advancement among minority men when compared to opportunities in the wage/salary sector. The evidence is less clear for the contribution of self-employment to economic mobility for black and Hispanic women. Unfortunately, these results do not provide an answer to the question of whether a randomly chosen minority individual will experience faster earnings growth in self-employment than in wage/salary work as they simply make comparisons between the actual experiences of minorities who are self-employed and employed in the wage/salary sector. Perhaps future research will shed light on this question.

Although self-employed black and Hispanic men earn more on average than their counterparts in the wage/salary sector, they earn considerably less than self-employed white men. The estimates from table 3 indicate that self-employed black and Hispanic men earn 35.5 and 18.9 percent less than self-employed white men, respectively. The differences in business equity, however, are even more striking. Average business equity for self-employed black men is 53.7 less than the average for self-employed whites and average business equity for self-employed Hispanic men is 52.0 percent less than for whites.[22] Among women, self-employed blacks and Hispanics also have substantially lower levels of business equity than do whites. These disparities are important in light of the controversy surrounding set-aside programs that target government contracts for disadvantaged and minority-owned firms. Many of these programs, which were created in the late 1970s to the mid 1980s, have been both judicially and legislatively challenged and dismantled in the past decade. In particular, the landmark 1989 *City of Richmond v Croson Co.* Supreme Court decision, invalidated the use of local and state programs unless they were used as narrowly tailored remedies for identified discrimination. More recently, the 1995 *Adarand Constructors, Inc. v. Peña* Supreme Court decision and state referendums passed in California (Proposition 209 in 1996) and Washington (1998) further jeopardize the future of government set-asides. The elimination of these programs may further exacerbate racial inequalities in small-business outcomes as well as in rates of business ownership.[23]

Acknowledgments

This research was funded by the Small Business Administration and the UC Institute for Labor and Employment. The views expressed here are those of the author and not necessarily those of the funders. I would like to thank Peter Gottschalk, Doug Holtz-Eakin, and participants at the Entrepreneurship and Public Policy Conference at Syracuse University and the 2001 United States Association for Small Business and Entrepreneurship Conference for helpful comments. Bill Koch and Gaofeng Han provided excellent research assistance.

Appendix Tables

Appendix table 1
Self-employment rates and earnings estimates, 1990 census. Sample consists of individuals (ages 22–41) who worked at least 300 hours in survey year. "White" includes all non-black, non-Hispanic individuals. All estimates use sample weights provided by Census Bureau.

	Men		Women	
	Self-employed	Wage/salary	Self-employed	Wage/salary
Blacks				
Rate	3.70%		1.91%	
Mean earnings	$35,523	$28,023	$23,617	$23,990
Sample size	3,435	102,949	1,779	103,732
Hispanics				
Rate	6.23%		4.24%	
Mean earnings	$37,649	$27,350	$24,039	$22,292
Sample size	6,886	107,720	2,388	62,749
Whites				
Rate	10.89%		6.30%	
Mean earnings	$49,337	$38,802	$26,745	$27,257
Sample size	133,113	1,055,508	44,797	744,168

Appendix table 2
Fixed effects earnings regressions (NLSY 1979–1998). Sample consists of youths who worked at least 1,400 hours in survey year. Standard errors are in parentheses below coefficient estimates. All specifications include individual fixed effects, marital status, number of children, and dummy variables for the local unemployment rate. "White" includes all non-black, non-Hispanic individuals.

	Men			Women		
	Black	Hispanic	White	Black	Hispanic	White
Time trend	0.0992	0.1478	0.1109	0.0949	0.1193	0.1020
	(0.0076)	(0.0083)	(0.0046)	(0.0088)	(0.0099)	(0.0049)
Time trend squared	−0.0034	−0.0052	−0.0039	−0.0028	−0.0039	−0.0031
	(0.0003)	(0.0004)	(0.0002)	(0.0004)	(0.0005)	(0.0002)
Self-employed	0.0093	−0.5720	0.0112	−0.5371	−0.4931	0.4196
	(0.2134)	(0.2003)	(0.0856)	(0.4218)	(0.4000)	(0.1325)
Time trend	−0.0195	0.0852	−0.0151	0.0835	0.0714	−0.2074
× Self-employed	(0.0396)	(0.0368)	(0.0168)	(0.0771)	(0.0683)	(0.0252)
Time trend squared	0.0014	−0.0022	0.0012	−0.0033	−0.0028	0.0093
× Self-employed	(0.0017)	(0.0016)	(0.0008)	(0.0034)	(0.0029)	(0.0011)
R^2	0.4157	0.4038	0.3864	0.3970	0.4272	0.4831
Sample size	105,63	8,062	22,770	8,652	5,638	16,971

Notes

1. See Glazer and Moynihan 1970; Light 1972, 1979; Sowell 1981; Moore 1983.

2. Recent studies: Bates 1997; Blanchflower, Levine, and Zimmerman 1998; Fairlie 1999; Fairlie and Meyer 2000; Hout and Rosen 2000.

3. The difficulty lies in finding a variable that affects the decision to become self-employed, but does not affect self-employment and wage/salary earnings patterns.

4. For additional details on the NLSY sample, see Center for Human Resource Research 1999.

5. Unpaid family workers are not counted as self-employed. The current or most recent job or "Current Population Survey (CPS) employer" is defined as the job with the most hours for those who worked during the survey week and as the most recent job for those who did not work during the survey week. For more details, see Center for Human Resource Research 1999.

6. In the most recent years of the NLSY, the average value of all top-coded observations is assigned to top-coded observations. These are generally close to $150,000.

7. The rates are generally similar when including only workers with at least 1,400 hours in the past calendar year.

8. Apparently, higher average self-employment earnings are not due to differences in observed characteristics. Controlling for age, education, family characteristics, region, urbanicity, local unemployment rates, and AFQT scores, I find that self-employed black men earn $6,039 more than black wage/salary workers and self-employed Hispanic men earn $13,143 more than Hispanic wage/salary workers. Both estimates are statistically significant. Portes and Zhou (1999) find similar results using data from the 1990 Census. They find higher actual and adjusted earnings among self-employed native-born blacks and Hispanic immigrants than their counterparts in the wage/salary sector.

9. I should note, however, that this problem is mitigated somewhat by the top coding described above.

10. The earnings differences between minorities and whites in the wage/salary sector has been documented and studied extensively in the literature. (For a recent review, see Altonji and Blank 1998.) Previous estimates also indicate that black- and Hispanic-owned businesses have lower profits and sales than do white-owned businesses. (See U.S. Small Business Administration 1999; U.S. Bureau of the Census 1997.)

11. Controlling for differences in observable characteristics, I find that self-employed Hispanic women earn $2,198 more than Hispanic wage/salary workers. The estimated difference in earnings among black women, however, is small and statistically insignificant.

12. For a thorough discussion of the issues, see Yuengert 1996. Using data on both total income from the business and reported labor income from the 1989 Survey of Consumer Finances, he finds that the self-employed, on average, understate their labor earnings by 38 percent and overstate their capital income.

13. The definition of small business used in the CBO is anyone who filed an IRS form 1040 Schedule C (individual proprietorship or self-employed person), 1065 (partnership), or 1120S (subchapter S corporation).

14. The instructions on the two questions were "'Market Value' is defined as 'how much the respondent would reasonably expect someone else to pay if the item(s) were sold today in its/their present condition: not the original price the respondent paid for the item(s)'" and "What is the total amount of debts or liabilities you ... owe on this operation or property? Include any unpaid mortgages. (Do not include any commodity credit loans.)"

15. I calculate the average annual real rate of return from 1985 to 1998 for both investments. The rate of return on the Treasury bond and S&P 500 are 4.8 and 10.4 percent, respectively.

16. I censor equity and adjust business income at 0.

17. I also calculate business equity for the sample of self-employed who report not owning any other real estate. These levels of equity are from 7.6 to 29.8 percent lower than the levels reported in table 3.

18. Another potential problem with reported business income is the ambiguity regarding how reinvested profits are treated. As the question in the NLSY is written, we do not know whether respondents incorrectly subtract reinvested profits from total self-employment income. To complicate issues further, this may differ depending on how the profits are reinvested. Purchases of small equipment may be considered expenses, whereas purchases of large items such as buildings or vehicles may be considered profits as they are more likely to be depreciated over a long period of time.

19. These studies generally find that being male, white, older, married, and an immigrant, and having a self-employed parent, higher asset levels, and more education increase self-employment. For a review of earlier studies in this literature, see Aaronson 1991; for recent examples, see Hout and Rosen 2000; Blanchflower and Oswald 1998; Dunn and Holtz-Eakin 2000; Fairlie 1999.

20. I include marital status, children, and local unemployment rates as time varying controls. The coefficient estimates on the time trend interactions are not overly sensitive to their inclusion.

21. Another approach would be to exclude these observations from the sample, which is similar to the common approach of removing "implausibly" low hourly wages (e.g., less than $2 per hour) in the estimation of log earnings regressions among wage/salary workers. In the case of the self-employed, however, it is more problematic because these low earnings may be perfectly plausible.

22. The differences are even larger when I include only those who report not owning other real estate.

23. Chay and Fairlie (1998) provide some evidence that the minority business set-aside programs created in many large cities in the 1980s led to an increase in the number of black-owned construction firms.

References

Altonji, Joseph G., and Rebecca M. Blank. 1998. Race and Gender in the Labor Market Institute for Policy Research. Working paper WP-98-18, Northwestern University.

Aronson, Robert L. (1991): *Self-Employment: A Labor Market Perspective*. ILR Press.

Baron, Salo W., et al. 1985. *Economic History of the Jews*. Schocken.

Bates, Timothy. 1997. *Race, Self-Employment & Upward Mobility: An Illusive American Dream*. Woodrow Wilson Center Press and Johns Hopkins University Press.

Blanchflower, David G., Phillip B. Levine, and David J. Zimmerman. 1998. Discrimination in the Small Business Credit Market. Working paper 6840, National Bureau of Economic Research.

Blanchflower, David G., and Andrew J. Oswald. 1998. What makes an entrepreneur? *Journal of Labor Economics* 16, no. 1: 26–60.

Bonacich, Edna, and John Modell. 1980. *The Economic Basis of Ethnic Solidarity in the Japanese American Community*. University of California Press.

Center for Human Resource Research. 1999. *NLSY79 Users' Guide*. Ohio State University.

Chay, Kenneth Y., and Robert W. Fairlie. 1998. Minority Business Set-Asides and Black Self-Employment. Working paper, University of California.

Dunn, Thomas A., and Douglas J. Holtz-Eakin. 2000. Financial capital, human capital, and the transition to self-employment: Evidence from intergenerational links. *Journal of Labor Economics* 18, no. 2: 282–305.

Evans, David, and Boyan Jovanovic. 1989. An estimated model of entrepreneurial choice under liquidity constraints. *Journal of Political Economy* 97, no. 4: 808–827.

Fairlie, Robert W. 1999. The absence of the African-American owned business: An analysis of the dynamics of self-employment. *Journal of Labor Economics* 17, no. 1: 80–108.

Fairlie, Robert W. 2000. Earnings Growth among Less-Educated Business Owners. Working paper 207, Joint Center for Poverty Research.

Fairlie, Robert W., and Bruce D. Meyer. 2000. Trends in self-employment among black and white men during the twentieth century. *Journal of Human Resources* 35, no. 4: 643–669.

Glazer, Nathan, and Daniel P. Moynihan. 1970. *Beyond the Melting Pot: The Negroes, Puerto Ricans, Jews, Italians, and Irish of New York City*, second edition. MIT Press.

Holtz-Eakin, Douglas, Harvey S. Rosen, and Robert Weathers. 2000. Horatio Alger meets the mobility tables. *Small Business Economics* 14: 243–274.

Hout, Michael, and Harvey S. Rosen. 2000. Self-employment, family background, and race. *Journal of Human Resources* 35, no. 4: 670–689.

Light, Ivan. 1972. *Ethnic Enterprise in America*. University of California Press.

Light, Ivan. 1979. Disadvantaged minorities in self-employment. *International Journal of Comparative Sociology* 20, no. 1–2: 31–45.

Loewen, James W. 1971. *The Mississippi Chinese: Between Black and White*. Harvard University Press.

Min, Pyong Gap. 1989. Some Positive Functions of Ethnic Business for an Immigrant Community: Koreans in Los Angeles. Final report submitted to National Science Foundation.

Min, Pyong Gap. 1993. Korean immigrants in Los Angeles. In I. Light and P. Bhachu, eds., *Immigration and Entrepreneurship*. Transaction.

Moore, Robert L. 1983. Employer discrimination: Evidence from self-employed workers. *Review of Economics and Statistics* 65, August: 496–501.

Portes, Alejandro, and Min Zhou. 1999. Entrepreneurship and economic progress in the 1990s: A comparative analysis of immigrants and African Americans. In *Immigration and Opportunity*, ed. F. Bean and S. Bell-Rose. Russell Sage Foundation.

Reardon, Elaine. 1997. Are the Self-Employed Misfits or Superstars? Working paper, Milken Institute.

Rees, Hedley, and Anup Shah. 1986. An empirical analysis of self-employment in the U.K. *Journal of Applied Econometrics* 1, no. 1: 95–108.

Sowell, Thomas. 1981. *Markets and Minorities*. Basic Books.

U.S. Bureau of the Census. 1993. *1990 Census of the Population, Social and Economic Characteristics, United States*. Government Printing Office.

U.S. Bureau of the Census. 1997. *1992 Economic Census: Characteristics of Business Owners*. Government Printing Office.

U.S. Bureau of the Census. 2001. *1997 Economic Census: Survey of Minority-Owned Business Enterprises*. Government Printing Office.

U.S. Small Business Administration. 1999. *Minorities in Business*. Office of Advocacy.

Yuengert, Andrew M. 1996. Left-Out Capital and the Return to Labor and Capital in Self-Employment. Working paper, Pepperdine University.

7 Entrepreneurial Activity and Wealth Inequality: A Historical Perspective

Carolyn M. Moehling and
Richard H. Steckel

The recent rise in wealth inequality in the United States has drawn even greater attention to the connection between entrepreneurial activity and the distribution of wealth. Entrepreneurs today hold a substantial fraction of the economy's wealth. In 1989, entrepreneurs accounted for only 8.7 percent of the U.S. population but held 37.7 percent of the nation's net worth (Gentry and Hubbard 2000, p. 7). Economists attribute the substantial wealthholdings of entrepreneurs, at least in part, to imperfections in capital markets that constrain would-be entrepreneurs to draw primarily on their own wealth for startup capital (Evans and Jovanovic 1989; Evans and Leighton 1989; Holtz-Eakin, Joulfaian, and Rosen 1994; Blanchflower and Oswald 1998; Quadrini 2000). This notion of constrained self-selection into entrepreneurship has been used to explain the "Kuznets curve," which proposes that inequality first increases and then decreases during periods of rapid economic growth. In the initial stages of growth, only a small number of fairly wealthy individuals are able to take advantage of the expanding opportunities and engage in entrepreneurial activity. Inequality rises as the gap between these wealthy entrepreneurs and the rest of the population increases. But as growth continues, average income increases and credit constraints ease, leading to greater equality.[1]

This chapter provides historical perspective by examining the relationship between entrepreneurial activity and wealth inequality in nineteenth-century Massachusetts. We make use of a unique data set that links information from the federal censuses to property tax records. The data allow us not only to examine the concentration of wealthholdings among entrepreneurs in an earlier period but also to examine the relationship among entrepreneurial activity, economic growth, and changes in the distribution of wealth. During the nineteenth century, Massachusetts experienced rapid industrialization,

tremendous economic growth, and a dramatic increase in wealth inequality. In a recent paper, we argued that standard explanations of changes in the returns to skills and changes in life-cycle saving behavior do not fit the changes in the distribution of wealth that we observe (Steckel and Moehling 2001). Here we consider the role of entrepreneurial activity in these changes. In particular, we consider how the distribution of wealth over this period was related to the fraction of the population engaged in entrepreneurial activity, the share of wealth held by entrepreneurs, and the inequality in wealthholdings among entrepreneurs.

The Data

Our data set links data from the federal censuses of 1820 to 1910 to data in the property tax records of Massachusetts.[2] After a survey of the available tax records, samples were taken from localities that had a complete set of records for the period 1820–1910: Boston, Salem, Lexington, Westminster, and Sturbridge. In each census year approximately 1,200 households, equally divided between the urban and rural areas, were randomly chosen from the census manuscript schedules.[3]

The tax records were maintained in alphabetical order of taxpayers by ward in cities or by town or township in rural areas. The census manuscript schedules were alphabetized accordingly to facilitate the search for a match. The tax records were searched for matches for only the household heads in the census samples. If a household head was not found in the tax records, it was assumed that he had no taxable property. This assumption may lead to errors in tabulating wealth in cases of garbled names or where individuals moved between the dates of the census and the tax enumeration.[4] Since matches were sometimes ambiguous, a coding procedure was devised to rate the confidence of the match, with categories of exact match, nearly exact match, probable match, improbable match, and duplicate (two or more people with the same name). The last two categories, which amounted to 2.1–4.1 percent of the sample (depending upon census year), were omitted from the analysis.

These linked data provide a valuable new tool for measuring and analyzing long-term trends in wealth inequality. Real and personal property taxes formed the backbone of state and local tax revenues until income and sales taxes were introduced in the twentieth century. According to Richard T. Ely (1888, p. 131), the antebellum period "wit-

nessed the complete establishment of the American system of state and local taxation. The distinguishing feature is … the taxation of all property, moveable and immovable, visible and invisible, real and personal, as we say in America, at one uniform rate." In 1796 the list of ratable property in Massachusetts was "so long as to include almost everything" (ibid., p. 138). All real and personal property not specially exempted was subject to taxation. Real estate included land and buildings, and personal estate included goods, chattels, money and effects (wherever they were); ships; money at interest; public stocks and securities; and stocks in turnpikes, bridges, and moneyed corporations, in or out of state. Over time, however, different types of property became exempt from taxation. In 1821, exemptions were introduced for household furniture not exceeding $1,000 in value, for wearing apparel, for farming utensils, and for mechanics' tools (Nichols 1938; Bullock 1916).[5] After 1860, concerns about double taxation led to exemptions for financial assets such as deposits in savings banks, stocks in Massachusetts corporations, and notes secured by mortgages of taxable real estate (Bullock 1916).

The "taxable wealth" data obtained from property tax records is clearly distinct from the wealth concepts used in studies of the wealth distribution in the current period. First, taxable wealth is a less comprehensive measure of wealth. The exemptions of furniture and clothing probably are of little consequence, since such items are typically not included in modern wealth measures.[6] The exclusion of certain forms of financial assets, however, represents a significant limitation of taxable wealth as a measure of total wealth. Even non-exempt financial assets were likely poorly captured in taxable wealth due to problems of tax avoidance. Intangible assets were fairly easy to hide from tax assessors. When the exemptions of the various types of financial assets went into effect, in fact, the total value of personal property in the state changed very little, suggesting that only a very small fraction of those assets were ever included on the tax rolls (Bullock 1916).[7] Moreover, the taxable wealth in Massachusetts during the period of study did not account for the debts and liabilities of wealthholders. The most commonly used measure of wealth in modern studies is "net worth," which subtracts the value of debts and liabilities from the value of assets. Up until 1916, however, taxpayers in Massachusetts could only subtract debts and liabilities from the value of their credit holdings. Another limitation arises from the way the linked data sets were constructed: our measure of taxable wealth does not include the value of

an individual's holdings of real estate and physical capital located in other tax jurisdictions.[8]

Taxable wealth is perhaps best interpreted as a measure of the value of physical capital holdings. For most individuals in our samples, taxable wealth understates net worth. The degree of understatement, however, likely varies greatly across individuals in any census year. The types of wealth excluded from taxable wealth—financial assets and real estate in located in other tax jurisdictions—were concentrated among the very rich in the nineteenth century perhaps even more so than they are today. More problematic, though, is that the degree of understatement likely varied over the sample period as capital markets expanded and financial assets became more important forms of wealthholding. However, to the extent that the value of an individual's holdings of physical capital was correlated with the value of his holdings of other assets, the analysis of taxable wealth can reveal factors influencing the distribution of wealth and the direction, if not the extent, of changes in inequality.

Other data on wealthholdings in the nineteenth century suffer from their own limitations. The two most prominent sources of historical wealth data are probate records and the federal population censuses of 1850, 1860, and 1870. Both are rich sources and have revealed a great deal about the determinants of wealth and its distribution in the past. But each has its own limitations. Probate data suffer from selection problems. Many individuals did not leave wills and their estates were never probated. The census wealth data come from self-reports and probably are very noisy measures of wealthholding. The 1860 instructions to the enumerators concerning the reporting of personal property wealth state: "Exact accuracy may not be arrived at, but all persons should be encouraged to give a near and prompt estimate for your information." The reported wealth levels are clustered on multiples of 100, indicating a strong tendency for rounding in the self-reports. Timothy Conley and David Galenson (1994) have also pointed out that the census wealth data appear to be censored, but the point of censoring is uncertain and may have varied across enumerators.

The real advantage of the linked data, however, is that they provide fairly consistent information on the wealth distribution over a long time span. Previous studies of trends in wealth inequality have been forced to interpolate between estimates of the distribution of wealth derived from probate records, census wealth data, and estate tax records (Williamson and Lindert 1980; Shammas 1994). These data

sources differ tremendously in both their coverage and their measurement of wealth. Differences in measured inequality across these sources, therefore, may represent differences in the nature of the underlying data rather than changes in the distribution of wealth. The linked data also provide information on the distribution of wealth during the periods between the standard benchmark estimates. Trends in wealth inequality in the late nineteenth century, for instance, have previously had to be inferred from interpolations between estimates from the 1870 census data and the estate tax data for 1922. Williamson and Lindert (1980) labeled this period "an empirical Dark Age for wealth distributions" (p. 47). The linked data, by providing estimates of the wealth distribution in 1880, 1900, and 1910, bring light into this dark age.

In this chapter, we limit our analysis to male heads of households.[9] Throughout the analysis, we weight the data in each sample so that the share urban equals the share urban in the state for that year.

Trends in Wealth Inequality

We use three measures of inequality to examine long-run changes in the distribution of wealth: the shares of total wealth held by the top fractiles of the wealth distribution, the Gini coefficient, and the Theil entropy measure.[10] Previous research on historical wealth distributions has focused on the first two of these measures. The shares of wealth held by the top 20 percent or the top 5 percent of the wealth distribution are straightforward and easy to calculate, but they do not capture the degree of dispersion within the top fractiles and ignore the lower fractiles of the wealth distribution. The Gini coefficient is an index measure based on the average absolute difference in wealth levels between all pairs of individuals or households. The Gini coefficient also has the more intuitive interpretation as twice the area between the Lorenz curve and the diagonal representing the case of "perfect equality" when all individuals have the same level of wealth.

Like the Gini coefficient, the Theil entropy measure is an index measure based on the entire wealth distribution. The Theil measure is given by

$$T = \frac{1}{n}\sum_{i=1}^{n}\frac{w_i}{\mu}\ln\left(\frac{w_i}{\mu}\right), \tag{1}$$

where n represents the number of observations, w_i represents the wealth of individual i, μ represents the full sample mean wealth, and [0 ln(0)] is taken to be 0 (Foster 1985, p. 55).[11] In the case of "perfect equality" when all individuals have the same level of wealth, the Theil measure, like the Gini coefficient, equals 0. In the case of "perfect inequality" when one individual owns all of the society's wealth, the Theil measure equals [ln(n)]. Although this means that the maximum value of the Theil measure varies with sample size, the rapidly diminishing slope of the natural log function makes this of little practical importance for samples, such as ours, that are fairly large and approximately the same size.

For all of the calculated inequality measures, we use bootstrap methods to estimate approximate standard errors, construct confidence intervals, and perform hypothesis tests.[12] For each sample of size n, we construct 1,000 resamples of size n by random draws with replacement from the original sample. Following Mills and Zandvakili (1997), we use the "percentile method" to construct confidence intervals, calculating tail probabilities directly from the bootstrapped distribution.[13]

Table 1 reports the aggregate measures of wealth inequality for the samples of male household heads. The measures are highly correlated: increases in the shares of wealth of the top 20 percent and the top 5 percent of the wealth distribution correspond to increases in the Gini coefficient and the Theil entropy measure. The correlation between the Gini coefficient and the Theil measure is 0.97. The choice of inequality measure, therefore, has little effect on the observed trends in wealth inequality.

Wealth inequality increased substantially between 1820 and 1910. The shares of wealth held by the top fractiles of the wealth distribution, the Gini coefficient, and the Theil entropy measure all indicate that wealthholdings were much more concentrated in 1910 than in 1820. The rise in inequality, though, was not steady. This can be best seen in figure 1, which plots the Gini coefficients and Theil entropy measures. The dashed lines represent 95 percent confidence intervals. Wealth inequality grew sharply between 1820 and 1850, leveled off between 1850 and 1870, and then rose steadily until 1900. The increases in both measures between 1820 and 1850 and between 1870 and 1900 are statistically significant at the 5 percent level.[14]

The growing concentration of wealthholdings was driven to a large extent by the increase in the percent of households with zero taxable wealth. The percentage of male household heads with zero taxable

Table 1
Distribution of total taxable wealth, male household heads, Massachusetts, 1820–1910. Numbers in parentheses are approximate standard errors obtained by bootstrapping. Data weighted so that share urban in sample equal to the share urban in the state each census year.

			Percentage of wealth held by			
	N	% zero wealth	Top 20%	Top 5%	Gini coefficient	Theil entropy measure
1820	1,016	34.2	72.0	40.5	0.720	1.125
			(1.7)	(2.7)	(0.015)	(0.086)
1830	989	34.1	77.6	49.2	0.775	1.486
			(2.2)	(4.5)	(0.020)	(0.175)
1840	977	37.7	78.3	45.0	0.771	1.282
			(1.6)	(2.8)	(0.013)	(0.072)
1850	1,023	51.2	85.8	55.7	0.836	1.761
			(1.7)	(4.5)	(0.016)	(0.147)
1860	1,005	54.7	88.1	55.7	0.844	1.679
			(1.4)	(3.5)	(0.012)	(0.096)
1870	1,017	58.2	90.1	56.7	0.856	1.730
			(1.2)	(3.3)	(0.011)	(0.086)
1880	1,020	66.2	93.7	60.3	0.877	1.924
			(1.2)	(4.1)	(0.012)	(0.136)
1900	977	70.1	97.3	70.5	0.911	2.264
			(0.9)	(4.0)	(0.011)	(0.157)
1910	1,003	73.6	98.3	68.7	0.910	2.207
			(0.8)	(3.8)	(0.010)	(0.124)

wealth more than doubled between 1820 and 1910 (from 34.2 percent to 73.6 percent). The Theil entropy measure can be decomposed to reveal how much of the measured inequality was due to the dispersion in wealth among individuals with positive wealth and how much was due to the fraction of the population with zero taxable wealth:

$$T = \frac{1}{n_+}\sum_{i=1}^{n_+}\frac{w_i}{\mu_+}\ln\left(\frac{w_i}{\mu_+}\right) - \ln\left(\frac{n_+}{n}\right), \tag{2}$$

where n_+ represents the number of observations with positive wealthholdings and μ_+ represents the mean wealth conditional on having positive wealthholdings. The first term on the right-hand side of equation 2 is just the Theil entropy measure calculated for only the observations with positive wealthholdings. The second term captures the inequality arising from the fraction of observations with zero wealth.

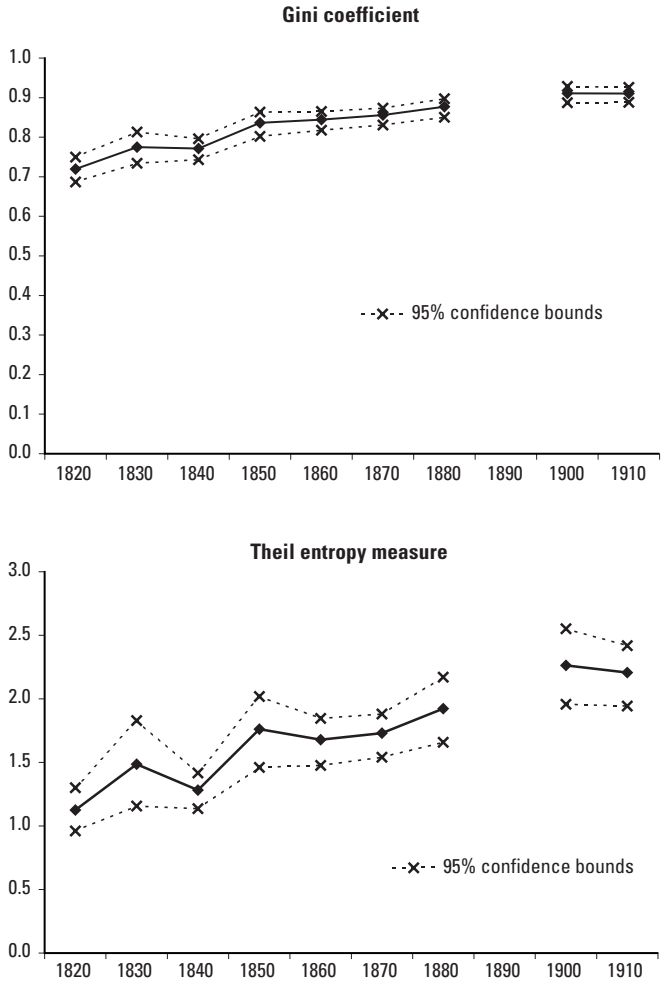


Figure 1
Wealth inequality, Massachusetts, 1820–1910.

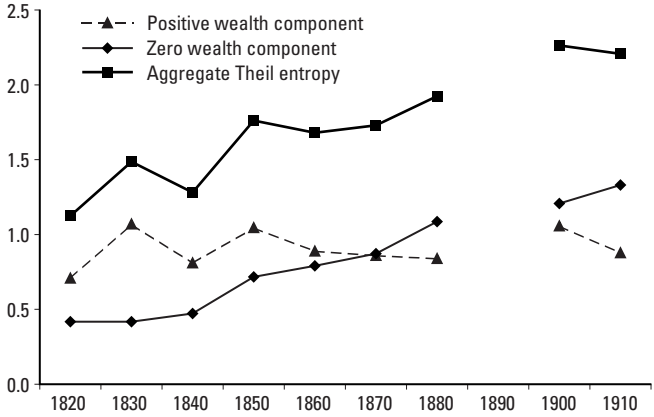


Figure 2
Decomposition of Theil entropy measure into "positive" and "zero" wealth components.

Figure 2 plots the Theil entropy measure and its "positive wealth" and "zero wealth" components for 1820 to 1910. The rise in the fraction of observations with zero wealth clearly played an important role in the rise in measured wealth inequality over the sample period. Inequality among those with positive wealthholdings did increase between 1820 and 1850 and between 1870 and 1900—the two periods of rising inequality identified from the data in table 1. But between those two periods, the positive wealth component decreased. Aggregate inequality was higher in 1910 than in 1820 only because of the dramatic increase over the century in the fraction of individuals with zero taxable wealth.

The prominent role played by the "zeros" raises some obvious concerns. In the linked data, having zero wealth simply meant that an individual could not be found in the property tax records. Therefore, the increase in individuals with zero wealth in the data could have been caused by something other than growing inequality. Perhaps fewer individuals were matched because migration rates increased or the growth in city size complicated the maintenance of accurate records. One test of this possibility is provided by data on the number of individuals assessed for only the poll tax (a head tax) relative to the total number of polls assessed. If slippage or disagreement between the census and the tax records was small, then this ratio should approximately equal the share of household heads not found in the tax

records. This is indeed the case. The ratio of the number assessed for the poll tax only to the total number of male polls assessed was 67.2 percent in 1886, 70.8 percent in 1890, 72.0 percent in 1900, and 77.9 percent in 1910.[15] These levels and the trend approximately agree with those in table 1 for the percent with no match near the turn of the century.[16] The data on polls, therefore, confirm that the fraction of men in Massachusetts with no taxable wealth was growing.

But was the increase in the fraction of men with no taxable wealth due to an increase in the fraction of men with relatively little wealth, or was it due to a decrease in the diligence or efficacy of tax assessors? This is a much more difficult question. Some insights can be gleaned from a comparison of the taxable wealth data to the self-reported wealth data collected in the 1850–1870 censuses. The census wealth data suffer from their own set of limitations as described above, but they do provide independent measures of the wealthholdings of individuals in our samples for 1850–1870. The census wealth data were intended to be used for informational purposes only. Although some individuals may have intentionally misrepresented their wealth in the census, the incentives to hide wealth from census enumerators would seem to be much weaker than the incentives to hide wealth from tax assessors. If tax assessors were effective in determining who did and did not have taxable wealth, we would expect that individuals not found in the tax records would have reported zero or very small wealthholdings in the census. More importantly for our current purpose, if the effectiveness of tax assessors was not deteriorating, the discrepancies between the taxable wealth and census wealth should not increase across the samples.

In general, the data indicate that tax assessors were fairly effective in identifying individuals with taxable wealth and that this effectiveness was not diminishing between 1850 and 1870. In the 1850 sample, 89 percent of the individuals with zero taxable real estate wealth reported zero real estate wealth in the census. In the 1860 and 1870 samples, that figure fell to approximately 83 percent, but an additional 3–5 percent reported less than $1,000 in real estate wealth in the census. Personal property wealth data were collected only in the 1860 and 1870 censuses. In the 1860 sample, 42 percent of those with zero taxable personal property wealth reported zero personal property wealth in the census and an additional 37 percent reported less than $500 of personal property wealth in the census. In the 1870 census, the corresponding percentages were 64 and 16.[17] The finding that discrepancies did

not increase between 1860 and 1870 is particularly significant because tax rates increased dramatically during this decade. In Boston, for instance, the tax rate per $1,000 of wealth increased from $8.90 in 1861 to $15.30 in 1870.[18]

Unfortunately, we have no similar independent data on wealth for the 1880, 1900, and 1910 samples. The increases in average tax rates over this period were small relative to the increase that took place between 1860 and 1870 (Bullock 1916, p. 24). But other factors such as increasing population and growing complexity in wealth portfolios may have affected the thoroughness of tax assessments.

Entrepreneurship and Wealth Inequality

In a recent paper (Steckel and Moehling 2001), we considered the received explanations of rising wealth inequality in the nineteenth century to see how well they fit the features of the postbellum rise in the concentration of wealthholdings in Massachusetts. We found that, as Simon Kuznets (1955) argued, compositional changes in the population resulting from rapid urbanization and the shift out of agriculture did contribute to rising wealth inequality. However, counter to explanations proposing increases in the returns to skill, wealth inequality between occupation groups was actually declining during this period in Massachusetts. The change in the distribution of wealth across age groups was also inconsistent with stories about changes in life-cycle saving patterns. Overall, we found that the increase in inequality was not driven by rising inequality between population groups but rather by rising inequality *within* population groups.

Here we examine the role played by entrepreneurial activity in the growing concentration of wealth in late-nineteenth-century Massachusetts.[19] Entrepreneurs today hold a substantial fraction of the economy's wealth. Economists attribute this concentration of wealth to the effects of imperfect credit markets. Credit constraints, it is argued, force would-be entrepreneurs to finance their ventures primarily out of their own wealthholdings. Evans and Jovanovic (1989), Evans and Leighton (1989), and Blanchflower and Oswald (1998) all have shown that the likelihood of becoming an entrepreneur is positively related to an individual's level of wealth. Even after entry, credit constraints appear to continue to bind and affect the profitability of enterprises. Holtz-Eakin, Joulfaian, and Rosen (1994) show that exogenous increases in wealth due to inheritances increase the likelihood an individual will remain

in entrepreneurship and increase the receipts from entrepreneurial activity.

This notion of constrained self-selection into entrepreneurship has been incorporated into theoretical models seeking to explain not only the highly skewed distribution of wealth observed in many industrialized economies today (Quadrini 2000) but also the perceived connection between economic growth and rising inequality (Banerjee and Newman 1993; Aghion and Bolton 1997). Under certain conditions, these models can generate the "Kuznets curve" which proposes that inequality first increases then decreases during periods of rapid economic growth. During the initial stages of growth, only a small number of fairly wealthy individuals are able to take advantage of the expanding opportunities and engage in entrepreneurial activity. Inequality rises as the gap between these wealthy entrepreneurs and the rest of the population increases. The fact that some entrepreneurs are credit constrained and others are not, also leads to an increase in inequality within the ranks of entrepreneurs. As growth continues, however, average income increases and credit constraints ease, leading to greater equality.[20]

Late-nineteenth-century Massachusetts would seem—at least superficially—to fit the first half of this story. The last few decades of the nineteenth century are referred to both as the Second Industrial Revolution and as the Gilded Age. The scale and scope of industrial production increased dramatically, as did the fortunes of such "captains of industry" as John D. Rockefeller and Andrew Carnegie. Massachusetts was at the forefront of this industrial expansion. Between 1870 and 1900, the amount of capital invested in manufacturing in the state (measured in constant dollars) increased by a factor of 5.[21] This expansion was accompanied by rapid wealth accumulation. Over the same period, the value of wealth assessed in the state increased by a factor of 3.[22] Were the changes in the distribution of wealth over this period consistent with those predicted by models of entrepreneurship with imperfect capital markets? These models and the empirical work on entrepreneurship suggest two sources of rising inequality: an increasing gap between entrepreneurs and the rest of the population, and a growing inequality within the ranks of entrepreneurs.

Defining Entrepreneurship

Joseph Schumpeter (1947, p. 151) claimed that the defining characteristic of an entrepreneur is "simply the doing of new things or the doing

of things that are already being done in a new way (innovation)." This is an intuitively appealing definition, but it provides little guidance for identifying entrepreneurs in the data. The most common definition of entrepreneurship used in empirical analysis is self-employment.[23] Self-employment, of course, is only a weak indicator of entrepreneurial activity. It is at the same time both too inclusive and too exclusive to capture Schumpeterian entrepreneurship. Most criticism, however, focuses on the fact that self-employment is too inclusive. Many in the ranks of the self-employed are clearly not involved in innovative activities. The benefit of using self-employment, though, is that information on employment status is available in a wide range of data sources. Unfortunately for our purposes, data on employment status were collected in only one of the censuses used to create the linked data sets. Only in 1910 did the census ask individuals whether they were employers, employees, or working "on their own account." The censuses before 1850, in fact, collected only data on the number of household members engaged in broadly defined industries such as agriculture and manufacturing.[24] Therefore, we cannot examine entrepreneurial activity and inequality before 1850. Starting in 1850, though, the census collected information on individuals' occupations. We use these occupation data to identify likely self-employment.

In this chapter, we focus on self-employment outside of agriculture. Farming was the most prevalent form of self-employment in the nineteenth century, and it remains so today (Gentry and Hubbard 2000, p. 6). Surely some farmers fit Schumpeter's definition of entrepreneurs. But agriculture in the nineteenth century was a sector in decline both in its share of employment and in its share of output. In Massachusetts in particular, the movement was out of, rather than into, farming during this period.[25]

Determining self-employment from occupation titles is not a straightforward task. Some occupation titles, such as "restaurant proprietor" and "glove manufacturer" clearly indicate self-employment; others, such as "assistant dressmaker" and "works in foundry," clearly indicate employee status. But many occupation titles do not reveal employment status. Were individuals who were reported as "barbers," "boilermakers," "tailors," or "millers" self-employed? To deal with this issue, we examined self-employment by occupation using data from the 1910 census available as part of the Integrated Public Use Microdata Series.[26] We classified as "likely self-employed" all occupations with self-employment rates of at least 25 percent. We chose this relatively low threshold to make our base-line classification fairly

inclusive. Over our sample period, self-employment was declining in many occupations. Using a threshold of 50 or 75 percent self-employed in 1910 would have led to the exclusion of occupations (such as blacksmith and tailor) that had high rates of self-employment in the middle of the nineteenth century. Even the threshold of 25 percent leads to the exclusion of some occupations that had high self-employment rates in the 1850s. By 1910, the artisanal shop had been all but eliminated in industries like meat packing and leather tanning and finishing. Yet in 1850 artisanal shops still produced substantial shares of the output in these industries. So we also include in the "likely self-employed" category skilled occupations in industries in which artisanal shops produced at least one-fourth of the industry's total value added in 1850 or 1870.[27]

The occupations we classify as likely self-employed fall into eight broad categories: *artisans*, such as blacksmiths, carpenters, and tailors; *contractors*; *manufacturers*; *personal service occupations*, such as barbers, manicurists, and laundrymen; *professional occupations*, such as lawyers, physicians, and veterinarians; *proprietors of service establishments*, such as restaurant, saloon, and hotel owners; *rentiers*, such as landlords, capitalists, and speculators; and *trade occupations*, such as merchants, brokers, and dealers.

The employment status information collected in the 1910 data allow us to test how well the occupation-based measures capture self-employment. Table 2 presents for the 1910 sample a cross-tabulation of inferred employment status based on occupation title and employment

Table 2
Cross-tabulation of inferred employment status based on occupation title and employment status as reported in 1910 census, 1910 linked sample. The top figure in each cell represents the number of observations in the category; the bottom figure in parentheses represents the percentage of all observations accounted for by the category. Sample consists of individuals with non-agricultural occupations and non-missing data on employment status.

Inferred employment status based on occupation title	Census-reported employment status		
	Employee	Self-employed	Total
Employee	521	19	540
	(63.1)	(2.3)	(65.4)
Self-employed	124	162	286
	(15.0)	(19.6)	(34.6)
Total	645	181	826
	(78.1)	(21.9)	(100.0)

status as reported in the census. The observations used in this tabulation include only individuals with non-agricultural occupations and non-missing data on employment status. The occupation-based assignments correctly classify 83 percent of the observations: 63 percent are correctly identified as employees and 20 percent are correctly identified as self-employed. The occupation-based assignments do, however, overstate self-employment in 1910. Fifteen percent of the observations were misclassified as self-employed by the occupation-title assignments. Of the 286 observations classified as self-employed by occupation title, only 162, or 57 percent, were reported as self-employed in the census. Most of the misclassified individuals were artisans. Of the 124 observations misclassified as self-employed, 104 had artisan occupations such as blacksmithing and leather tanning. Although these occupations had fairly low rates of self-employment in 1910, they likely had high rates of self-employment in the middle of the nineteenth century. Including these occupations in the self-employed category leads to the overstatement of self-employment in 1910, but excluding them would lead to the understatement of self-employment in 1850.

This issue together with the general problem that self-employment is only a weak indicator of entrepreneurial activity leads us to consider as well a subset of the self-employment occupations. This subset includes all occupations in the following categories: contractors, manufacturers, proprietors of service establishments, rentiers, and trade occupations. We refer to these occupations as *employer occupations* because in the 1910 census more than half of the individuals in these occupations were reported as "employers."[28] For many of these, the occupation title alone indicates self-employment. But more importantly, these occupations seem to have greater potential for meeting the Schumpeterian definition of entrepreneurship than other self-employment occupations.

Entrepreneurship and the Distribution of Wealth

Table 3 presents for Massachusetts for 1850 to 1910, the population and wealth shares of the self-employed in non-agricultural occupations and the subset of employer occupations. The fraction of the population self-employed in non-agricultural occupations was essentially the same in 1910 as it had been in 1850. However, between 1860 and 1900 this fraction declined, primarily owing to a decline in some traditional artisan occupations. In contrast, the fraction of the population in employer occupations increased its population share by 50 percent between 1850

Table 3
Population and wealth shares of the self-employed, Massachusetts, 1850–1910.

	Self-employed in non-agricultural occupations		Employer occupations	
	% pop	% wealth	% pop	% wealth
1850	37.4	47.6	9.7	31.5
1860	42.2	45.9	10.1	26.3
1870	37.1	50.7	11.7	29.9
1880	33.1	40.6	9.2	16.3
1900	32.7	50.6	13.0	42.9
1910	36.7	68.3	14.6	45.9

and 1910. This increase did not come steadily, however. The big jump came between 1880 and 1900. In view of the occupations in this category, this jump is at least suggestive of a surge in entrepreneurial activity in Massachusetts in the last two decades of the nineteenth century. These decades certainly seem to be the period with the highest rates of entry into the manufacturing industry. Between 1870 and 1880, the number of manufacturing establishments in the state increased from 13,212 to 14,352. But by 1890 the number of establishments had jumped to 26,923, and by 1900 it had increased further to 29,180 (U.S. Census Office 1883, p. 5; U.S. Census Office 1902, p. 359).

Just as is observed today, the self-employed held a disproportionate share of wealth. The concentration of wealth among all self-employed in non-agricultural occupations was growing over the period. Although the population share of this group was approximately the same in 1910 as it had been in 1850, its wealth share increased from 48 percent to 68 percent. The share of wealth held by the employer occupations also grew over the sample period, but this growth was proportional to the growth in the group's population share, indicating that the average share of wealth held by each individual in this group did not change over this period. In other words, the wealth gap between these employer occupations and the rest of the population apparently was not widening between 1850 and 1910. Once again, however, comparing just the endpoints obscures some interesting trends. The concentration of wealth among the self-employed declined between 1870 and 1880 and then experienced a big jump between 1880 and 1900.

The overall picture that emerges from table 3 is that self-employment was strongly associated with wealth accumulation in postbellum

Table 4
The self-employed in the top of the wealth distribution, Massachusetts, 1850–1910.

	Share of top 20%		Share of top 5%	
	Self-employed non-ag. occ.	Employer occ.	Self-employed non-ag. occ.	Employer occ.
1850	33.7	17.4	45.8	34.0
1860	44.7	21.0	56.9	34.0
1870	46.9	19.8	51.7	35.1
1880	41.1	17.1	54.2	24.5
1900	38.2	22.4	61.5	53.6
1910	48.6	32.1	72.6	53.9

Massachusetts. This picture can be brought more sharply into focus by examining self-employment in the top fractiles of the wealth distribution. Table 4 presents the fractions of self-employed and employer occupations in the top 20 percent and the top 5 percent of the wealth distribution for Massachusetts from 1850 to 1910. The self-employed—particularly the employer occupations—were over-represented in the wealthiest groups. In 1850, when employer occupations accounted for 10 percent of the total population, they accounted for 17 percent of the top 20 percent of the wealth distribution and 34 percent of the top 5 percent of the wealth distribution. Over time, the shares of these occupations in the wealthiest groups grew. By 1910, employer occupations accounted for almost one-third of the individuals in the top 20 percent and more than half of those in the top 5 percent.

The changes in the population and wealth shares of the self-employed between 1850 and 1910 were accompanied by changes in the occupational composition of self-employment. Table 5 presents the distribution of the self-employed in non-agricultural occupations across the occupational categories described above. The top panel presents the occupational distribution of all self-employed and the bottom panel presents the occupational distribution of the self-employed in the top 20 percent of the wealth distribution. The changes in the occupations of the self-employed are reflective of the changes in the Massachusetts economy in the late nineteenth century. Artisans made up the largest group of self-employed in all years, but their share—especially among the wealthiest self-employed—was declining. The fraction of professionals (doctors, lawyers, veterinarians, etc.), however, rose. This rise was particularly pronounced among the self-employed in the wealthiest 20 percent of the population. The other

Table 5
Percentages of self-employed in non-agricultural occupations, Massachusetts, 1850–1910.

	Artisans	Contractors	Manufac-turers	Personal service	Professional	Proprietors Service ind.	Rentiers	Trade
All self-employed								
1850	67.7	0.0	3.1	0.8	5.6	0.6	0.5	21.8
1860	71.3	0.0	2.4	0.6	4.3	0.4	0.0	21.1
1870	61.5	0.0	4.2	0.4	6.6	0.7	0.0	26.6
1880	62.8	0.5	3.0	2.0	7.4	1.4	0.0	23.0
1900	48.3	1.4	4.4	3.3	8.6	3.3	3.1	27.6
1910	48.5	2.6	2.9	2.8	9.1	2.9	7.8	23.4
Self-employed in top 20% of wealth distribution								
1850	42.4	0.0	4.3	0.0	5.8	1.5	1.4	44.6
1860	43.6	0.0	2.3	0.0	9.4	0.8	0.0	43.9
1870	45.0	0.0	8.9	0.0	12.8	0.0	0.0	33.3
1880	46.7	0.0	7.2	0.0	11.5	1.2	0.0	33.5
1900	27.9	1.8	7.0	0.0	13.4	4.6	6.4	39.0
1910	17.9	8.4	4.8	1.4	14.6	0.6	18.9	33.3

groups of self-employed that grew over this period were contractors, personal service workers, and proprietors of service industry establishments (restaurants, hotels, bars, etc.). These increases reflect the rapid urbanization of this period, which led to a construction boom as well as to the commercialization of many tasks (such as laundry and meal preparation) that were formerly done within the home. Also increasing was the number of individuals with rentier occupations. This increase itself is an indicator of the expansion of rent-seeking activities in the economy. As observed in the earlier tables, the period of greatest change was between 1880 and 1900.

Tables 3–5 indicate that the nature and scale of self-employment—and perhaps the nature and scale of entrepreneurial activity—changed in late-nineteenth-century Massachusetts. But how did these changes affect the distribution of wealth? First, we consider the distribution of wealthholdings among the self-employed. Figure 3 plots the Theil entropy measures of wealth inequality for the two groups of self-employed. For both groups, wealth inequality appears to have fallen between 1850 and 1860 and then increased between 1880 and 1900. The increase between 1880 and 1900 is consistent with the changes observed in the earlier tables for this period. As self-employment activities became more diverse, the dispersion of wealthholdings among the self-employed increased.

Of primary interest, though, is how, if at all, did the changes in self-employment contribute to the increase in wealth inequality in Massachusetts between 1870 and 1900? The implicit counterfactual question

Figure 3
Theil entropy measures for self-employment categories.

is: what would have happened to aggregate wealth inequality in the absence of these changes in self-employment? A true counterfactual is not feasible, but we can take a more mechanical approach to addressing this question by examining decompositions of the change in the Theil entropy measure. First, it is useful to discuss the most common decomposition of the Theil measure into within-group and between-group inequality. For any exhaustive collection of mutually exclusive subsets of observations $1, 2, \ldots, G$, the Theil measure can be rewritten as

$$T = \sum_{g=1}^{G} \frac{n_g \mu_g}{n\mu} T_g + \sum_{g=1}^{G} \frac{n_g \mu_g}{n\mu} \ln\left(\frac{\mu_g}{\mu}\right), \tag{3}$$

where n_g represents the number of observations in sub-group g, μ_g represents the mean wealth of sub-group g, and T_g represents the measure in equation 1 calculated for sub-group g. The first term on the right-hand side of equation 2 is the weighted sum of the Theil entropy measures for the sub-group wealth distributions where the weights are the sub-group shares of total wealth. This term represents the component of measured inequality due to inequality in the distribution of wealth *within* population sub-groups. The second term is simply the Theil entropy measure of equation 1 calculated from a wealth distribution in which each person is assigned the mean wealth of their sub-group, and, therefore, represents the component of measured inequality due to inequality in the distribution of wealth *between* population sub-groups.

Examination of equation 3 reveals that changes in the Theil entropy measure can arise from changes in three factors: the population shares of sub-groups (n_g/n), the relative mean wealth of subgroups (μ_g/μ), and the dispersion of wealth within subgroups (T_g). The change in the Theil entropy measure between two periods may be decomposed into the contributions of these three factors. The contributions of each of these elements to the change between two periods, s and t, can be calculated as follows:

$$\Delta T_n^{s,t} = \sum_{g=1}^{G} \left[\left(\frac{n_g^t}{n^t}\right) - \left(\frac{n_g^s}{n^s}\right) \frac{\mu_g^t}{\mu^t} T_g^t \right] + \sum_{g=1}^{G} \left(\frac{n_g^t}{n^t} - \frac{n_g^s}{n^s}\right) \frac{\mu_g^t}{\mu^t} \ln\left(\frac{\mu_g^t}{\mu^t}\right),$$

$$\Delta T_\mu^{s,t} = \sum_{g=1}^{G} \left(\frac{\mu_g^t}{\mu^t} - \frac{\mu_g^s}{\mu^s}\right) \frac{n_g^s}{n^s} T_g^t + \sum_{g=1}^{G} \left[\frac{\mu_g^t}{\mu^t} \ln\left(\frac{\mu_g^t}{\mu^t}\right) \frac{\mu_g^s}{\mu^s} \ln\left(\frac{\mu_g^s}{\mu^s}\right) \right] \frac{n_g^s}{n^s}.$$

Table 6
Decompositions of change in Theil entropy measure, 1870–1900.

	Self-employed in non-agricultural occupations	Employer occupations
Total change in T	0.534	0.534
Change in T due to		
change in population share of self-employed	−0.098	0.101
change in relative mean wealth of self-employed	0.128	−0.060
change in within group inequality	0.504	0.493
in inequality among self-employed	0.216	0.045
in inequality among rest of population	0.287	0.448

Such decompositions allow us to quantify the contributions to the overall rise in inequality of changes in the fraction of the population who were self-employed, changes in the relative wealth of the self-employed, and the increase in wealth inequality within the ranks of the self-employed. It is important to note, however, that these decomposition are simply mathematical relationships that ignore interactions between the different components. For instance, they ignore the possibility that the change in the population share of the self-employed may have had an effect on the distribution of wealth within the self-employed group or even within the non-self-employed group. We will return to this point below.

Table 6 presents decompositions of the change in the Theil measure over the period of rising aggregate inequality: 1870–1900.[29] The decompositions are calculated twice, defining the self-employed first as all self-employed in non-agricultural occupations and then as only the employer occupations. Between 1870 and 1900, the Theil entropy measure rose by 0.534, an increase of more than 30 percent. Only a small portion of this change can be attributed to changes in the population share of the self-employed and changes in the wealth gap between the self-employed and the rest of the population. The signs of these effects differ in the two decompositions. The decrease in the population share of the self-employed in non-agricultural occupations decreased aggregate inequality whereas the increase in the population share of employer occupations increased aggregate inequality; the wealth gap between the self-employed in non-agricultural occupations and the rest of the population grew whereas that between employer occupations and the rest of the population decreased slightly. But in both

decompositions, the composition and relative mean wealth effects essentially cancel each other out.

Both decompositions pin the increase in aggregate inequality on the increase in inequality within population subgroups. The decomposition using the broader definition of self-employment indicates that even if the population share and relative mean wealth of the self-employed had remained constant between 1870 and 1900, the increase in wealth inequality within population groups would have led to an increase in the Theil measure of 0.504 (29 percent). The decomposition using the narrower definition of self-employment indicates that the increase in within-group inequality alone would have led to a 28 percent increase in the Theil measure.

Some of this increase was due to the growing dispersion in the distribution of wealth among the self-employed found in figure 3. But inequality was also increasing in the non-self-employed population. The last two rows of table 6 break down the within-group effect into the separate effects of changes in wealth inequality within the two population groups. Both decompositions indicate that the changes in the distribution of wealth within the non-self-employed population had a larger effect on aggregate inequality than the changes in the distribution of wealth within the self-employed population. The increase in inequality among all self-employed in non-agricultural occupations did account for more than 40 percent of the increase in within-group inequality between 1870 and 1900. But the increase in inequality within the employer occupations accounted for less than 10 percent of the overall increase in within-group inequality.

The results of the decompositions are not entirely consistent with predictions of the models of entrepreneurship with imperfect credit markets. The decomposition using the broader definition of self-employment do support the two predictions stated above: the wealth gap between the self-employed and the rest of the population did increase as did inequality within the self-employed. But the decomposition using the narrower definition of employer occupations indicates a narrowing wealth gap and only a small effect of rising inequality within the self-employed group. In view of the closer association of the employer occupations to entrepreneurial activities, this decomposition might have been expected to fit better the features of the models. Most significantly, however, both decompositions indicate that an important factor in the growing concentration in wealth in late nineteenth century Massachusetts was the increase in inequality within the

non-self-employed population—a finding not anticipated by the standard models of entrepreneurship with imperfect credit markets.

Discussion

The linked data sets indicate that, just as today, entrepreneurs held a disproportionate share of wealth in nineteenth century Massachusetts. The data also suggest that entrepreneurial activity was increasing in the last decades of the century just as industry was expanding and wealth inequality was rising. But a closer analysis of the changes in the wealth distribution during this period indicate that the rise in inequality had more to do with what was happening within the non-entrepreneur population than what was happening within the entrepreneur population. Much of the rise in inequality in Massachusetts between 1870 and 1900 was due to growing dispersion in wealthholdings within the non-entrepreneurial population. Granted, this population is by its nature very heterogeneous. But that heterogeneity, at least in terms of wealthholdings, was growing in the last decades of the nineteenth century. Even when we divide this population into more narrowly defined occupational categories, we still observe growing inequality within groups. Figure 4 plots the Theil entropy measures for non-self-employed unskilled, skilled and white-collar workers between 1850 and 1910. Inequality for all of these groups increased between 1870 and 1900. For the unskilled and white-collar workers, this increase represented a reversal of a downward trend in

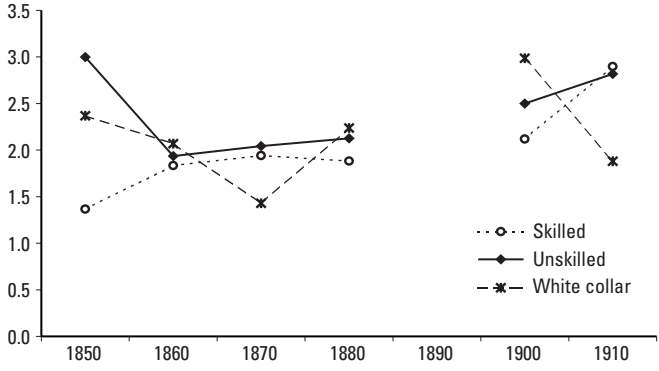


Figure 4
Theil entropy measures for non-self-employed population by occupational categories.

inequality from 1850 to 1870. But inequality among non-self-employed skilled workers was increasing throughout the sample period. The rise in inequality between 1870 and 1900 in Massachusetts was driven by growing dispersion in the wealthholdings of men with similar skills.

Here we find a strong parallel to the rise in inequality in the last few decades. Some of the growing disparity is due to changes in the returns to skills and education that have increased the resource gaps between groups. But much is due to rising inequality within fairly narrowly defined populations. For instance, Gottschalk (1997, p. 28) found that even controlling for race, education, experience, and geographic region, within-group inequality accounted for 50 percent of the rise in wage inequality among males and 23 percent of the rise in wage inequality among females between 1973 and 1994. Much of the literature on trends in inequality—be it wage inequality or wealth inequality—focuses on changes in the distribution of resources between groups. But the evidence indicates that changes in the distribution of resources within groups contributes greatly to trends in inequality. Clearly, more attention needs to be devoted to understanding the factors that lead to changes in the distribution of resources within groups.

The answer may yet lie with entrepreneurial activity. Decompositions of the Theil entropy measure do not represent conclusive tests of the links between entrepreneurship and rising wealth inequality in postbellum Massachusetts. As noted above, such decompositions ignore the possibility of interactions between different components. For instance, they ignore the possibility that changes in self-employment could have affected the distribution of wealth in the non-self-employed population. Changes in self-employment may change the distribution of inheritances or altered patterns of occupational mobility across and within generations. Entrepreneurial activity may also change the dispersion in the returns to particular skills in the economy. Workers in entrepreneurial firms, for instance, may receive higher wages or experience greater wage growth than other workers. Not all of the individuals who made their fortunes during the "dot com" phenomenon of the recent past were the entrepreneurs; many were the employers in those firms. In the late nineteenth century, wages for unskilled workers were higher in larger and more capital-intensive firms (Atack, Bateman, and Margo 2000). Understanding these phenomena may provide greater insight into the links between entrepreneurship and inequality.

Acknowledgments

The authors benefited greatly from the comments and suggestions of William Gentry, Kevin Hassett, Doug Holtz-Eakin, and the participants in the Maxwell Policy Research Symposium on Entrepreneurship and Public Policy.

Notes

1. See e.g. Aghion and Bolton 1997.

2. Steckel (1994) provides details on sampling procedures, additional characteristics of the samples, detailed definitions of occupations, information on the collection of taxes, and comparisons with wealth reported in the censuses of 1850, 1860, and 1870.

3. Nearly all the schedules of the 1890 census were destroyed in a fire. Therefore, there is no sample for that year. The sample sizes reflect our evaluation of the tradeoffs between costs of data collection and the sensitivity of results in small samples to outliers in the wealth distribution. In a judgment call, it was felt that roughly 600 observations in each of rural and urban areas would be adequate to depict and analyze the wealth distribution in a particular census year.

4. The tax lists were compiled twice a year (late spring and late fall), and the list prepared closest to the date of the census was used.

5. Initially mechanics' tools were exempted to an unlimited value. A $300 limit was imposed sometime after 1875 but rescinded in 1931. See Street 1863, p. 217; Commonwealth of Massachusetts 1875, p. 153; Commonwealth of Massachusetts 1902, p. 6; Nichols 1938, p. 253.

6. This exclusion is justified not only by the difficulty in assessing the value of such items but also by the fact that these items are not readily converted to cash. If the interest in wealth comes from its role as a source of potential consumption, as Edward Wolff argues (1994, p. 144), then wealth should be measured as the value of fungible assets.

7. For instance, the aggregate value of notes secured by mortgages of taxable real estate was estimated to be $48 million in 1881. When such notes became tax exempt in 1882, total personal property assessed in the state fell by only $3.6 million (Bullock 1916, p. 21).

8. Other problems associated with measures of taxable wealth relate to methods of tax assessment. In many jurisdictions, tax assessments were reevaluated infrequently. In Massachusetts, though, new valuations were prepared annually from lists of taxable property submitted by property owners, reducing the distortions of obsolete property evaluations that might occur in times of rapid changes in asset prices (Bullock 1909; Huse 1916). Another issue is that of underassessment. Often, assets are assessed for tax purposes at values greatly below their market values. Such underassessment poses a problem for the analysis of the distribution of wealth, however, only if the degree of underassessment varies over time and across types of assets.

9. Female household heads accounted for approximately 10% of the observations in the full sample for each census year. The property of female household heads was subject to

different tax exemptions than that of male household heads. Accordingly, the taxable wealth of male heads and female heads are not directly comparable in these data.

10. For an excellent discussion of the theory and application of these and other inequality measures, see Foster 1985.

11. This is often referred to as the "Theil T." As will be shown below, this measure weights population groups by their wealth shares. Theil also proposed an alternative measure known as the "Theil L" which weights population groups by their population shares. The Theil L is only defined for distributions with no non-zero observations, however, and therefore cannot be used with the taxable wealth data.

12. Asymptotic approximations of the variances of the Gini coefficient and the Theil entropy measure do exist, but little is known of their small sample properties. Statistical inference based on bootstrap methods has been shown to be superior to asymptotic approximations both on theoretical grounds and in a variety of applications. See Mills and Zandvakili 1997.

13. For more information on the theory and application of bootstrapping, see Efron and Tibshirani 1993.

14. These tests were conducted by using bootstrap analysis to calculate approximate standard errors and confidence intervals for the *difference* in each of the measures between periods.

15. The results are taken from various years of the Commonwealth of Massachusetts Aggregate of Polls, Property, Taxes, Etc.

16. Moreover, the variation across towns in the fraction with no match in the tax records is very similar to the variation across towns in the fraction of males assessed for the poll tax only. For example, in 1900, the fraction of males assessed for the poll tax only was 30.1% in Westminster and 90.0% in Boston. For the same year, the 32.4% of the individuals from Westminster and 90.5% of the individuals in Boston in the census sample had no match in the tax records.

17. Steckel (1994) uses scatter diagrams and regressions to compare census wealth with taxable wealth for the 1850, 1860, and 1870. In the case of discrepancies, census wealth often exceeded taxable wealth, but the differences were not systematically associated with socioeconomic variables, such as occupation or age, that were reported by the census. There are several plausible explanations for the differences, including assessments below market value, exemptions, and inclusion of property owned by the spouse or children in census wealth. However, the differences in the Gini coefficients calculated from the census and tax data for male household heads are small (< 0.02) and not statistically different from 0.

18. Data on tax rates by jurisdiction are available in the Commonwealth of Massachusetts Aggregate of Polls, Property, Taxes, Etc.

19. It would also be interesting to study the connection between entrepreneurship and the antebellum rise in wealth inequality. Unfortunately, as will be described below, the limited information available in the pre-1850 censuses precludes such an investigation.

20. See for example Aghion and Bolton 1997.

21. Data on the amount of capital invested in manufacturing is available in published volumes of the Census of Manufacturing. (See U.S. Census Office 1883, 1902.) These data

were converted to constant dollars using Composite Consumer Price Index presented in McCusker 1992.

22. Data on the value of wealth assessed in the state are available in the Commonwealth of Massachusetts Aggregate of Polls, Property, Taxes, Etc. These data were converted to constant dollars again using Composite Consumer Price Index presented in McCusker 1992.

23. See Evans and Leighton 1989; Fairlie 1996; Blanchflower and Oswald 1998. Alternative definitions of entrepreneurship are based on business ownership. Holtz-Eakin, Joulfaian, and Rosen (1994) define as entrepreneurs individuals who filed schedule C ("Profit or Loss from Business (Sole Proprietorship)") on their federal tax returns. Gentry and Hubbard define entrepreneurial households as households which own one or more active businesses with a total market value of at least $5,000.

24. This industry data was collected in both the 1820 and 1840 censuses. The 1830 census collected no information on market activity.

25. Studies of self-employment in the current period differ in how they treat agricultural occupations. Some exclude farmers from the self-employed category, arguing that the determinants of entry into farming seem to be quite different than the determinants of entry into other types of self-employment. See e.g. Fairlie 1996.

26. The Integrated Public Use Microdata Series (IPUMS) is a collection of national random samples of households drawn from the federal censuses. Information on the IPUMS data is available at http://www.ipums.umn.edu/.

27. Atack (1985) provides data on the shares of industry value added produced by different types of firms for both 1850 and 1870. Artisanal shops were defined as establishments with 1–6 employees and no inanimate power source.

28. The only exceptions to this are the rentier occupations. Data on employment status was generally missing for these occupations because they were considered "non-occupational" responses by the census.

29. We also performed decompositions of the changes between 1880 and 1900—the period of the most dramatic changes in self-employment. The results of these decompositions reveal the same patterns as the 1870–1900 decompositions.

References

Aghion, Philippe, and Patrick Bolton. 1997. A theory of trickle-down growth and development. *Review of Economic Studies* 64: 157–172.

Atack, Jeremy. 1985. Industrial structure and the emergence of the modern industrial corporation. *Explorations in Economic History* 22: 29–52.

Atack, Jeremy, Fred Bateman, and Robert A. Margo. 2000. Rising Wage Dispersion across American Manufacturing Establishments, 1850–1880. Working paper 7932, National Bureau of Economic Research.

Banerjee, Abhijit, and Andrew Newman. 1993. Occupational choice and the process of development. *Journal of Political Economy* 101: 274–298.

Blanchflower, David G., and Andrew J. Oswald. 1998. What makes an entrepreneur? *Journal of Labor Economics* 16: 26–60.

Bullock, Charles J. 1909. *The General Property Tax in the United States*. International Tax Association.

Bullock, Charles J. 1916. The taxation of property and income in Massachusetts. *Quarterly Journal of Economics* 31: 1–61.

Commonwealth of Massachusetts. Various years. Aggregate of Polls, Property, Taxes, Etc.

Commonwealth of Massachusetts. 1875. *Report of the Commissioners Appointed to Inquire into the Expediency of Revising and Amending the Laws Relating to Taxation and Exemption Therefrom*. House Doc. No. 15. Boston: Wright & Potter, State Printer.

Commonwealth of Massachusetts. 1902. *Chapter 12 of the Revised Laws, Regulating Taxation by the Local Assessor in Massachusetts, Including Statutes Relating to the Collection of Taxes*. Tax Doc. No. 1. Boston: Wright & Potter Printing Company, State Printers.

Conley, Timothy G., and David W. Galenson. 1994. Quantile regression analysis of censored wealth data. *Historical Methods* 27: 149–165.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.

Ely, Richard T. 1888. *Taxation in American States and Cities*. Crowell.

Evans, David S., and Boyan Jovanovic. 1989. An estimated model of entrepreneurial choice under liquidity constraints. *Journal of Political Economy* 97: 808–827.

Evans, David S., and Linda S. Leighton. 1989. Some empirical aspects of entrepreneurship. *American Economic Review* 79: 519–535.

Fairlie, Robert W. 1996. *Ethnic and Racial Entrepreneurship: A Study of Historical and Contemporary Differences*. Garland.

Foster, James. E. 1985. Inequality measurement. *Proceedings of Symposia in Applied Mathematics* 33: 31–68.

Gentry, William M., and R. Glenn Hubbard. 2000. Entrepreneurship and Household Saving. Working paper 7894, National Bureau of Economic Research.

Gottschalk, Peter. 1997. Inequality, income growth, and mobility: The basic facts. *Journal of Economic Perspectives* 44: 21–40.

Holtz-Eakin, Douglas, David Joulfaian, and Harvey S. Rosen. 1994. Sticking it out: Entrepreneurial survival and liquidity constraints. *Journal of Political Economy* 102: 53–75.

Huse, Charles P. 1916. *The Financial History of Boston from May 1, 1822, to January 31, 1909*. Harvard University Press.

Kuznets, Simon. 1955. Economic growth and income inequality. *American Economic Review* 45: 1–28.

McCusker, John J. 1992. *How Much Is That in Real Money? A Historical Price Index for Use as a Deflator of Money Values in the Economy of the United States*. American Antiquarian Society.

Mills, Jeffrey A., and Sourushe Zandvakili. 1997. Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics* 12: 133–150.

Nichols, Philip. 1938. *Taxation in Massachusetts: A Treatise on the Assessment and Collection of Taxes, Excises, and Special Assessments under the Laws of the Commonwealth of Massachusetts*. Third Edition. Financial Publishing Co.

Quadrini, Vincenzo. 2000. Entrepreneurship, saving, and social mobility. *Review of Economic Dynamics* 3: 1–40.

Schumpeter, Joseph A. 1947. The creative response in economic history. *Journal of Economic History* 7: 149–159.

Shammas, Carole. 1993. A new look at long-term trends in wealth inequality in the United States. *American Historical Review* 98: 412–431.

Steckel, Richard H. 1994. Census manuscript schedules matched with property tax lists: A source of information on long-term trends in wealth inequality. *Historical Methods* 27: 71–85.

Steckel, Richard H., and Carolyn M. Moehling. 2001. Rising inequality: Trends in the distribution of wealth in industrializing New England. *Journal of Economic History* 61: 160–183.

Street, Alfred Billings. 1863. *A Digest of Taxation in the States*. Albany: Weed, Parsons.

U.S. Census Office. 1883. *Report on the Manufactures of the United States at the Tenth Census*. Government Printing Office.

U.S. Census Office. 1902. *Twelfth Census of the United States, Taken in the Year 1900. Manufactures. Part II: States and Territories*. Government Printing Office.

Williamson, Jeffrey G., and Peter H. Lindert. 1980. *American Inequality: A Macroeconomic History*. Academic Press.

Wolff, Edward N. 1994. Trends in household wealth in the United States, 1962–83 and 1983–89. *Review of Income and Wealth* 40: 143–174.

Index