The Journal of Economic Perspectives

A journal of the
American Economic Association

Spring 2013

The Journal of
Economic Perspectives

A journal of the American Economic Association

Editorial offices:
Journal of Economic Perspectives
American Economic Association Publications
2403 Sidney St., #260
Pittsburgh, PA 15203
email: jep@jepjournal.org

The *Journal of Economic Perspectives* gratefully acknowledges the support of Macalester College.
Registered in the US Patent and Trademark Office (®).

Composed by American Economic Association Publications, Pittsburgh, Pennsylvania, USA
Printed by R. R. Donnelley Company, Jefferson City, Missouri, 65109, USA
No responsibility for the views expressed by the authors in this journal is assumed by the editors or by the American Economic Association.
THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Spring 2013, Vol. 27, No. 2. The *JEP* is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $20.00, $30.00, or $40.00 depending on income; for an additional $15.00, you can receive this journal in print. E-reader versions are free. For details and further information on the AEA go to http://www.vanderbilt.edu/AEA/.
Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203.   Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 27 • Number 2 • Spring 2013

Symposia

The Growth of the Financial Sector

Early and Later Interventions

Articles

Features

Statement of Purpose

The *Journal of Economic Perspectives* attempts to fill a gap between the general interest press and most other academic economics journals. The journal aims to publish articles that will serve several goals: to synthesize and integrate lessons learned from active lines of economic research; to provide economic analysis of public policy issues; to encourage cross-fertilization of ideas among the fields of economics; to offer readers an accessible source for state-of-the-art economic thinking; to suggest directions for future research; to provide insights and readings for classroom use; and to address issues relating to the economics profession. Articles appearing in the journal are normally solicited by the editors and associate editors. Proposals for topics and authors should be directed to the journal office, at the address inside the front cover.

Policy on Data Availability

It is the policy of the *Journal of Economic Perspectives* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Details of the computations sufficient to permit replication must be provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the above requirements cannot be met.

Policy on Disclosure

Authors of articles appearing in the *Journal of Economic Perspectives* are expected to disclose any potential conflicts of interest that may arise from their consulting activities, financial interests, or other nonacademic activities.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 3–28

The Growth of Finance[†]

Robin Greenwood and David Scharfstein

During the last 30 years, the financial services sector has grown enormously. This growth is apparent whether one measures the financial sector by its share of GDP, by the quantity of financial assets, by employment, or by average wages.

At its peak in 2006, the financial services sector contributed 8.3 percent to US GDP, compared to 4.9 percent in 1980 and 2.8 percent in 1950. The contribution to GDP is measured by the US Bureau of Economic Analysis (BEA) as value-added, which can be calculated either as financial sector revenues minus nonwage inputs, or equivalently as profits plus compensation. Figure 1, following the methodology of Philippon (2012) and constructed from a variety of historical sources, shows that that the financial sector share of GDP increased at a faster rate since 1980 (13 basis points of GDP per annum) than it did in the prior 30 years (7 basis points of GDP per annum).[1] The growth of financial services since 1980 accounted for more than a quarter of the growth of the services sector as a whole. Figure 1 shows

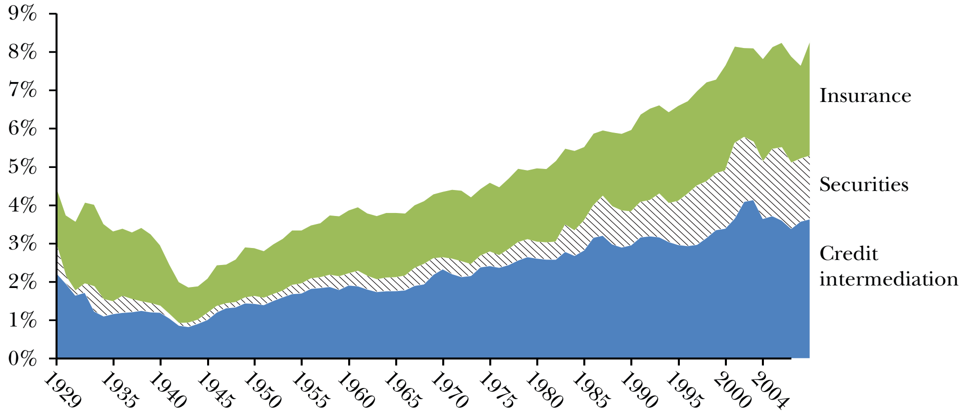
[1] Online Appendix Table 1, which is available with this article at http://e-jep.org, covers the period 1980–2007 and is based on the national income account published by the BEA. It shows the contribution to GDP of the industries comprising the financial services sector: securities, credit intermediation, and insurance. Details on all data sources and calculations are provided in the online Appendix.

■ *Robin Greenwood is George Gund Professor of Finance and Banking and David Scharfstein is Edmund Cogswell Converse Professor of Finance and Banking, both at Harvard Business School, Harvard University, Cambridge, Massachusetts. Greenwood and Scharfstein are Research Associates at the National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are rgreenwood@hbs.edu and dscharfstein@hbs.edu.*

[†] To access the Appendix and disclosure statements, visit http://dx.doi.org/10.1257/jep.27.2.3.

doi=10.1257/jep.27.2.3

Figure 1
The Growth of Financial Services
(value added share of GDP)



Source: Authors' calculations using data from National Income and Product Accounts (1947–2009) and the National Economic Accounts (1929–1947).
Notes: The finance sector includes the insurance, securities, and credit intermediation subsectors. The securities subsector includes the activities typically associated with investment banks and asset management firms, and it comprises two different categories in later sample years ("Securities" and "Funds, trusts, and other vehicles"); we combine them into one category for consistency.

that the securities and credit intermediation, subsectors of finance are responsible for the acceleration of financial sector growth since 1980; insurance, by contrast, has grown at a steady pace since the 1940s.

The growth of the financial sector is also evident in the growth of financial claims and contracts, including stocks, bonds, derivatives, and mutual fund shares. Drawing on the Flow of Funds Accounts published by the Federal Reserve, the value of total financial assets was approximately five times US GDP in 1980; by 2007, this ratio had doubled. Over the same period, the ratio of financial assets to tangible assets (like plant and equipment, land, and residential structures) increased as well. This growth was not simply the continuation of a trend that started in the 1950s; rather, something appears to have changed in the early 1980s.

The US economy was not the only one to experience dramatic growth in financial services. Other than the relatively small economy of Switzerland, where financial services play an outsized role, there is a group of English-speaking countries including the United States, Great Britain, and Canada that stand out for the share of their economy devoted to finance.

Workers in the financial sector have shared impressively in this growth: in 1980, the typical financial services employee earned about the same wages as his counterpart in other industries; by 2006, employees in financial services earned an average of 70 percent more (Phillipon and Reshef 2009). Attracted by high wages, graduates of elite universities flocked into the industry. In 2008, 28 percent of Harvard College graduates went into financial services, compared to only 6 percent between 1969 and 1973 (Goldin and Katz 2008). Graduates from the Stanford MBA program who entered financial services during the 1990s earned more than three times the wages of their classmates who entered other industries (Oyer 2008).

Has society benefited from the recent growth of the financial sector? There is a large literature dating back at least to Schumpeter (1911) that sees a vibrant financial sector as critical to capital allocation and economic growth. Seminal empirical contributions include Goldsmith (1969), King and Levine (1993), and Rajan and Zingales (1998), which document the relationship between financial development and growth in cross-country studies. It is natural to think therefore that the more recent period of financial development has also been economically beneficial. Yet, many are skeptical about its value, particularly in light of the recent financial crisis. Indeed, Rajan (2005), whose research has emphasized the value of financial development, famously called into question the value of more recent financial sector growth at a symposium of central bankers just before the financial crisis erupted. And Adair Turner (2010), the top financial regulator in the UK, has written: "There is no clear evidence that the growth in the scale and complexity of the financial system in the rich developed world over the last 20 to 30 years has driven increased growth or stability, and it is possible for financial activity to extract rents from the real economy rather than to deliver economic value." Similarly, Philippon (2012; see also Phillipon and Reshef in this issue) argues that the period of recent growth has come with a puzzling increase in the cost of financial intermediation.

In this paper, we try to shed light on these competing perspectives by first documenting the ways in which finance changed during the period from 1980 to 2007. We take this approach because surprisingly little is known about which activities contributed to the rapid growth of the financial sector. With a better understanding of how the financial sector changed, we provide some perspectives on the social benefits and costs of financial sector growth.

Our main finding is that much of the growth of finance is associated with two activities: asset management and the provision of household credit. The value of financial assets under professional management grew dramatically, with the total fees charged to manage these assets growing at approximately the same pace. A large part of this growth came from the increase in the value of financial assets, which was itself driven largely by an increase in stock market valuations (such as the price/earnings multiples). There was also enormous growth in household credit, from 48 percent of GDP in 1980 to 99 percent in 2007. Most of this growth was in residential mortgages. Consumer debt (auto, credit card, and student loans) also grew, and a significant fraction of mortgage debt took the form of home equity lines used to fund consumption (Mian and Sufi 2012). The increase in household credit contributed to the growth of the financial sector mainly through fees on loan origination, underwriting of asset-backed securities, trading and management of fixed income products, and derivatives trading.

Thus, any assessment of whether and in what ways society benefited from the growth of the financial sector depends in large part on an evaluation of professional asset management and the increase in household credit. In our view, the professionalization of asset management brought significant benefits. The main benefit was that it facilitated an increase in financial market participation and diversification, which likely lowered the cost of capital to corporations. Young firms benefited in particular, both because they are more reliant on external financing and because their value depends more on the cost of capital. At the same time, the cost of professional asset management has been persistently high. While the high price encourages more active asset management, it may not result in the kind of active asset management that leads to more informative securities prices or better monitoring of management. It also generates economic rents that could draw more resources to the industry than is socially desirable.

While greater access to credit has arguably improved the ability of households to smooth consumption, it has also made it easier for many households to overinvest in housing and consume in excess of sustainable levels. This increase in credit was facilitated by the growth of "shadow banking," whereby many different types of nonbank financial entities performed some of the essential functions of traditional banking, but in a less-stable way. The financial crisis that erupted late in 2007 and proved so costly to the economy was largely a crisis in shadow banking.

To develop these points we follow the US Bureau of Economic Analysis in breaking out the financial services sector into two subsectors: "securities" and "credit intermediation." We do not consider insurance, the other main subsector of financial services, because its steady growth is less of a puzzle.[2] The securities subsector (or "industry" in the terminology of the BEA) includes the activities typically associated with investment banks (such as Goldman Sachs) and asset management firms (such as Fidelity). These activities include securities trading and market making, securities underwriting, and asset management for individual and institutional investors. The credit intermediation industry performs the activities typically associated with traditional banking—lending to consumers and corporations, deposit taking, and processing financial transactions. After describing what drove the growth of these industries over the course of the 1980–2007 period, we evaluate the benefits and costs of this growth.

[2] Changes in the value added of insurance since 1980 have been driven mainly by a slight decline in life insurance revenues as a percentage of GDP and increases in property and casualty insurance and private health insurance. Property and casualty insurance tends to grow mechanically with the stock of tangible assets, as households insure more automobiles and larger and more expensive houses. The growth of private health insurance, while important for many reasons, is driven by factors outside the scope of this article.

The Growth of the Securities Industry

Components of Growth

Figure 1 shows that the growth of the securities industry accounts for almost half the overall (3 percentage point) growth of the financial sector relative to GDP from 1980–2007. In particular, the securities industry grew from 0.4 percent of GDP in 1980 to 1.7 percent of GDP in 2007, having peaked at 2.0 percent of GDP in 2001 during the Internet boom.

To get a better sense of the components of growth within the securities industry, ideally we would break out value added by activity. Unfortunately, there are no published data on the input costs at the activity-level needed to calculate value added. Instead, we use data on the output of the various activities of the securities industry. This output measure, calculated by the US Bureau of Economic Analysis for 1997 and 2002 and the US Census Bureau for 2007, is essentially the revenues of each of the activities of the industry. Detailed breakdowns are only available in these years. Later in this section we will discuss our own estimates of activity-level outputs for the complete 1980–2007 period. For the remainder of the paper, we focus more on industry output rather than on value added.

As Table 1 shows, in 2007, securities industry output was $676.1 billion, while value added was $241.2 billion. Asset management was by far the largest component of output, totaling $341.9 billion, well over four times its level in 1997. What we call asset management "output" includes fees from investment advisory and management services (the largest component), the administration of mutual and pension funds, and trust and custody services.

Table 1 shows that three revenue sources traditionally associated with investment banking—trading fees and commissions, trading gains, and securities underwriting fees—fell as a percentage of GDP between 1997 and 2007. These declines occurred despite a fourfold increase in stock-market trading. At the same time, two other activities grew substantially: brokering and dealing in debt products with 2007 output of $36 billion, and derivatives trading with output of $45 billion. Most of the revenues from derivatives trading appear to be associated with fixed income products, and as such, can be understood as a by-product of the growth of credit intermediation, which we discuss in the next section.[3] In 1997, the derivatives category was not even reported, suggesting that it was not significant enough to warrant its own category.

Panning Back to 1980

Because the Bureau of Economic Analysis does not provide detailed activity-level data prior to 1997, we use a variety of sources to break out securities industry

[3] For example, a large fraction of Goldman Sachs' derivatives revenues appear to be tied to fixed income trading. See http://fcic-static.law.stanford.edu/cdn_media/fcic-docs/0000-00-00%20Goldman%20Sachs%20Estimated%20Revenue%20Analysis.pdf.

Table 1
Value Added and Output from Securities Firms, Selected Years

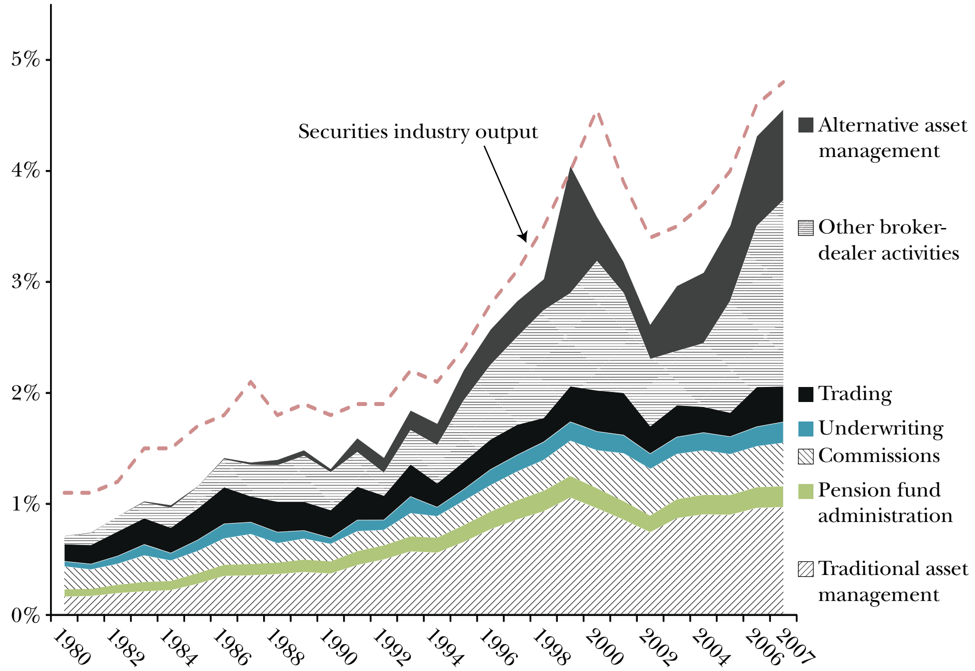
	$ billions			*% of GDP*		
Industry outputs, by activity	*1997*	*2002*	*2007*	*1997*	*2002*	*2007*
Asset management	82.8	199.2	341.9	0.99%	1.87%	2.43%
Fees and commissions from trading equities	55.6	57	74.1	0.67%	0.54%	0.53%
Trading gains	33.8	19	45.1	0.41%	0.18%	0.32%
Securities underwriting	28.3	22.1	35.1	0.34%	0.21%	0.25%
Profits from derivative contracts		16.3	45.3		0.15%	0.32%
Brokering and dealing debt products—debt instruments			36.5			0.26%
Management of financial market and clearing products			22.9			0.16%
Other broker-dealer revenue	18.4	40.6	56.2	0.22%	0.38%	0.40%
Other	2.6	1.7	19.0	0.03%	0.02%	0.14%
Total securities outputs	**221.5**	**355.9**	**676.1**	**2.66%**	**3.34%**	**4.81%**
By-products produced by securities firms (revenues collected by securities firms for other activities)	5.5	7.6	11.7	0.07%	0.07%	0.08%
Total inputs	**89.4**	**131.8**	**364.6**	**1.07%**	**1.24%**	**2.59%**
Revenues collected by nonsecurities firms for securities-related activities	9.4	52.8	82.1	0.11%	0.50%	0.58%
Value added by securities firms	**128.1**	**179.0**	**241.2**	**1.54%**	**1.68%**	**1.72%**
Value added for all securities-related activities	**129.2**	**206.4**	**284.0**	**1.55%**	**1.94%**	**2.02%**

Source: Bureau of Economic Analysis, Economic Census of the United States, and authors' estimates.
Notes: Asset Management consists of financial planning and investment management services, direct expenses associated with mutual funds and pension funds, and trust services. Other broker-dealer revenue includes brokering and dealing investment company securities, foreign currency, brokerage correspondent fees, and other fees. Missing cells indicate that the item was either zero or grouped into another category.

output back to 1980. Figure 2 shows annual estimates of the revenues from several key activities: traditional asset management (mutual funds, pension funds, and exchange-traded funds), alternative asset management (hedge funds, private equity, and venture capital), and a variety of broker-dealer activities (underwriting, customer trading, and proprietary trading). Although our estimates are imperfect and these categories do not correspond exactly to the product line outputs shown in Table 1, Figure 2 shows that we match the time-series of securities industry output reasonably well.

Fees earned from traditional asset management along with administration costs of pension funds are the largest component of output for the securities industry and are generally an increasing share of output until 1998. We estimate total fees using assets under management reported by the Investment Company Institute (ICI) and percentage fees reported by French (2008) and ICI. The largest

Figure 2
The Growth of the Securities Industry, 1980–2007
(revenues from different activities as a percent of GDP)



Source: Data are compiled by authors and described further in the text.
Notes: "Other broker-dealer activities" include revenues from derivatives and commodities trading, as well as other unclassified broker-dealer activities. Alternative asset management includes management of hedge funds, private equity, and venture capital. Traditional asset management includes management of mutual funds, money market funds, and exchange traded funds.

component of fees from traditional asset management comes from mutual funds (including money market mutual funds), which grew assets under management from $134 billion in 1980 to over $12 trillion in 2007. Fees on equity mutual funds dropped steadily during this period, from over 2 percent of assets to approximately 1 percent of assets, a decline largely driven by less use of mutual funds with up-front fees ("loads"). Absent the drop in loads, the average expense ratio would have risen slightly during this time, despite the increasing availability of low-fee index funds such as the Vanguard Standard & Poor's 500 mutual fund. Because percentage fees dropped slowly, total fees in each year were largely driven by the value of assets under management. For example, total fees fell in 2001 with the bursting of the Internet bubble, rose to hit their prior peak in 2004, and continued to grow thereafter. Overall, despite year-to-year fluctuations, there was enormous growth in fees from traditional asset management between 1980 and 2007.

The fees collected by alternative asset managers—hedge funds, private equity funds, and venture capital funds—also rose substantially over this period. Most of these funds charge a management fee of 1.5–2.5 percent of assets under management, plus "carried interest," a percentage of realized gains in the range of 15–25 percent. In most years, the combination of the management fee and carried interest is between 3 and 5 percent of assets under management, considerably higher than the fees charged by mutual funds. To compute aggregate fees collected by hedge funds, we apply percentage fees reported in French (2008) to the complete universe of US hedge funds, as reported by Hedge Fund Research. For private equity and venture capital, we use total fees reported by Kaplan and Rauh (2010), which we update to 2007 using data on assets under management provided by Thomson Financial.

Hedge fund, private equity, and venture capital fees were all near-zero in 1990 because assets under management were low. However, by 2007, approximately $854 billion of assets was managed by private equity firms, $258 billion by venture capital firms, and another $1.46 trillion by US-domiciled hedge funds. Hedge fund fees peaked at $69 billion in 2007. Fees for private equity and venture capital were more volatile, spiking in 1999 at $66 billion, driven by a record number of exits in both private equity and venture capital. In 2007, private equity fees were $26 billion and venture capital fees were $14 billion. Together, fees for these alternative investments are comparable to the $91 billion that was collected by mutual fund managers, who managed more than five times as many assets.

Our estimates of asset management fees are conservative because we do not capture growth in fees charged by investment advisors (although these are included in the data shown in Table 1). These services introduce another layer of fees on top of the management fees that go to traditional and alternative investment managers. We estimate that these advisors collect at least another $30–$40 billion of revenues not reflected in Figure 2.[4] Including these fees helps bridge the gap between the combined total of estimated management fees across investment vehicles (from hedge funds, mutual funds, and so on) and the revenue numbers for asset management reported by the US Bureau of Economic Analysis in 2007.

Combining the fees paid to traditional and alternative asset managers, the average fee has fluctuated between 1.1 and 1.6 percent of assets under management, with the exception of 1999, when venture capital exits took the average fee to 2.3 percent. In 2007, fees were 1.3 percent of assets under management. In short, although the composition of asset managers has changed over time—with high fee alternative asset managers gaining market share—the average fee paid to the industry per dollar of assets under management has not declined. French (2008)

[4] Historically, investment advisors charged commissions based on the number of trades they execute on behalf of their clients. However, a large number of advisors now mainly charge fees based on assets under management. For example, the US division of UBS Wealth Management reported income of $6.1 billion on end-of-year assets under management of $764 billion, implying a fee of 0.79 percent. In 2007, the total assets under management of investment advisors was approximately $3.6 trillion, suggesting another $30–$40 billion of revenues not reflected in Figure 2.

reaches this same conclusion. However, our estimates for total fees are higher than those reported by French (2008) because we also include fees earned by US asset managers for assets other than US-listed stocks.

All told, during the period 1980–2007, total asset management fees grew by 2.2 percentage points of GDP, which is over one-third of the growth in financial sector output. By contrast, drawing on data broker-dealers file with the Securities and Exchange Commission, Figure 2 shows that the other main activities of the securities industry—underwriting, trading, and commissions—do not appear to explain a significant share of growth in the securities industry and the financial sector.[5] However, these filings do reveal significant growth in a catchall miscellaneous category, "other," which showed large growth during the period. Based on the BEA and Census Bureau data in Table 1 it is reasonable to infer that the growth of this category is related to other unmeasured asset management fees (perhaps advisory fees as described directly above), as well as growth in fixed-income market-making and derivatives trading.

Since asset management fees as a percentage of assets did not fluctuate by much, what then explains the growth in these fees relative to GDP? This growth was driven by two factors: increases in the total outstanding amount of financial assets, and increases in the share of these assets that were professionally managed. We describe each of these changes below.

The bottom two series in Figure 3 show the value of traded equity and fixed income securities over time, both scaled by GDP, on the left y-axis. Taken together, these assets increased from 107 percent of GDP in 1980 to 323 percent of GDP by 2007. The figure shows that securities industry output (the dashed line, with values read off the right y-axis) closely tracks the total value of these assets.

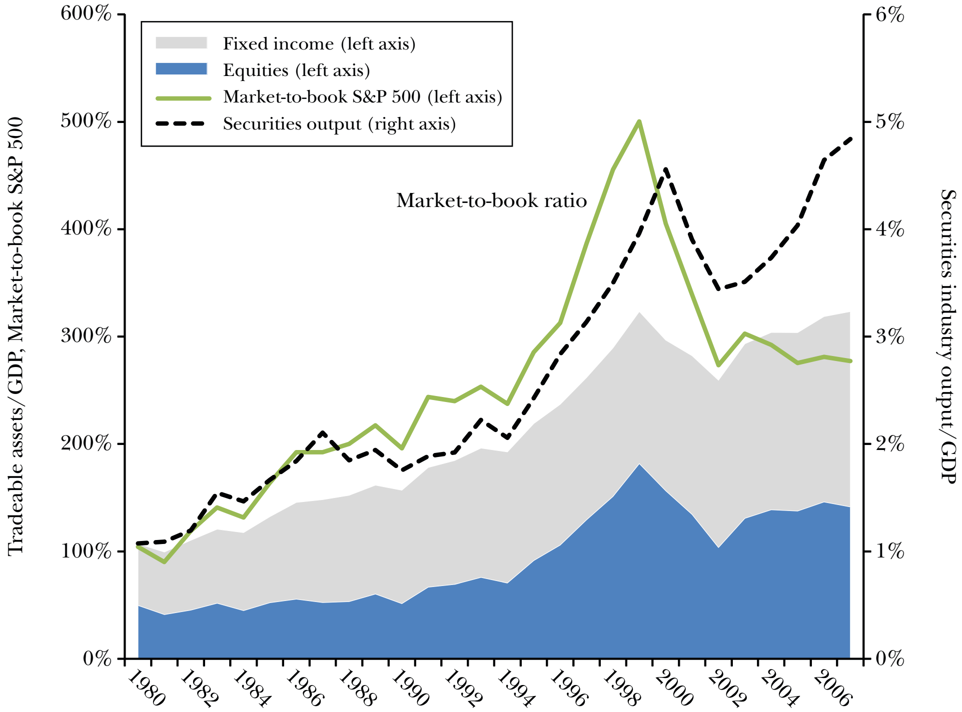
In fixed income, much of the growth came from securitization, whereby assets that were once held as illiquid loans on bank balance sheets were pooled into securities that could be traded and managed by professional investors. Fixed income securities grew from 57 percent of GDP in 1980 to 182 percent of GDP in 2007; approximately 58 percentage points of this growth came from securitization.[6]

In equities, much of the growth came from an increase in valuation ratios. Figure 3 shows that the value of publicly traded equity relative to GDP tracks the market-to-book ratio of the Standard and Poor's 500 (read off the left y-axis). Market capitalization of equities nearly tripled as a share of GDP between 1980 and 2007, growing from 50 percent to 141 percent of GDP. At the same time, the market-to-book ratio of the S&P 500 grew from 1.04 to 2.77 (from 104 to 277 percent on the graph),

[5] These filings are Financial and Operational Combined Uniform Single reports (commonly referred to as FOCUS reports).

[6] While fixed income assets increased dramatically, outside of hedge fund vehicles, the fees for managing fixed income assets are much lower than for equities and thus did not contribute much to the overall growth of asset management fees. Data provided by Greenwich Associates suggest that fees for active management of fixed income assets were 30 basis points in 2008, compared to 55 basis points for domestic equities and 66 basis points for international equities.

Figure 3
Tradable Assets and Securities Industry Output



Source: Flow of Funds Accounts of the United States, Bureau of Economic Analysis, and authors' estimates.
Notes: Figure 3 show the values of traded equity and of fixed income securities over time as a percentage of GDP (left axis); the market-to-book ratio of the Standard and Poor's 500 (left axis); and securities industry output as a percentage of GDP (right axis).

almost entirely explaining the growth. By contrast, the *book value* of equity of publicly-traded firms normalized by GDP was essentially flat during the same period.

In addition to increases in the amount of financial assets, the share of these assets under professional management has also increased. According to the Flow of Funds data from the Federal Reserve, 53 percent of household equity holdings were professionally managed in 2007, compared with only 25 percent in 1980. Lewellen (2011) reports that from 1980 to 2007, the share of US common stocks that were held by institutional investors increased from 32 percent to 68 percent of aggregate market capitalization. We do not have comparable statistics for the broader universe of fixed income assets, but the Flow of Funds suggests similar increases in the share of these assets that were professionally managed.[7]

[7] For example, direct household holdings of US Treasury bonds fell during this period from 14 percent of outstanding bonds to less than 1 percent.

Evaluation of the Growth of Professional Asset Management

The direct cost of professional asset management, at 1.3 percent of assets, is high. The present value of this fee paid over 30 years amounts to approximately one-third of the assets initially invested—a large price to pay a manager who does not outperform passive benchmarks. Moreover, paying managers as a percentage of assets under management rewards them when overall asset values rise, even if the manager does not outperform.[8] Indeed, as shown above, asset management fees during the 1980–2007 period rose in large part because valuation ratios increased.

Has society benefited from the growth of professional asset management despite these high fees? In the standard competitive model, the growth of an industry would seem to imply increased value to consumers and to society. But in the case of asset management, this implication does not follow immediately because of two important deviations from the competitive benchmark. The first deviation is that most of the potential benefits (and some of the costs) of professional asset management do not accrue directly to users. The second deviation is that many users have trouble assessing the quality and cost of professional asset management services or are influenced by agency considerations in choosing and compensating asset managers.

There are two related direct benefits of professional asset management: household participation in financial markets and diversification. Mutual funds, for example, enable individuals to buy a basket of securities in one transaction rather than construct a portfolio of securities through multiple transactions. Employer-based retirement plans also make it easier to participate and diversify. And, as Gennaioli, Shleifer, and Vishny (2012) point out, professional asset management facilitates participation to the extent that excessively risk-averse individuals trust professional asset managers (rightly or wrongly) to invest their money wisely.

According to modern finance theory, participation and diversification bring significant direct benefits to households. Participating in financial markets enables individuals to save and to earn a premium from holding risky assets—a premium that has historically been very high (Mehra and Prescott 1985). Diversifying enables individuals to more efficiently bear financial risk.

There is evidence that professional asset management has indeed increased household participation. During the 1980–2007 period of growth in asset management, the share of household financial assets held in marketable securities or mutual funds grew from 45 percent to 66 percent. According to the Survey of Consumer Finances, the percentage of households that owned stock increased from 32 percent

[8] An influential argument by Berk and Green (2004) might be interpreted as rationalizing the payment of fees as a percentage of assets. They suggest that active asset managers have the ability to outperform, but that this ability is scarce and increasingly difficult to achieve when a manager invests a larger portfolio of assets. Because the ability to outperform is scarce, in a competitive equilibrium, larger asset pools should pay higher dollar fees because they use up managers' outperformance ability. But this theory does not square with the facts. Active mutual fund managers underperform passive benchmarks even before netting out fees (Fama and French 2010).

in 1989 to 51 percent in 2007. There is also evidence that households increasingly diversified their portfolios. For example, holdings of foreign equities rose from 2 percent of US residents' portfolios in 1980 to 27.2 percent in 2007 (French 2008).

In theory, there is a positive externality from an increase in participation and diversification. Increasing households' willingness and capacity to take market risk should reduce investors' overall required rates of return. It is therefore possible—but hard to verify—that the growth of professional asset management was indirectly responsible for the large increase in stock market valuation ratios between 1980 and 2007 (Heaton and Lucas 1999; Fama and French 2002). This, in turn, may have led to a decline in the cost of capital to corporations. The greatest beneficiaries would have been young entrepreneurial firms—those most dependent on equity financing and whose values depend more on the cost of capital because of their more distant cash flows. Consistent with this interpretation, Fama and French (2004) show that young firms list their equity on the stock market at an increasing pace after 1979. The enhanced ability of young firms to go public could also help explain the growth of venture-capital backed entrepreneurship after 1980.

Much of professional asset management, however, is not explicitly directed at participation and diversification but rather at beating the market—that is, earning excess risk-adjusted returns or "alpha." Here the evidence on mutual fund performance strongly indicates that such active management is not directly beneficial to investors on average. Most studies document that active investment managers underperform, especially after taking into account fees. Fama and French (2010) show that mutual funds underperform passive benchmarks, even before taking out fees. Ibbotson, Chen, and Zhu (2011) suggest that hedge funds have produced modest alpha for their investors, but Jurek and Stafford (2011) point out that there is no alpha once returns are properly adjusted for tail risk. Of course, in the aggregate, there can be no outperformance of the market on average, since one investor's positive alpha must be another's negative alpha. Thus, beating the market cannot be a direct social benefit of professional management.[9]

However, from a social benefit perspective, the critical question is not whether active management leads investors to earn excess returns—it does not. Rather what matters is whether the *pursuit* of excess returns produces social benefits. One such benefit is more accurate ("efficient") securities prices, which enable firms to raise new capital at prices that better reflect their fundamental value. If prices are closer to fundamental value, firms have greater incentives to invest in the most productive projects, and to choose the appropriate scale of investment, thereby improving the economy's overall allocation of capital. One area in which information is particularly

[9] An exception is private equity and venture capital where alpha could come from improving firm performance rather than trading on information. The evidence is mixed on whether private equity and venture capital generate alpha. A recent study by Harris, Jenkinson, and Kaplan (2012) suggests that reporting bias has understated returns. In their study, private equity appears to generate consistently strong returns while venture capital does not. However, they do not adjust for risk and do not identify whether the returns come from improving firm performance or buying undervalued assets.

valuable is in the funding of start-up firms, where uncertainty and information asymmetries are large. Active asset managers—particularly venture capital firms, private equity firms, and hedge funds (and to a lesser extent mutual funds)—can also play a role in monitoring management to make sure that they are taking actions consistent with shareholder value maximization. Indeed, when venture capital firms fund new investments they typically have significant control over the firm, as do private equity investors involved in leveraged buyouts (Gompers 1995; Kaplan and Strömberg 2003, 2008). Hedge funds often pressure the boards of public companies to change corporate policies (Brav, Jiang, Partnoy, and Thomas 2008; Greenwood and Schor 2009), although there is some debate about whether such pressure actually enhances economic value.

Although it may be *socially* beneficial for active managers to acquire information and monitor firms, it is puzzling that they are able to attract funds despite their underperformance. There are few satisfying answers to explain why. The two most promising explanations stem from a lack of sophistication among households, along with agency problems at pension funds and other institutional investors. In the case of households, there is evidence that many households do not understand the financial products they buy (Capon, Fitzsimons, and Prince 1996; Alexander, Jones, and Nigro 1998) or their costs (Choi, Laibson, and Madrian 2010). As a result, such households also probably do not understand that it is hard to identify managers who can consistently generate risk-adjusted excess returns. Gennaioli, Shleifer, and Vishny (2012) suggest that trust is at least as important for manager selection as the desire for outperformance.

In the case of institutions, pension fund and endowment managers are more sophisticated than households, and some of these institutions have been able to earn high returns through their use of high-fee alternative managers (Swensen 2000). However, agency problems appear to have led the vast majority of institutions to overpay for active management. Lakonishok, Shleifer, and Vishny (1992) show that institutional managers underperform the Standard and Poor's 500 by 2.6 percentage points per year, which they attribute to agency problems.[10] Goyal and Wahal (2008) show that investment management firms hired by pension plan sponsors typically underperform when compared to investment management firms that were recently terminated by the same sponsors. Novy-Marx and Rauh (2009) point out that public pension funds have incentives to invest in riskier asset classes because this enables them to report higher return forecasts and thereby discount reported liabilities at a higher rate. And many institutions seek advice from banks and investment advisors, which typically recommend private equity investments that subsequently underperform (Lerner, Schoar, and Wongsunwai 2007).

[10] One of the main agency problems pointed out by Lakonishok, Shleifer, and Vishny (1992), is that the Treasurer's office prefers active management, because these managers need to be monitored and selected, and thus it helps support the perceived need to have a Treasurer's office in the first place.

One could argue that the behavior of unsophisticated and agency-prone investors generates a positive externality: there are surely more resources spent gathering information and monitoring managers than there would be in a world in which investors refused to overpay for active asset management. Absent investors' willingness to overpay, equilibrium securities prices could have less than the socially efficient amount of information, and corporate managers would be subject to insufficient monitoring. For example, venture capital funding of start-up firms, which has arguably brought significant positive externalities, would have been less robust if investors in venture capital funds had required adequate compensation for the risks they were taking.

While one could make this sort of argument, it is not entirely convincing. One important reason is that not all information collection performed by active asset managers is socially valuable. For example, a hedge fund may be willing to pay $20,000 to form a more accurate prediction of a company's earnings to be released in the next week. To the extent that this information allows the hedge fund to profit at the expense of other less-informed market participants, the fund earns an excess return. Hirshleifer (1971) calls information of this type "foreknowledge," but explains that it has no social value. More specifically, the $20,000 expenditure should be regarded as a social loss because getting this information into prices one week earlier is unlikely to lead to a more efficient allocation of real resources. Modern financial markets are rife with examples of such socially wasteful investments. For example, consider the costs of "co-location hosting services," which enable electronic orders to arrive milliseconds faster because of their geographical proximity to trading centers. These investments lend support to Paul Samuelson's view, originally cited in Shiller (2001, p. 243), that modern financial markets display "considerable micro efficiency"—perhaps facilitated by active asset management—while at the same time retaining large "macro inefficiency." We find it noteworthy that over the last 15 years, despite increased resources devoted to asset management, there have been two large and socially costly valuation errors: the Internet bubble at the end of the 1990s and the overvaluation of mortgage-backed securities during the 2000s.

Another reason to question the social benefits of information production by active managers is the evidence that they cater to the preferences of unsophisticated investors. For example, mutual fund managers channel investor flows into the sorts of securities that investors want to own (say, Internet stocks at certain times, high-yield bonds at other times, and so on) rather than allocating capital to its best use (Frazzini and Lamont 2008). Gennaioli, Shleifer, and Vishny (2012) suggest that investment managers cater to unsophisticated investors' preferences to earn their trust.[11] Thus, we think there is good reason to question whether the marginal

[11] Also, Scharfstein and Stein (1990) and Froot, Scharfstein, and Stein (1992) show that reputational concerns can lead active asset managers to herd in their investment decisions. Thus, the inefficiency in active asset management does not depend on there being unsophisticated investors.

dollar of active management makes securities prices more informative. Indeed, Bai, Philippon, and Savov (2012) present evidence suggesting that securities prices have not become more informative since the 1960s.

Finally, when investors overpay for active management, it creates rents in the sector. These rents lure talented individuals away from potentially more productive sectors (Baumol 1990; Murphy, Shleifer, and Vishny 1991).[12] Indeed, during the period of rapid growth in asset management, finance attracted more talent, at least as measured by the number of students entering finance from elite universities. The cost of this reallocation of talent depends, in large measure, on the industries that top students would have otherwise entered and the marginal value of additional talent entering finance. If, for example, students shifted into finance from science and engineering, where rents are low and marginal productivity potentially higher, then the talent reallocation is costly to society. By contrast, the social costs are much lower if the marginal entrant into finance would have otherwise sought a career in other rent-seeking sectors, such as parts of legal services. In a recent study of MIT undergraduates, Shu (2013) shows that finance attracts the best students, as measured by their characteristics at the time of admission.[13]

The Growth of Credit Intermediation

Components of Growth

As illustrated in Figure 1, the credit intermediation industry (as defined by the BEA) grew on a value-added basis from 2.6 percent of GDP in 1980 to 3.4 percent in 2007, having peaked at 4.1 percent of GDP in 2003. The growth of credit intermediation accounted for roughly one-quarter of the growth in the financial sector, which is less than the contribution of the securities industry to financial sector growth and about equal to that of the insurance industry.

As with the securities industry, we examine in more detail the activities that drove the growth of credit intermediation. Again due to data limitations, we look at the output of these activities rather than their value-added. Table 2, using data from the Bureau of Economic Analysis and Economic Census, breaks out credit intermediation into its main components: *traditional banking* (lending and deposit-taking) and *transactional services* related to credit card accounts, deposit accounts, ATM usage, and loan origination. The distinction between these broad categories is admittedly imprecise.

[12] Murphy, Shleifer, and Vishny (1991) argue that talent flows to large markets, where there are weakly diminishing returns, and talent is measurable and contractible. These are all features of asset management.

[13] However, Shu (2013) also shows that the students who go into finance are not the best ones at the time of graduation. The best students at graduation go to graduate school in science and engineering. Thus, it is possible that the lure of finance induces the best MIT students at the time of admission to invest less in coursework and focus more on preparing themselves for a career in finance.

Table 2
Value Added and Output from Credit Intermediation Firms, Selected Years

	$ billions			*% of GDP*		
Industry outputs, by activity	*1997*	*2002*	*2007*	*1997*	*2002*	*2007*
Traditional banking (imputed output)	**179.1**	**253.9**	**328.9**	**2.15%**	**2.39%**	**2.34%**
Lending	76.8	99.2	102.2	0.92%	1.32%	1.34%
Deposit-taking	102.3	79.9	76.9	1.23%	1.07%	1.00%
Transactional services (fees)	**186.1**	**328.0**	**487.7**	**2.25%**	**3.08%**	**3.47%**
Deposits and cash management	24.7	57.5	78.4	0.30%	0.54%	0.56%
Credit card accounts	23.8	23.7	29.6	0.29%	0.22%	0.21%
Other products supporting financial services	17.8	55.0	76.3	0.21%	0.52%	0.54%
Loan origination, nonresidential	14.0	20.2	27.9	0.17%	0.19%	0.20%
Loan origination, consumer residential	11.3	76.8	62.3	0.14%	0.72%	0.44%
ATM and electronic transactions	3.0	6.2	8.6	0.04%	0.06%	0.06%
Other	91.5	88.6	204.6	1.10%	0.83%	1.46%
Total credit outputs	**365.2**	**582**	**816.6**	**4.38%**	**5.47%**	**5.82%**
Bank revenues from activities other than credit intermediation	67.3	109	130.3	0.81%	1.02%	0.93%
Total inputs	**180.8**	**239.9**	**455.2**	**2.17%**	**2.25%**	**3.24%**
	3.8	15.2	14.9	0.05%	0.14%	0.11%
Revenues collected by nonbanks for credit-related activities						
Value added by credit intermediation firms	**247.9**	**436**	**476.9**	**2.97%**	**4.10%**	**3.40%**
Value added for all credit intermediation activities	**211.2**	**374.7**	**415.1**	**2.53%**	**3.52%**	**2.96%**

Source: Bureau of Economic Analysis, Economic Census of the United States, and authors' estimates.
Note: Firms engaged in credit intermediation are mostly banks, but also include credit unions and other savings and lending institutions.

The output from transactional services is simply measured as the fees collected for these services. Measuring the output from traditional banking, which is divided into lending and deposit-taking, is more complex. The output from lending is imputed as the difference between the interest earned on bank loans (that is, loans on bank balance sheets including commercial, consumer, and real-estate loans) and the interest that would have been earned, had the funds been invested in Treasury and Agency securities (those guaranteed by government agencies such as the Federal Housing Administration or government-sponsored enterprises such as Fannie Mae). These calculations use the average interest rate earned on banks' holdings of these securities: that is, *Lending Output* = *Bank Loans* × (*Interest Rate on Loans* – *Interest Rate on Treasury and Agency Securities*). This is meant to capture the ongoing services provided by banks in managing and monitoring loans on their balance sheets, as well as the value of identifying the loans in the first place. However, this basic measure could overstate or understate the value of these services. It overstates the value to the extent that it also includes the credit risk and maturity premium that banks (or any other investors) earn by holding risky long-term loans (Ashcraft and Steindel 2008). The measure could understate the value of these services to the extent that the fees associated with loan origination are included in our transactional services category.

The imputed output from deposit-taking is measured as the quantity of deposits multiplied by the difference between the rate earned on Treasury and Agency securities and the rate paid on those deposits; that is, *Deposit Services Output* = *Deposits* × *(Treasury Interest Rate* – *Average Interest Rate Paid to Depositors).* Depositors presumably accept yields below those of US Treasuries and equivalent government guaranteed securities because they use deposits for transactional purposes.

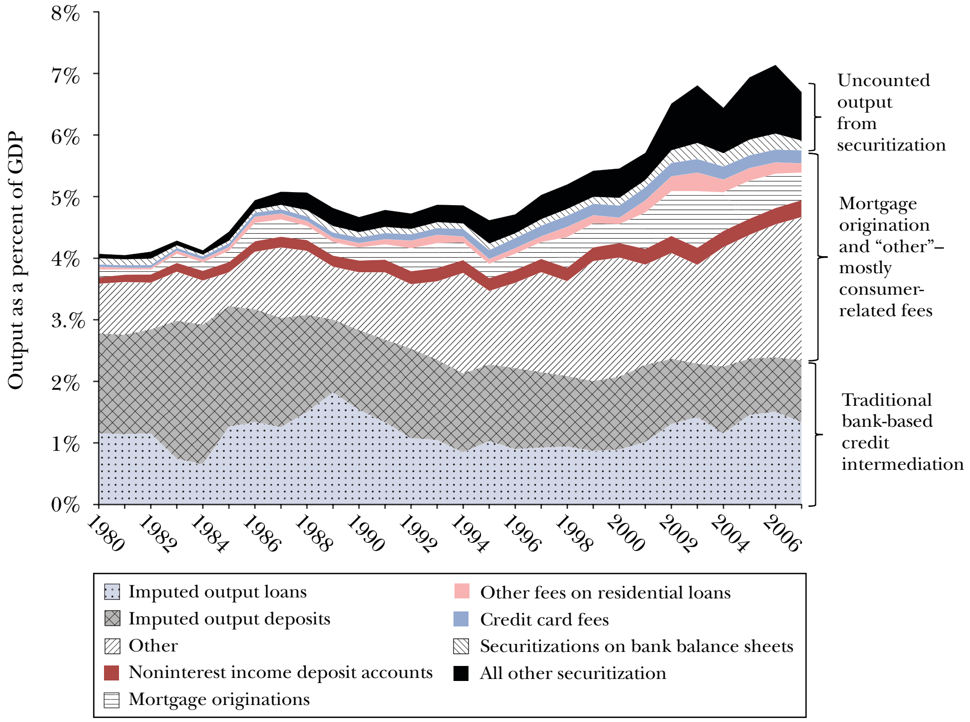
Table 2 shows that the output from traditional banking as a percentage of GDP was roughly the same in 2007 as it was in 1997. However, substantial growth occurred in transactional services, which in turn were largely reflected in fees associated with deposits, residential loan origination, and the catchall category of "other products supporting financial services." In 2002, in particular, residential loan origination fees spiked as part of the largest mortgage-refinancing wave in US history. These fees totaled $76.8 billion—0.7 percent of GDP, or 2.7 percent of the $2.85 trillion of residential mortgages issued in that year.

As in the previous section, we form our own estimates of the sector's outputs going back to 1980. Here, we follow the methodology of the Bureau of Economic Analysis and use data from the Call Reports, which all regulated financial institutions must submit to the Federal Deposit Insurance Corporation at the end of each quarter. As a consistency check, we verify that we can replicate the total output numbers in the years in which the Economic Census is carried out (that is, every five years starting in 1982).[14]

As can be seen from Figure 4, imputed output from lending as a share of GDP has fluctuated around its mean of 1.2 percent of GDP. Much of the variation comes from changes in the ratio of bank loans to GDP, which fell from about 60 percent at the end of the 1980s to under 50 percent at the end of 1990s. During the housing boom in the 2000–2006 period, bank loans rose back to about 60 percent of GDP.

[14] Output from lending and deposit-taking is calculated using data from Federal Reserve's *Call Reports*, and from the *Historical Statistics on Banking* of the Federal Deposit Insurance Corporation. Fees on mortgage loans are imputed from BEA benchmark year estimates using annual mortgage origination totals. Fees on credit card accounts are imputed combining Flow of Funds data on total credit card debt outstanding with Government Accountability Office data on average credit card fees. Data on service charges on deposit accounts are from FDIC's *Historical Statistics on Banking*.

Figure 4
Credit Intermediation Output 1980–2007



Source: Call Reports, Flow of Funds Accounts of the United States, Bureau of Economic Analysis, and authors’ estimates.
Note: For imputed output, we follow the BEA’s methodology.

Figure 4 also shows that output from deposit-taking has generally been falling over time. Some of the decline stems from reductions in spreads between securities and deposits, but the main source of the decline is a reduction in deposits relative to GDP, from its peak of about 70 percent at the beginning of the 1980s to under 50 percent in the early 2000s. This decline mostly reflects a shift of saving into money market funds, bond funds, and the stock market. While traditional banking has declined slightly as a share of GDP, Figure 4 illustrates that essentially all of the growth in the credit intermediation industry has come from transactional services, largely reflected in fees associated with consumer and mortgage credit. A sizable share of the fees can be traced to the refinancing of existing mortgages. Mortgage origination, in turn, is highly dependent on the path of nominal interest rates, which were falling for most of the period we study here and led to extraordinarily high levels of refinancing for a number of years during the period.

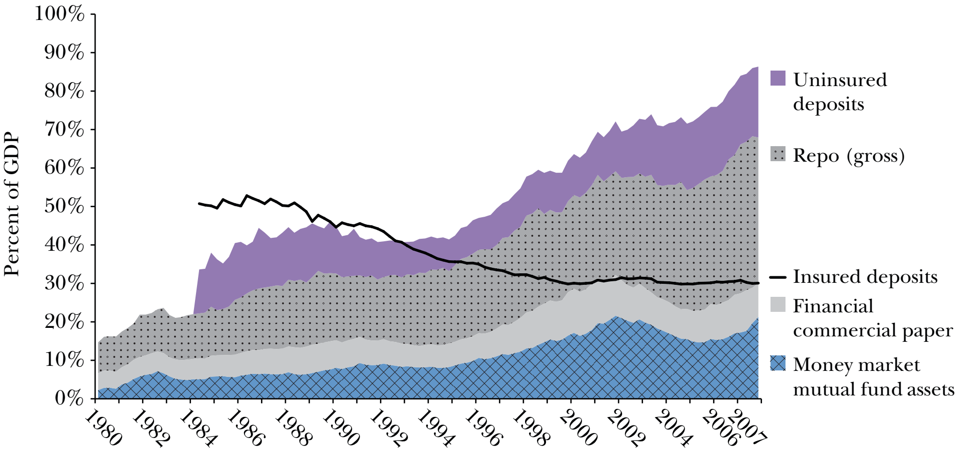
Increase in Household Credit and the Development of the Shadow Banking System

Even with the decline in traditional banking, corporate and household credit rose as a share of GDP from 1980–2007. Overall corporate credit grew from 31 percent of GDP in 1980 to 50 percent in 2007, while corporate loans on bank balance sheets fell slightly, from 14 percent of GDP in 1980 to 11 percent in 2007. Household credit, mainly mortgage debt, grew more dramatically from 48 percent of GDP in 1980 to 99 percent, with the steepest rise occurring during the housing boom of 2000–2006. Despite this growth, banks held roughly the same amount of household credit as a share of GDP—approximately 40 percent—at the beginning and end of the period. All of the incremental growth in household credit as a share of GDP was securitized. That is, instead of banks holding the additional mortgages and consumer loans directly on their balance sheets, these loans were packaged into asset-backed securities. Indeed, as early as 1995, more than half of all outstanding single-family mortgages and a sizeable share of commercial mortgages and consumer credit were securitized.

The growing importance of securitization during the period is not reflected in the Bureau of Economic Analysis measure of output from lending; if a loan is securitized, the interest rate spread is not included in the measure. If instead we incorporate asset-backed securities in the measure by assigning them the same interest rate spread as loans on bank balance sheets, we estimate that imputed output from lending would have been approximately 0.9 percentage points of GDP higher in 2007. The growth in output from securitization is reflected in the top shaded area of Figure 4. Not surprisingly, it increased significantly during the credit boom of 2000–2006.

It is difficult to know whether securitization was driven by an increased demand for credit by households and firms, or by an increase in supply stemming from changes in technology that allowed for easier administration of large pools of securities or lax regulation. Regardless of the cause, securitization surely facilitated the growth of credit. Importantly, securitization also went hand-in-hand with the growth of "shadow banking," in which key functions of traditional banking are provided by a host of nonbank financial entities (though often in conjunction with traditional banks). Pozsar, Adrian, Ashcraft, and Boesky (2010) define shadow banks as "financial intermediaries that conduct maturity, credit, and liquidity transformation without explicit access to central bank liquidity or public sector credit guarantees." Like banks, these entities issue short-term, liquid claims and hold longer-term, riskier, and less-liquid assets. But unlike banks, they cannot issue insured deposits and do not have guaranteed access to the Federal Reserve's lender-of-last-resort credit facilities. Examples of shadow banks include structured investment vehicles that hold loans and asset-backed securities while being funded with short-term asset-backed commercial paper. Money market funds are also shadow banks; they issue short-term claims and hold somewhat longer-term securities. And the government-sponsored entities like Fannie Mae and Freddie Mac hold mortgages and mortgage-backed securities, funded, in part, by issuing

Figure 5
Short-term Funding of the Financial Sector



Source: Author using data from the Flow of Funds Accounts of the United States.
Note: Insured deposits and uninsured deposits are only available starting in 1984.

short-term debt instruments. Figure 5 shows that short-term instruments typically associated with the shadow banking sector—including repurchase agreements (which are effectively secured loans and are often called "repo"), money market funds, and commercial paper—rose significantly as a share of GDP.

Shadow banking institutions do not operate in isolation, but rather are connected to each other in the credit intermediation process. For example, money market funds hold asset-backed commercial paper, which itself holds asset-backed securities comprised of loans that are sometimes guaranteed by other entities. Pozsar et al. (2010) provide a graphical depiction and detailed account of relationships between the various entities of the shadow banking system.

As noted by Adrian and Shin (2010) and others, shadow banking has increased the number of interconnected steps in the credit intermediation process. Combined with short-term leverage, this new approach to banking may have increased financial system fragility. We attempt to measure the increase in the number of credit intermediation steps with a summary statistic, which we call the Credit Intermediation Index. This measure seeks to estimate the average number of steps a dollar takes as it passes from households to the final end-users, with data from the Flow of Funds accounts. For example, when a household makes a direct loan to a business, this direct finance involves one step. If a household deposits funds in a bank, which then makes a loan directly to a business, there are two intermediation steps. More broadly, the ratio of total liabilities (including those of the financial sector which is not an end-user of credit) to liabilities of the household, government, and nonfinancial business sectors (which are end-users of credit) is mathematically equivalent to the expected number of intermediation steps taken by a dollar on the way to its end-user. Thus, the Credit Intermediation Index is defined as:

Credit Intermediation Index = (Total Liabilities of All Sectors)/(Total End-User Liabilities).

Financial sector liabilities, which are a key component of the numerator, include the liabilities of the banking sector: deposits, commercial paper, long-term debt, and repo. They also include money market fund assets, debt of the government-sponsored entities, mortgage pools of the government-sponsored entities, private asset-backed-securities, and the investments of pension funds and mutual funds in credit instruments.[15] The financial sector liabilities that experienced the largest growth are asset-backed securities, borrowing by government-sponsored entities like Fannie Mae and Freddie Mac and government-sponsored entity pools.

Our Credit Intermediation Index captures the increasing number of steps involved in credit creation as shown in Figure 6, with most of the increase occurring during the 1990s.[16] This increase is related to the growth of securitization because most asset-backed securities are held by financial intermediaries rather than by households directly. For example, in 2007 approximately 73 percent of outstanding mortgage-backed securities were held by financial intermediaries, including commercial banks (15 percent), government-sponsored entities (16 percent), and mutual funds (11 percent). These intermediaries, in turn, often fund their purchases of mortgage-backed securities with debt, thereby increasing the number of steps in credit intermediation.

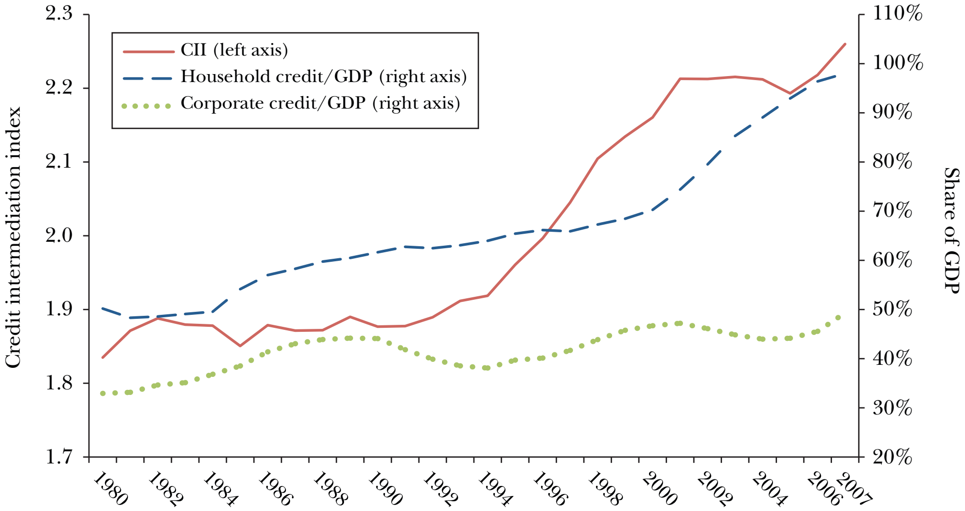
Evaluation of the Growth of Credit Intermediation

A sizable share of the growth of the financial sector can be attributed to the growth in household credit. This growth was likely facilitated by the advent of shadow banking, which expanded the supply of credit to a wider set of households. As noted above, shadow banking also brought fundamental changes in the way credit is delivered.

[15] We are including securitizations in financial sector liabilities. While one could argue that these securities are a form of direct finance like a corporate bond, they rely much more heavily on the ongoing involvement of a variety of financial intermediaries than would a corporate bond. For example, mortgage pools created by the government-sponsored entities like Fannie Mae and Freddie Mac receive a credit guarantee from those entities. Other asset-backed securities require servicers and collateral managers to make payments to bondholders, deal with defaulted loans, ensure that covenants are not violated, and in some cases move collateral in and out of the securitization vehicle.

[16] As constructed, however, this Credit Intermediation Index understates the steps in of the credit intermediation process for a variety of reasons including: our inability to measure intrasector intermediation activity; ignoring approximately $15 trillion of credit derivatives, which transfer risk in the credit intermediation process; understating repos from nonbank entities; and omitting key steps in the credit intermediation chain such as origination by mortgage brokers and mortgage insurance.

Figure 6
Credit Intermediation Index



Source: Flow of Funds and author's calculations.
Notes: The Credit Intermediation Index (CII) is equal to the ratio of gross credit to net credit to end users (government, households, and nonfinancial firms). Household credit and corporate credit are from Table L1 of the Flow of Funds.

It is tempting to argue that society must be better off if, by lowering costs, financial innovation expands the supply of credit and households choose to borrow more. In the standard competitive model, expanding supply is welfare enhancing. But, as in our discussion of asset management, a number of considerations suggest that this logic is incomplete.

First, while credit can play an important role in enabling households to smooth consumption and fund investments, it can also lead to excessive consumption. Laibson (1997) shows that when individuals have self-control problems—which he models with a hyperbolic discount rate—then financial innovation that increases the availability of credit can make these individuals worse off. The steep increase in indebtedness of many low- and moderate-income households above sustainable levels arguably made many of these households worse off. Many houses financed during the 2000–2007 housing boom now sit empty, and many households that increased their credit card borrowing during the credit boom have defaulted (Mian and Sufi 2012).

Second, much of the growth in household credit took the form of an increase in mortgage debt. As is well known, the US tax code already biases households towards investments in housing over other types of investments (Sinai and Gyourko 2004). Making mortgage credit cheaper and more available may have exacerbated a preexisting bias.

Third, an increase in household indebtedness may have adverse consequences for macroeconomic stability. For example, Lamont and Stein (1999) show that household leverage increases house price volatility. Mian and Sufi (2012) show that greater availability of mortgage credit led to large increases in durables consumption, followed by large decreases in consumption when house prices fell during the financial crisis. Households do not take these macroeconomic externalities into account when they choose how much to borrow.

Finally, as noted above, the growth of household credit went hand-in-hand with the growth of shadow banking. While shadow banking offers a number of theoretical benefits—like greater liquidity and the sharing of risk across the financial system—the financial crisis revealed significant financial stability costs of shadow banking. As noted above, these costs stem from the issuance of short-term financial claims without explicit government guarantees by entities that do not have access to the Federal Reserve's lender-of-last-resort facilities, which in turn exposes these entities to runs when investors become concerned about the entities' solvency (Gorton and Metrick 2011). As Stein (2012) argues, market participants do not internalize the full cost that the possibility of these runs may impose on the financial system, resulting in socially excessive issuance of short-term claims. Shadow banking may have also reduced the stability of the financial system by increasing the number of steps in the credit intermediation process, which makes it harder for market participants to understand the risk exposures of their counterparties. Separating credit intermediation into distinct components can provide benefits like intermediary specialization and more liquid financial markets during ordinary times. However, market participants are unlikely to internalize the impact of a longer intermediation chain on financial stability.

Conclusions

Our objective in this paper has been to understand the activities that contributed to the growth of finance between 1980 and 2007, and to provide a preliminary assessment of whether and in what ways society benefited from this growth.

One large part of the growth of finance is asset management, which facilitated increased diversification and household participation in securities markets. As a result, it is likely that required rates of return on risky securities have fallen, valuations have risen, and the cost of capital to corporations has decreased. The biggest beneficiaries were likely young firms. On the other hand, asset management has been very costly. While some amount of active asset management is necessary for informational efficiency and adequate monitoring, there are many reasons to believe that there is too much of it on the margin.

The other major source of growth in the financial sector was in credit intermediation. Financial innovation changed the process of credit delivery in a way that especially facilitated the expansion of household credit, mainly residential mortgage credit. While there may be benefits of expanding access to mortgage credit, there are a number of societal costs from such an expansion, including instability from excessive household leverage. Moreover, the shadow banking system that facilitated this expansion made the financial system more fragile. This runs counter to the traditional "functional" view of finance, which suggests that a primary function of the financial sector is to dampen the effects of risk by reallocating it efficiently to parties that can bear risks the most easily (Merton and Bodie 1995). In evaluating the implications of the growth of the financial sector, such concerns need to be weighed against the many benefits that we have identified.

■ *We thank Jared Dourdeville and Toomas Laarits for excellent research assistance. We are grateful to Lewis Alexander, John Campbell, Darrell Duffie, Sam Hanson, Chang-Tai Hsieh, Anil Kashyap, Ulrike Malmendier, Morgan Ricks, Andrei Shleifer, Jeremy Stein, Adi Sunderam, Paul Tucker, Bob Turley, Luigi Zingales, and especially David Autor and Timothy Taylor for very helpful suggestions. We also thank Jared Dourdeville, Erin Ludlow, James Green, Rodger Smith, Karen Lanzetta, Justyna Podziemka, Covie Edwards-Pitt for their help and advice on financial services data, and the Securities Industry and Financial Markets Association (SIFMA) and Greenwich Associates for providing some of the data.*

References

Adrian, Tobias, and Hyun Song Shin. 2010. "The Changing Nature of Financial Intermediation and the Financial Crisis of 2007–09." Federal Reserve Bank of New York Staff Report 439.

Alexander, Gordon J, Jonathan D. Jones, and Peter J. Nigro. 1998. "Mutual Fund Shareholders: Characteristics, Investor Knowledge and Sources of Information." *Financial Services Review* 7(4): 301–16.

Ashcraft, Adam B., and Charles Steindel. 2008. "Measuring the Impact of Securitization on Imputed Bank Output." February 4. http://www.vanderbilt.edu/AEA/AEAStat/papers/Ashcraft_and_Steindel_1-2008.pdf.

Bai, Jennie, Thomas Philippon, and Alexi Savov. 2012. "Have Financial Markets Become More Informative?" Federal Reserve Bank of New York Staff Report 578, October.

Baumol, William. 1990. "Entrepreneurship: Productive, Unproductive, and Destructive." *Journal of Political Economy* 98(5): 893–921.

Bergstresser, Daniel B., John Chalmers, and Peter Tufano. 2009. "Assessing the Costs and Benefits of Brokers in the Mutual Fund Industry." *Review of Financial Studies* 22(10): 4129–56.

Berk, Jonathan B., and Richard C. Green. 2004. "Mutual Fund Flows and Performance in Rational Markets." *Journal of Political Economy* 112(6): 1269–95.

Brav, Alon, Wei Jiang, Frank Partnoy, and Randall Thomas. 2008. "Hedge Fund Activism, Corporate Governance, and Firm Performance." *Journal of Finance* 63(4): 1729–75.

Capon, Noel, Gavin J. Fitzsimons, and Russ Alan Prince. 1996. "An Individual Level Analysis of the Mutual Fund Investment Decision." *Journal of Financial Services Research* 10(1): 59–82.

Chevalier, Judith A., and Glenn D. Ellison. 1997. "Risk Taking by Mutual Funds as a Response to Incentives." *Journal of Political Economy* 105(6): 1167–1200.

Choi, James J., David Laibson, and Brigitte C. Madrian. 2010. "Why Does the Law of One Price Fail? An Experiment on Index Mutual Funds." *Review of Financial Studies* 23(4): 1405–32.

Fama, Eugene F., and Kenneth R. French. 2002. "The Equity Premium." *Journal of Finance* 57(2): 637–59.

Fama, Eugene F., and Kenneth R. French. 2004. "New Lists: Fundamentals and Survival Rates." *Journal of Financial Economics* 73(2): 229–69.

Fama, Eugene F., and Kenneth R. French. 2010. "Luck versus Skill in the Cross-Section of Mutual Fund Returns." *Journal of Finance* 65(5): 1915–47.

Frazzini, Andrea, and Owen Lamont. 2008. "Dumb Money: Mutual Fund Flows and the Cross-Section of Stock Returns." *Journal of Financial Economics* 88(2): 299–322.

French, Kenneth. 2008. "Presidential Address: The Cost of Active Investing." *Journal of Finance* 63(4): 1537–73.

Froot Kenneth A., Davis S. Scharfstein, and Jeremy C. Stein. 1992. "Herd on the Street: Informational Inefficiences in a Market with Short_term Speculation." *Journal of Finance* 47(4): 1461–84.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. 2012. "Money Doctors." NBER Working Paper 18174.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. Forthcoming. "A Model of Shadow Banking." *Journal of Finance.*

Goldin, Claudia, and Lawrence F. Katz. 2008. "Transitions: Career and Family Life Cycles of the Educational Elite." *American Economic Review* 98(2): 363–69.

Goldsmith, Raymond. 1969. *Financial Structure and Development.* New Haven: Yale University Press.

Gompers, Paul A. 1995. "Optimal Investment, Monitoring, and the Staging of Venture Capital." *Journal of Finance* 50(5): 1461–89.

Gorton, Gary, and Andrew Metrick. 2011. "Securitized Banking and the Run on Repo." *Journal of Financial Economics* 104(3): 425–51.

Goyal, Amit, and Sunil Wahal. 2008. "The Selection and Termination of Investment Management Firms by Plan Sponsors." *Journal of Finance* 63(4): 1804–47.

Greenwood, Robin, and Michael Schor. 2009. "Investor Activism and Takeovers." *Journal of Financial Economics* 92(3): 362–75.

Harris, Robert S., Tim Jenkinson, and Steven N. Kaplan. 2012. "Private Equity Performance: What Do We Know?" NBER Working Paper Series 17874.

Heaton, John, and Deborah Lucas. 1999. "Stock Prices and Fundamentals." Chap. 4 in *Macroeconomics Annual 1999*, edited by Ben Bernanke and Julio Rotemberg. Cambridge, MA: National Bureau of Economic Research, MIT Press.

Hirshleifer, Jack. 1971. "The Private and Social Value of Information and the Reward to Inventive Activity." *American Economic Review* 61(4): 561–74.

Ibbotson, Roger G., Peng Chen, and Kevin X. Zhu. 2011. "The ABCs of Hedge Funds: Alphas, Betas, and Costs." *Financial Analysts Journal* 67(1): 15–25.

Investment Company Institute. 2012. *Investment Company Fact Book*, 52st edition. Washington: Investment Company Institute. http://www.icifactbook.org.

Jurek, Jakub W., and Erik Stafford. 2011. "The Cost of Capital for Alternative Investments." Harvard Business School Working Paper 1910719.

Kaplan, Steven, and Joshua Rauh. 2010. "Wall Street and Main Street: What Contributes to the Rise in the Highest Income." *Review of Financial Studies* 23(3): 1004–1050.

Kaplan, Steven N., and Per Strömberg. 2003. "Financial Contracting Theory Meets the Real World: An Empirical Analysis of Venture Capital Contracts." *Review of Economic Studies* 70(2): 281–315.

Kaplan, Steven N., and Per Strömberg. 2008. "Leveraged Buyouts and Private Equity." *Journal of Economic Perspectives* 23(1): 121–46.

King, Robert G., and Ross Levine. 1993. "Finance and Growth: Schumpeter Might Be Right." *Quarterly Journal of Economics* 108(3): 681–737.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112(2): 443–77.

Lakonishok, Josef, Andrei Shleifer, and Robert Vishny. 1992. "The Structure and Performance of the Money Management Industry." *Brookings Papers on Economic Activity: Microeconomics 1992*, p. 339–91.

Lamont, Owen, and Jeremy C. Stein. 1999. "Leverage and House-Price Dynamics in U.S. Cities." *RAND Journal of Economics* 30(3): 498–514.

Lerner, Josh, Antoinette Schoar, and Wan Wongsunwai. 2007. "Smart Institutions, Foolish Choices: The Limited Partner Performance Puzzle." *Journal of Finance* 62(2): 731–64.

Lewellen, Jonathan. 2011. "Institutional Investors and the Limits of Arbitrage." *Journal of Financial Economics* 102(1): 62–80.

Mehra, Rajnish, and Edward Prescott. 1985. "The Equity Premium: A Puzzle." *Journal of Monetary Economics* 15(2): 145–61.

Merton, Robert C., and Zvi Bodie. 2005. "Design of Financial Systems: Towards a Synthesis of Function and Structure." *Journal of Investment Management* 3(1): 1–23.

Merton, Robert C., and Zvi Bodie. 1995. "A Conceptual Framework for Analyzing the Financial Environment." Chap. 1 in *The Global Financial System: A Functional Perspective*, edited by Dwight B. Crane et al. Cambridge: Harvard Business School Press.

Mian, Atif, and Amir Sufi. 2012. "What Explains High Unemployment? The Aggregate Demand Channel." NBER Working Paper 17830.

Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny. 1991. "The Allocation of Talent: Implications for Growth." *Quarterly Journal of Economics* 106(2): 503–30.

Novy-Marx, Robert, and Joshua Rauh. 2009. "The Liabilities and Risks of State-Sponsored Pension Plans." *Journal of Economic Perspectives* 23(4): 191–210.

Oyer, Paul. 2008. "The Making of an Investment Banker: Stock Market Shocks, Career Choice, and Lifetime Income." *Journal of Finance* 63(6): 2601–28.

Philippon, Thomas. 2012. "Has the US Finance Industry Become Less Efficient? On the Theory and Measurement of Financial Intermediation." NBER Working Paper 18077.

Philippon, Thomas, and Ariell Reshef. 2009. "Wages and Human Capital in the U.S. Financial Industry: 1909–2006." NBER Working Paper 14644.

Pozsar, Zoltan, Tobias Adrian, Adam Ashcraft, and Haley Boesky. 2010. "Shadow Banking." Federal Reserve Bank of New York Staff Report 458.

Rajan, Raghuram. 2005. "Has Financial Development Made the World Riskier?" NBER Working Paper 11728.

Rajan, Raghuram G., and Luigi Zingales. 1998. "Financial Dependence and Growth." *American Economic Review* 88(3): 559–86.

Scharfstein, David, and Jeremy C. Stein. 1990. "Herd Behavior and Investment." *American Economic Review* 80(3): 465–79.

Schumpeter, Joseph A. 1911. *The Theory of Economic Development.* Cambridge, MA: Harvard University Press.

Shu, Pian. 2013. "The Impact of the Financial Crisis on Selection into Jobs in Finance: Evidence from MIT Graduates." Working Paper.

Shiller, Robert J. 2001. *Irrational Exuberance.* Princeton University Press.

Sinai, Todd, and Joseph Gyourko. 2004. "The (Un)changing Geographical Distribution of Housing Tax Benefits: 1980–2000." *Tax Policy and the Economy*, vol. 18, pp. 175–208.

Stein, Jeremy C. 2012. "Monetary Policy as Financial Stability Regulation." *Quarterly Journal of Economics* 127(1): 57–95.

Swensen, David F. 2000. *Pioneering Portfolio Management: An Unconventional Approach to Institutional Investment.* New York: The Free Press.

Turner, Adair. 2010. "What Do Banks Do? Why do Credit Booms and Busts Occur and What Can Public Policy Do about It?" Chap. 1 in *The Future of Finance*, edited by Adair Turner et al. London School of Economics.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 29–50

Finance: Function Matters, Not Size

John H. Cochrane

The US economy spends $170 billion a year on advertising, just to trick people into buying stuff they don't need. What a waste!

There are 2.2 people doing medical billing for every doctor that actually sees patients, costing $360 billion—2.4 percent of GDP. Talk about an industry that is too big!

Wholesale and retail trade and transportation cost 14.6 percent of GDP, while all manufacturing is only 11.5 percent of GDP. We spend more to move goods around than to make them!

My wife asked me to look at light fixtures. Do you know how many thousands of different kinds of light fixtures there are? The excess complexity is insane. Ten ought to be plenty.

It's ridiculous how much people overpay for brand names when the generic is so much cheaper. People are pretty naive.

Business school finance professors are horribly overpaid. Ask an anthropologist! We get paid almost a half a million bucks, and work a grand total of 10 weeks a year, all to teach students that they can't make money trading in the stock market.

It's fun to pass judgment on waste, size, usefulness, complexity, naiveté, and excessive compensation, isn't it? But as economists, we have an analytical structure for thinking about these questions. We start with supply, demand, and competition,

■ *John H. Cochrane is the AQR Capital Management Distinguished Service Professor of Finance, University of Chicago Booth School of Business, Chicago, Illinois. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts; Senior Fellow, Hoover Institution, Stanford, California; and Adjunct Scholar, Cato Institute, Washington, DC. His email address is john.cochrane@chicagobooth.edu, and his website is http://faculty.chicagobooth.edu/john.cochrane.*

http://dx.doi.org/10.1257/jep.27.2.29. doi=10.1257/jep.27.2.29

and with the suggestion of the first welfare theorem that these forces usually lead to socially beneficial arrangements. When outcomes seem puzzling using this analysis, we embark on a three-pronged investigation. First, we work harder to find how supply and demand might really operate, in the humble knowledge that initially puzzling institutions and outcomes have often taken us years to comprehend. Second, maybe there is a "market failure"—an externality, public good, natural monopoly, asymmetric information situation, or missing market—that explains our puzzle. Third, we often discover a "government failure," that the puzzling aspect of our world is a consequence of laws or regulation, either unintended or the result of capture.

Only then can we begin to diagnose a divergence between reality and socially desirable outcomes, and only then can we start to think of how to improve reality. "I don't understand it" doesn't mean "it's bad," or "regulation will improve it." And since that attitude pervades policy analysis in general and financial regulation in particular, economists do the world a disservice if we echo it.

I belabor this point, because I do not offer a competing black box. I don't claim to estimate the socially optimal "size of finance" at, say, 8.267 percent of GDP. It's just the wrong question. Hayek and the failure of planning should teach us a little modesty: Pronouncing on socially optimal industry size is a waste of time. Is the finance industry functioning well? Are there identifiable market or government distortions? Will proposed regulations help or make matters worse? These are useful questions.

With a rather catastrophic failure behind us and other crises bubbling on the back burner, it also seems a bit strange to be arguing whether 5 or 8 percent of GDP is the right "size" of finance, and whether it needs to be nudged to become larger or smaller. Many of us might happily accept an additional 3 percentage points of GDP in the financial sector in return for a financial system that is not prone to runs and crises. Our political system *has* accepted a big increase in resources devoted to financial regulation and compliance, and a potentially larger reduction in the efficiency, innovation, and competitiveness of financial institutions and markets, in the quest—misguided or not—for stability. The run-prone nature of the US financial system, together with its massive regulation, subsidies, government guarantees, and regulatory capture, looks to be a more fertile fishing ground for trying to understand market and government failures than does mere size.

Still, the size of finance represents a contentious issue, and my plea that we ask different questions isn't going to silence the debate, so let us think about it. Let us use size as an organizing principle for studying function and dysfunction.

Greenwood and Scharfstein nicely review the key facts and ideas in their paper in this issue. Their most basic story is: quantity increased a lot, but prices didn't fall. This description suggests a simple economic interpretation: The demand for financial services shifted out. People with scarce skills supplying such services made a lot of money. A system with proportional fees, which is a common structure in professional services, interacted with stock-price and home-price increases (a different surge in demand) to produce increased financial sector revenue. *Why* demand shifted out, and why house and stock prices rose (temporarily, it turns out) are good questions—but they don't have much to do with the structure of the finance industry. This story also suggests that, like the weather, if you don't like the size of finance, just wait a while. Finance has contracted rather dramatically since 2007.

Many puzzles remain, however, and the current academic literature paints an interesting and quite novel picture of how the finance industry functions—and maybe does not function.

The Controversy over Active Management Fees

Management fees are a big part of the "size of finance." Fees aren't GDP, of course, but they are much more easily measured. The large overall rise in fee revenue reflects several offsetting trends. Individuals moved investments from direct holdings to mutual funds, and then to index funds or other passive funds. This trend continues. New investors in defined contribution plans invest almost exclusively in mutual funds or exchange-traded funds.

Mutual fund fee *rates* came down sharply, in part reflecting the slow shift to very low-fee index and semi-passive funds, and in part reflecting competitive pressure. French (2008) reports that the average actively managed equity mutual fund fee fell from 2.19 percent in 1980 to 1 percent in 2007. Greenwood and Scharfstein (2012) report that average bond fund fees fell from 2.04 to 0.75 percent. Some index funds charge as little as 0.07 percent. Fee-based advisers and wealth managers are lowering fees, and bundling larger arrays of services, including tax and estate planning.

Funds are far more efficient vehicles for individual investors than holding individual stocks. The measured GDP of the fund industry is at least in part a benefit rather than a cost, as it displaces inefficient and unmeasured home production of financial management services. Hiring a (legal) house cleaner also raises measured GDP.

Thus, mutual fund fee revenues reflect declining rates multiplied by a much larger share of assets under management. This market does reflect sensible forces, if one is willing to grant a rather long time span for those forces to affect industry structure. But after all, the moves to low-cost airlines and big-box retailers took a while too.

However, at the same time that individuals were moving to passive funds and those funds were expanding, high-wealth individuals and institutions (pensions, endowments, sovereign-wealth funds, and so forth) moved their investments to hedge funds, private equity, venture capital, and other even higher-fee and more-active investment vehicles. Hedge fund fee rates are reportedly stable over time, and surprisingly large: managers charge 1.5–2.5 percent of assets each year, and also 15–25 percent of profits. This part of the market offers the more puzzling behavior.

The Traditional View

High-fee active management and underlying active trading have been deplored by academic finance for a generation. French (2008) offers a comprehensive summary. French estimates that equity investors in aggregate, between 1980 and 2006, paid 0.67 percent per year in active management fees, whose present value he estimated to equal 10 percent of their investments. French eloquently conveys the view that these investors wasted their money.

The standard analysis divides investment returns into "alpha" and "beta." We run a regression of a fund's returns on the returns of a low-cost index, both returns in excess of the risk-free rate. Beta is the slope coefficient. Beta times the index return is the component of the fund's return that is earned for passively shouldering systematic risk, and can be synthesized by the investor without paying fees. Alpha is the intercept in this regression, and gives the mean of that part of the fund's return that cannot be easily replicated. Alpha is conventionally interpreted as the extra return that the fund earns, on average, from the manager's talent or superior information, and therefore potentially worth paying a fee to obtain. Both alpha and beta are, conceptually, one's best estimate of this decomposition of returns going forward, of course. Estimated alphas from past history contain a great deal of luck.

The average alpha of all equity mutual funds, before fees, is very nearly zero. This result follows almost by accounting, since the portfolio of equity mutual funds, taken as a whole, is almost exactly the value-weighted market portfolio.

The evidence on hedge fund, private equity, venture capital, and other returns is complicated by survivor bias (funds that perform badly tend to drop out of the data) and by difficulties of calculating benchmarks that appropriately reflect the risks, time horizons, and illiquidity of these investments. But the academic argument over whether such funds as a class provide substantial alpha ends up arguing over a few percentage points one way or the other—hardly the promised gold mines.

Mediocre average results for actively managed investments might not be surprising. Entry into the business is relatively free. The average artist isn't that good, either.

But one might expect that, as in every other field of human endeavor, the good managers would be reliably good. Michael Jordan's past performance was a good forecast of what would happen in the next game. Yet the nearly universal conclusion of the academic literature is that there are no reliably "good" managers.

To evaluate this question, we must separate skill from luck. "Why did Warren Buffet earn so much money?" is not a productive question. The classic technique is to examine rules by which one might have chosen funds in the past, and then study the subsequent returns of all such funds. Study after study finds no reliable rule that one can use to identify funds that will perform well in the future, after controlling for betas. (Carhart 1997 is an excellent example.) Fama and French (2010) pursue a clever measurement that does not require one to hypothesize such a rule. They show that the distribution of estimated alpha across mutual funds is only very slightly wider than what one would expect if sample alphas were just due to luck. Fama and French estimate (p. 1935) that the distribution of true alpha has a standard deviation of only 1.25 percent on an annual basis, meaning that only about one-sixth of funds have true alphas (gross, before fees) of 1.25 percent or greater—while another one-sixth have "true alphas" of negative 1.25 percent or worse. (True negative alpha is a bit of a puzzling concept. You should not be able to reliably underperform the market either, as all I have to do is short what you buy.) And all of this before fees.

A Supply-and-Demand View of Active Management and Its Fees

It seems the average investor should save 60 basis points a year and just buy a passive index such as Vanguard's Total Stock Market Portfolio. It seems that the stock pickers should do something more productive, like drive cabs. Active management and its fees seem like a total private, and social, waste.

Yet his hallowed view—and its antithesis—do not completely make sense. After all, active management and fees have survived 40 years of efficient-market disdain. Economists who would dismiss "people are stupid" as an "explanation" for a pricing anomaly that lasts 40 years surely cannot use the same "explanation" for the persistence of active management. Economists who think the evidence favors lots of "inefficiencies" in the market are even less well placed to deplore active management. They should conclude that we need more, or at least better, active management to correct the market's inefficiencies. Their puzzle is the inability of existing managers to pick low-hanging fruit.

Progress is being made at last. Berk and Green (2004) have created a supply-and-demand economic model that explains many of the basic facts of mutual fund performance, flows, and fees. (Berk 2005 offers a simple exposition.)

Suppose that some fund managers do have alpha. Alpha, however, has diminishing returns to scale. Traders report that many strategies apply only to smaller stocks (see evidence in Fama and French 2006) or that prices move against them if they try to execute trades that are too large. As an example, suppose that a manager can generate 10 percent risk-free alpha with $10 million in assets under management. Suppose also that the manager's fees are 1 percent of assets under management, and suppose that the market does not go up or down. Then, in his first year, the manager makes $1 million abnormal return. The manager pockets $100,000 and investors in the fund receive $900,000.

Seeing these good results, investors rush in. But the manager's idea cannot scale past $10 million of assets, so the manager invests extra money in an index. With $20 million under management, the manager generates $1 million alpha on the first $10 million and nothing on the rest. The manager again receives 1 percent of assets under management, which is now $200,000. But investors still get $800,000 alpha. More investors pour in.

The process stops when the manager has $100 million under management. The manager still generates $1 million alpha, but now he collects $1 million in fees. His investors get exactly zero alpha, the competitive rate of return. Everyone is acting rationally.

Berk and Green's (2004) model is much more sophisticated than this simple example. They include uncertainty in returns and a signal extraction problem for investors, which give rise to interesting dynamics. A large literature has followed.

This model explains many puzzling facts: In equilibrium, returns to investors are the same in active and passively managed funds. Funds earn only enough alpha to cover their fees. Good past fund returns do not forecast good future returns. Investors chase managers with good past returns anyway, seemingly irrational behavior and thus one of the most famous puzzles in the mutual fund literature (for example, Chevalier and Ellison 1997). Returns to investors do not measure alpha. Fees do. Managers with good track records get paid a lot.

This model is the focus of the current debate. Fama and French (2010) complain that the average alpha before fees is nearly zero—and negative, not zero, after fees. Berk and Van Binsbergen (2012) answer that Fama and French's benchmarks are not tradable, and skill should be measured as alpha times assets under management, as 0.1 percent alpha on a billion dollars is a lot. Using these measures, they find investors just about breaking even, and a good deal of positive skill. (Fama and French's Table AI agrees.) The model needs to be brought to the data quantitatively: Does the magnitude of fund flows following performance follow the model's predictions? Does it describe fund exit, the persistence of negative alpha, and the shift to passive management? Like all models, one can explore deeper foundations. What is this alpha, anyway? Why are fees a flat percentage of assets under management? If the manager could simply charge a $1 million fee to start with, the fund would not need to expand.

And all that is how it should be. After 40 years, the research agenda is finally about how to fit the facts into a supply and demand framework. Arguing about benchmarks, calibration, and optimal contracts is a lot more productive than deploring the financial industry as folly, or declaring that if it survives, markets must be working. The answer will surely not end up all on one side or another: Surely some investors have overpaid for pointless trading. Surely there is some durable value in an industry that has lasted so long. Surely there are some understandable distortions. On this path, we may finally understand how this market works, and maybe, humbly, suggest some improvements. This is a great example of how the economic framework operates—and a sobering reminder of how long it often takes to see that a straightforward economic analysis is possible.

Is It Silly to Pay a Proportional Fee?

Much of the argument that "finance is too big" rests on the view that fees based on a proportion of assets under management are a suboptimal contract. Assets under management went up, fees went up, and managers laughed all the way to the bank. This is a big part of Greenwood and Scharfstein's story in this issue. On closer examination, this argument seems awfully strained.

First, we have seen in fact a substantial decline in many fees and migration to lower fee vehicles, in mutual funds, exchange traded funds, and many wealth management services. Competition does seem to be working, though more slowly than we may like.

Second, fee revenue is not a good measure of the "size" of finance. Fees are a transfer, like gambling losses, not a measure of resources consumed or output produced. Policy may and obviously does care a lot about transfers, but that is a conceptually different question than worrying about wasted resources. Moreover, fees vary based on outcomes. If the fund gains or loses money, fee income rises and falls as well. Hedge fund fees, usually 2 percent of assets and 20 percent of profits, vary enormously. The same fees that were puzzlingly high in 2006 were a lot lower in 2008. Fees have much of the character of a risk-sharing arrangement among co-investors, rather than an expense for professional services.

Third, if the fund doubles in value because everything else in the economy doubles—capital stock, earnings, and so on—then surely by constant returns to scale, the value of investment management (whatever that is) also doubles.

And finally, I'd like to see a specific claim as to what the alternative, realistic, and privately or socially optimal contract is. Funds cannot bill by the hour, passing on "cost" as lawyers do (or rather, used to do), for obvious monitoring and principal–agent reasons. Should we agree to pay a fraction of initial investment, regardless of subsequent performance? It's obvious why we don't do that. Accounting for different vintages of investment would be a nightmare. It would also violate the regulatory principle that all investors must be treated equally.

Proportional fees seem almost inescapable in funds that allow investors to withdraw money and invest freely. Suppose funds charge 1 percent for new money, but do not lower dollar fees after losses. Then after a fund has lost half its value, its investors face 2 percent fees going forward. They will quickly withdraw their money and give it to a new fund. Funds that lost money would quickly spiral out of existence, or investors would undermine the fee by withdrawing and reinvesting the next day as new money. Venture capital, private equity, and some hedge funds do not allow free withdrawal so for them, this argument does not apply as strongly—and they have more complex fee structures.

Percentage fees pervade professional services. Real estate agents charge percentage fees, and do better when house prices rise. Architects charge percentage fees. Contingency-fee lawyers take a percentage of winnings. Salesmen get percentage commissions. Even corrupt officials often take percentage bribes.

Perhaps the argument boils down to the claim that there is no alpha, so nobody should pay any fees at all for active management. That's a different question. If there is alpha or some other function of active management, its optimal contract is a difficult (and much-studied, though I do not review it here) principal–agent problem. Skill is hard to measure, and a fund's actions are hard to monitor. It seems a big jump to conclude that percentage fees came into existence and have persisted for decades, across a wide range of industries, while inflicting important private and social costs, just because people are naive or irrational in some unspecified way.

Are Fee-Payers Naive?

Delegating active management and paying large fees is common and increasing among large, completely unconstrained, and very sophisticated investors. For example, the Harvard endowment was in 2012 about two-thirds externally managed by fee investors and was 30 percent invested in "private equity" and "absolute return," largely meaning hedge funds.[1] The University of Chicago endowment is similarly invested[2] in private equity and "absolute return." Apparently, whatever qualms some of its curmudgeonly faculty express about alphas, fees, and active management are not shared by the endowment. Its most recent annual report states: "The majority of TRIP's [Total Return Investment Portfolio] assets are managed by external managers specializing in a specific asset class, geography, or strategy. These asset managers outperformed their respective benchmarks in every asset class, adding over 500 basis points of performance versus the strategic benchmark." Five hundred basis points! Put that in your pipe and smoke it, efficient marketers. At least we know one active manager's perception of what they get for their fees.

These endowments' approach to portfolio management is pretty much standard at endowments, nonprofits, sovereign wealth funds, family offices, pension funds, and so forth—anywhere there is a big pot of money to invest. These investors pay a lot of attention to allocation among name-based buckets, as represented in the pie charts, "domestic equity," "international equity," "fixed income," "absolute return," "private equity," and the like. Then, they allocate funds in the buckets to groups of fee-based active managers.

This approach bears no resemblance to standard portfolio theory, in which an investor pays attention only to means and covariances, not buckets. And don't even ask how often hedge fund manager A is shorting what B is buying; what happens to fees when you give a portfolio of managers 2+20 compensation and half of them win and half lose; or why one would pay the manager of a growth-oriented fund to buy the same stock that the manager of the value-oriented fund just sold.

Why have these decision procedures become standard practice? Vague reference to "agency problems" and "naiveté" seem unpersuasive. Harvard's endowment was overseen by a high-powered board, including its president Larry Summers, possibly the least naive investor on the planet. The picture that Summers and his board, or the high-powered talent on Chicago's Investment Committee are simply too naive to demand passive investing, or that they really want the endowments to be invested in the Vanguard total market index, but some "agency problem" with the managers they hire and fire with alacrity prevents that outcome from happening, simply does not wash. (Yes, delegated portfolio management is a classic principal-agent problem. But no, it's hard to conceive that it produces this result.) Perhaps instead, we should admit that standard portfolio theory is not much help in situations of any real-world complexity, try to understand what these rough and ready procedures achieve, and offer more helpful advice.

As for "excessive" compensation, in the first layer of fees (fees to the manager who pays fees to the other managers) Harvard endowment's CIO Jane Mendillo

[1] See the Harvard Management Company website: http://www.hmc.harvard.edu/investment-management.
[2] See the University of Chicago's *Annual Report*, "The Endowment": http://annualreport.uchicago.edu/page/endowment.

was paid $4.7 million, most of which was straight salary.[3] The University of Chicago's Mark Schmid gets only $1.8 million, though our measly $5.6 billion assets under management relative to Harvard's $27.6 billion may have something to do with it. If major nonprofit university endowments are paying this much, is it really a puzzle that pension funds do the same thing?

Finding Alpha? Implications for Active Trading

To justify fees for active management, one must explain why active trading is worthwhile. The *average investor theorem* is an important benchmark: The average investor must hold the value-weighted market portfolio. Alpha, relative to the market portfolio, is by definition a zero-sum game. For every investor who overweights a security or invests in a fund that earns positive alpha, some other investor must underweight the same security and earn the same negative alpha. Collectively, we cannot even rebalance. And each of us can protect ourselves from being the negative-alpha mark with a simple strategy: hold the market portfolio, buy or sell only the portfolio in its entirety, and refuse to trade away from its weights, no matter what price is offered. If every uninformed trader followed this strategy, informed traders could never profit at our expense.

Alphas and Multiple Factors

Alpha seems a dicey proposition. But the last 20 years of finance research is as clear as empirical research in economics can be: There is alpha relative to the market portfolio—there are strategies that deliver average returns larger than the covariation of their returns with the market portfolio justifies—lots of it, and all over the place. In Cochrane (2011), I provide a summary of this huge literature; I won't provide a separate citation for each fact here.

Examples of such strategies include value (stocks with low market value relative to accounting book value), momentum (stocks that have risen in the previous year), stocks of companies that repurchase shares, stocks of companies with accounting measures of high expected earnings, and stocks with low betas. The "carry trade" in maturities, currencies and, credit—buy high-yield securities, sell low-yield securities—and writing options, especially the "disaster insurance" of out-of-the-money put options, all generate alpha. Expected returns on the market and most of the anomaly strategies vary predictably over time, implying profitable dynamic trading strategies.

Many of these anomalies lead to new "factors," new dimensions of "systematic" risk and rewards. For example, if one buys a large portfolio of "value" (low-price) stocks, engineered to have zero correlation with the market, thinking that one will reap the value-stock alpha and diversify away the risks, one soon discovers the

[3] See "Chart: Top Paid CIOs of Tax-Exempt Institutions," http://www.pionline.com/article/20111107/CHART04/111109905.

tendency of all value stocks to rise and fall together. The portfolio remains risky no matter how many stocks one adds. In this way, pursuing the "value" alpha requires one to take on this additional dimension of undiversifiable risk.

As formalized in Fama and French's (1996) three-factor model and its larger successors, the world appears to have many such "factors," acting as the market return factor did in our early understanding, each offering orthogonal dimensions of risk and a return premium to those investors who are willing to take the risks. Those "factor premiums" capture most of observed "alpha" relative to the market portfolio.

Large risk premiums opened up in the recent financial crisis, as prices of very nearly identical securities diverged. For example, corporate bonds traded at lower prices than their synthetic replication by a Treasury bond and a credit default swap. The "covered interest parity" condition failed: You could earn money by borrowing dollars, buying euros, investing in European money markets, and converting back to dollars in the futures markets. If you could borrow dollars! These events and other price movements in the crisis suggest to the researchers studying them "fire sales," "financial constraints," "financial frictions," "price pressure," and "limits to arbitrage"—all of which are ways of saying that the active managers of the time were insufficient to equalize prices of nearly identical securities, and active traders could have made alphas. Similar pricing divergences and insufficient arbitrage appeared in the trading frenzies of the Internet boom (for example, Lamont and Thaler 2003; Cochrane 2003).

There are multiple dimensions of risk, and bearing these risks generates expected-return rewards, rewards that change over time. These facts are not really under debate. Their interpretation is. These alphas might represent imperfect risk sharing and (often temporary) market segmentation, or "sentiment," irrational attachment or aversion to broad categories of securities. They might also reflect a multidimensional and time-varying nature of risk premiums in a fully-integrated and informationally efficient market. They certainly look less and less like "information" about individual securities that is somehow improperly reflected in prices.

These facts and interpretations lend a quite new color to our central questions: Is the financial sector too large or too small? How should investors behave in a world with multiple dimensions of systemic risk? What is the economic function of active management, and the economic value of management fees?

Multidimensional Risk-Sharing

The conventional disdain for active financial management is based on a conventional perspective: The market portfolio is the one and only source of "systematic" risk which generates a premium. It is accessible through low-cost passive investments. The investor understands this opportunity and knows how much market risk he or she wishes to take. Alpha represents the trader's knowledge of information not reflected in market prices.

But the dozens of semi-passive strategies, each of which produce alpha (relative to the market), each of which exposes the investor to new dimensions of undiversifiable risk, and many of which are poorly understood, changes the picture completely.

Each investor needs to decide which of the many sources of risk he or she is best able to bear, or needs to avoid despite their attractive premiums.

Investors need to consider the even larger set of asset market risks that do not bear premiums. Before chasing alphas, investors should hedge the risks of their jobs, businesses, outside income streams, real estate, or peculiar liability streams by setting up portfolios of assets whose returns are negatively correlated with those risks. You should want a portfolio that rises when there is bad news about your future income. Curiously, academic finance has done little to characterize these nonpriced risks and prescribe hedging strategies.

One can see this process beginning. Many pension funds are moving towards bond-like investments to match their liabilities. University endowments are beginning to recognize how their liability streams affect investments. They thought of themselves as "long term" investors able to reap the premiums of illiquid investments, and able to wait patiently through market downturns, until many in the crisis realized they were supporting a bond-like liability stream in salaries of tenured professors, and were leveraged by bond-financed construction. They found themselves trying to sell illiquid assets at the bottom like everyone else. Now, they are thinking about matching endowment funding to projects that can bear risk and adapting portfolios to their cash flows, including the implicit beta that alumni donations rise when the stock market goes up. Endowments are recognizing that their objectives include an important tournament relative to other universities (Goetzmann and Oster 2012). The wealth-management arms of big banks help to set up hedge portfolios for executives who have large unsaleable stock or option positions, to help them come as close to shorting their own business as possible. Websites available to individual investors are starting to emphasize intelligent and individual-specific choice of "style" rather than promise generic "alpha."

But none of this is easy. Merton (1971) described state-variable hedging demands 40 years ago. Yet, with thousands of following papers, academic portfolio theory still really does not offer clear-cut real-world advice (Cochrane forthcoming).

The nature and amount of multidimensional systematic risk one should take is also much more nebulous and difficult to assess than the traditional question of how much market risk one should take. Should you write put options, to earn the premium? Or maybe you should buy put options as disaster insurance? Are you positioned to buy value stocks? To take on the credit risks of default? To take the risk that high-interest rate foreign currencies depreciate against the dollar? Do the alpha premiums these strategies offer compensate for the risks you will suffer when they lose money? The whole alpha/beta definition is falling apart.

Even then, taking advantage of time-varying multidimensional risks requires technical knowledge. Do you know how to write a credit default swap contract, how to make stock momentum strategy work without drowning in transactions costs, how to take advantage of temporarily high put option premiums in the euro-zone, or even how reliably to buy a "value" portfolio? Because such questions are not easy, portfolio problems like this might certainly benefit from professional and specialized management, and such management ought to be able to charge a fee.

Perhaps some of the puzzling features of investment practice might be understood as a rough and ready way of adapting to this more realistic portrait of risks and returns. If so, some active management and dynamic trading represents a form of socially beneficial insurance provision.

Hedge funds might make more sense in this investment world. They can move to and from asset classes as risk premiums change, and by using leverage and derivatives they can alter overall exposures quickly without incurring the transactions cost of buying and selling large portfolios.

Many of these alpha-generating strategies and new "factors" suggest needed institutional developments. As a concrete and recent example, consider the "betting against beta" anomaly reexamined by Frazzini and Pedersen (2011a, b). They document that low-beta stocks get higher average returns than they should, and high beta stocks get lower returns than they should. Their interpretation is that many investors want more risk than the market portfolio provides, yet leverage is costly to obtain. These investors buy high-beta stocks instead of leveraging, driving up the prices of high-beta stocks, and vice versa for low-beta stocks. In this setting, arbitrageurs cannot help. The problem is the price of risk, needing wider risk-sharing, not an arbitrage (riskless profit) opportunity. To bring prices back to what they should be, we need low-cost vehicles to bring leveraged low-beta investments to the part of the investing public that wants them—which, perhaps not coincidentally, Frazzini and Pedersen's company provides.

We have seen this kind of institutional development before. Small stocks were one of the first prominent anomalies, generating (it appeared) higher average returns than their betas justified. But it was hard for individual investors to hold a diversified portfolio of small stocks. Arbitrageurs could only do so much, because small stocks move together, so a concentrated portfolio bears undiversifiable risk. Small stock mutual funds were started, which allowed a mass of investors to participate. Fees and expenses of those funds contributed to revenue and measured GDP, in a way that the activities of individual investors holding small stocks did not. But they allowed the risk of small stocks to be widely shared and the small stock premium to decline.

So far I have made no mention at all of informational inefficiency, exploiting mispricings, superior information, or winning the zero-sum alpha game. I have not violated the average investor theorem. Given the new facts of empirical finance, a large role for active management exists without any of that at all. Of course, I do not claim that current portfolio practice, and especially hiring many different high-fee hedge funds, is an optimal strategy. But it isn't necessarily as "naive" or "agency conflicted" as it otherwise seems.

Marketing

In the quest to explain the persistence of active management and its fees, one other analogy seems worth pursuing: marketing. Marketing and advertising have long been a puzzle to economists, along with readers of *Consumer Reports* and coupon-clippers everywhere. Why buy the brand name when the generic is nearly identical, and costs a lot less?

The money-management industry is essentially a marketing industry. Its practitioners take generic ingredients, package, label, advertise, and market them. Yes, it's puzzling that people don't buy the generic at Vanguard. It's puzzling that they don't buy the pieces and assemble their own, with E*TRADE. It's puzzling that they pay so much for the slight differences in ingredients that the active managers deliver. And it is equally puzzling that they pay for Coke, Clorox, Bayer, or bottled water; that they shop at Macy's not Target, Whole Foods not Costco, and a hundred other brand names.

This is not the place to digress into the "rationality" of marketing and advertising. Simply dismissing centuries worth of branding and advertising as naiveté and folly seems, well, its own form of naiveté. Perhaps by thinking of active fund management as an instance of this larger pattern, we may make some progress to understanding how it actually works.

Information Trading and Price Discovery

Much trading and active management, however, is clearly aimed at bringing information to the market, not at better sharing of time-varying and multidimensional risk. The first welfare theorem does not clearly apply to information production, so we have little a priori reassurance that the quest for trading profits produces the "right" amount—or, perhaps more importantly, the right *kind*—of information.

It is possible that *not enough* social resources are devoted to trading, because information is a public good. As French (2008) wrote, despite deploring the private costs of alpha-chasing: "I offer no evidence on whether society is buying too little or too much of this good. Price discovery, however, is an externality—each active investor pays the full cost of his efforts but captures only a tiny slice of the benefit—so there is no reason to think active investors purchase the optimal amount of price discovery."

The common complaints "the financial crisis proves markets aren't efficient," or that tech and mortgages represented "bubbles," are at heart complaints that there was not *enough* active information-based trading. All a more "efficient" market could have done is to crash sooner, by better expressing the pessimist's views. Remember, "efficiency" means that prices incorporate all available information, not that markets are clairvoyant. The definition of "efficiency" is widely misunderstood. I once told a newspaper reporter that I thought markets are pretty "efficient," and he quoted me as saying markets are "self-regulating!"

If information is *not* incorporated into market prices and to such an extent that simple strategies with big alphas can be published in the *Journal of Finance*, there are not enough arbitrageurs. If asset prices fall in "fire sales," only to rebound later, there are not enough buyers following the fire trucks. If credit constraints are impeding the flow of capital, there is a social benefit to loosening those constraints.

The literature on short-selling is revealing on this point. Short sellers uncover far more financial fraud than the Securities and Exchange Commission. Conversely, some of the biggest alphas and "inefficiencies" occur when there is a technical or regulatory impediment to short seller's activities. Lamont (2012) finds 2.4 percent monthly alpha to a portfolio of short-selling-constrained stocks, a large informational inefficiency. This is a concrete example of inadequate (because constrained) information-based trading.

Information trading produces more informationally-efficient prices, which are socially useful. With better market signals, companies raise capital more easily for valuable projects, and are signaled not to invest in poor projects or at poor times. True, the simple *q* theory, which predicts that corporate investment should be a perfect function of stock price relative to book value, is formally rejected, but its glass is also half full: There are strong correlations between stock prices and investment, over time (through the tech boom and bust of the 1990s and through the financial crisis (see Cochrane 1991; 2011, Figure 10)) and across industries (Google versus, say, GM). When issuing stock generates a lot of money, companies do it, and build factories or websites. Those who view asset market booms and following busts as "irrational" or "bubbles" point to the consequent investment booms and busts as examples of the social costs of inefficient markets, thereby endorsing the social value of more efficient markets.

Even without investment, more efficient prices provide better risk sharing, as in an endowment economy. If the owner of an apple tree and that of a pear tree hedge their risks by trading stock in the other tree, their risk-sharing improves when stock prices are more efficient. (Hirshleifer's, 1971, famous analysis stating that efficiency is only socially beneficial if production is involved did not treat such risk-sharing).

Information trading is central to "liquidity provision" and thus the success of markets for risk sharing. Markets such as Consumer Price Index, GDP futures or hurricane catastrophe options failed because there was not enough information trading. This is an important external benefit. Indeed, in the public forum, hedge funds and high-frequency traders primarily defend their activities by touting their "market making" and "liquidity provision" for small investors. (Of course, they are also pandering to their regulators' tastes here.)

The Puzzle of Information Trading

Still, the cacophony of trading seems like a lot of effort for these goals. The classic theory of finance predicts that information is perfectly reflected in prices, with no trading volume needed. Suppose Apple is trading at $500 per share, but you know that the iPhone 6 will make Apple worth $1,000 per share. If you approach an uninformed investor with an offer to buy Apple at $600 per share, the index investor should answer: "No, you must know something I don't know. I only buy and sell the entire index, so I don't lose to people like you." If you offer $700, the index investor answers: "I don't think you heard me. I only buy and sell the entire index." You keep trying, bidding the price up all the way to $1,000 per share, at which point you give up. The price rises, reflecting your information, but no trade occurred. This is a colloquial version of Milgrom and Stokey's (1982) famous no-trade theorem.

The theory that prices reflect information with zero trading volume is of course dramatically at odds with the facts. The classic theory also ignores costs. If information traders cannot earn positive alpha, and if producing information and trading on it takes any time and resources, the information traders won't bother, and nobody is left to make prices reflect information. For this reason, as Grossman and Stiglitz (1980) wrote, informationally efficient markets are impossible.

The standard compromise model (Grossman and Stiglitz 1980; Kyle 1985; and a huge literature) posits "informed" traders who receive a signal about a firm's value, "liquidity" traders who for unspecified reasons must trade, and "market makers" who intermediate, charging a bid-ask spread to defend themselves against the informed traders.

Now, all current theories of trading rely on some sort of "irrationality" or other artificial assumptions. "Liquidity traders" are the classic example. Other models, like Scheinkman and Xiong (2003), posit slightly irrational dogmatic beliefs so each information trader can believe he or she is smarter than average. Many models, such as Acharya and Pedersen (2005), write down overlapping generations of agents without bequests who die every week or so, forcing them to trade.

But these assumptions are convenient shortcuts for getting trading into the model for other purposes, such as studying price discovery and liquidity. They are not there to describe microfoundations of socially destructive trading that needs remediation by policy. The "irrationality" that breaks the no-trade theorem, or the irrationality of the liquidity traders, is not typically deeply micro-founded in the psychology literature, as in true behavioral finance. People live more than a week, and leave bequests.

The fact staring us in the face is that "price discovery," the process by which information becomes embedded in market prices, uses a *lot* of trading volume, and a lot of time, effort, and resources. And we are only beginning to understand it.

The empirical literature offers tantalizing glimpses of this process. A very small taste of this vast literature: The period after a news announcement often features high price volatility and trading volume, in which markets seem to be fleshing out what the news announcement actually means for the value of the security. For example, Lucca and Moench (2012, Figure 6) show a spike in stock-index trading volume and price volatility in the hours just *after* the Federal Reserve announcements of its interest rate decisions. The information is perfectly public. But the process of the market digesting its meaning, aggregating the opinions of its traders, and deciding what value the stock index should be with the new information, seems to need actual shares to trade hands.[4] Perhaps the common model of information—essentially, we all agree on the deck of cards, we just don't know which one was picked—is wrong.

Securities such as "on the run" or benchmark bonds, where "price discovery" takes place, have higher prices than otherwise identical securities. Traders are

[4] Banerjee and Kremer (2010) and Kim and Verrecchia (1991) offer models in which such disagreement about public information leads to trading volume.

willing to suffer lower average returns in order to participate in the information-trading game, in much the same way as money holders suffer lower returns for the transactions services money provides (see Cochrane 2003 and references therein). Similarly, "liquidity" seems to be extremely valuable to investors and has been so for a long time, even though none of us feel the need to trade every 10 minutes.

Markets in financial securities are set up, and exist, almost entirely to be markets for *information trading,* and high-frequency "liquidity provision," that we find hard to fathom. They are not really markets for the *securities* themselves. We could easily handle individuals' lifetime saving and dissaving needs, and firms' need to issue and retire equity, with orders-of-magnitude less volume, in much sleepier bank-like institutions. Yes, we could each avoid being the negative-alpha part of price discovery by only buying index funds. It's a bit of a puzzle that we don't. It's also a good thing we don't, or there would be no traders making prices efficient.

But as with active management, perhaps we should work just a little harder before dismissing the hundreds of years of trading activity, and the entire existence of the New York Stock Exchange, Chicago Mercantile Exchange, and other markets, as monuments to human folly, or before advocating regulations such as transactions taxes—the perennial favorite answer in search of a question—to reduce trading volume whose size, function, and operation we do not understand. Are we sure that they should not be transactions subsidies?

And before we deplore, it's worth remembering just how crazy passive indexing sounds to any market participant. "What," they might respond, "would you walk in to a wine store and say 'I can't tell good from bad, and the arbitrageurs are out in force. I sure won't pay you 1 percent for recommendations. Just give me one of everything'?"

High-Frequency Trading and Market-Making

It's especially hard to see why high-frequency trading is needed. Price discovery every millisecond doesn't seem necessary to guide corporate investment or individual risk sharing and hedging.

High-frequency trading reminds us in the extreme that the amount of trading based on a well-understood or "fundamental" piece of information about a company's cash flow is minuscule. Models in which an informed trader possesses a "signal" about the value of a liquidating dividend just don't describe the vast majority of trading. High-frequency traders do not trade on earnings reports 20 milliseconds ahead of the market.

Instead, high-frequency traders—and even most "low-frequency" day and week traders—look at patterns of prices, volumes, and past trading activity, not "information" or opinion about firm fundamentals.

They may describe their strategy as "statistical arbitrage," removing the small predictability of high-frequency price movements (and grossly misusing the term "arbitrage.") Sometimes they defend their social function as "market makers" or "liquidity providers." If so, market making is a far more dynamic process than simply posting bid-ask spreads, as the standard theory envisions! If you ask their critics, they are artfully front-running demand from less-sophisticated investors, subtracting "liquidity," worsening "price impact," choking bandwidth with quickly-canceled orders and removing the economic rewards to genuine information trading. Their activity may also answer the interesting question of how information spreads from one informed trade to the whole market. Somebody has to notice the price pattern and pile in.

However we come to understand these issues, the social costs and benefits of high-frequency trading are clearly not at all related to the minor (as a fraction of GDP) resources devoted to them—the cost of possibly useless fiber-optic cable, co-located servers, and the time of smart programmers who could be developing better iPhone games. The social question for high-frequency trading—like all of finance, really—is whether it screws up markets or makes them more efficient and "liquid."

There isn't yet much evidence or theory on this point, but isolated events suggest doubts about liquidity-provision and efficiency. For example, in the May 6, 2010 "Flash Crash," the Standard and Poor's 500 fell 6 percent in a few minutes after a large sell order arrived, and promptly recovered in less than an hour, only after a five minute trading halt. Kirilenko, Kyle, Samadi, and Tuzun (2011) who study this event (see their Figure 1) document that high-frequency traders absorbed demand for about four seconds before turning around and selling along with everyone else. On July 19, 2012, Coke, McDonalds, IBM, and Apple saw price sawtooths: sharp rises exactly on each hour, reversed by the next hour. Vigna and Lauricella (2012) offer some amazing graphs.[5] These movements were widely attributed to an algorithm placing big orders exactly on the hour—and other algorithms not picking up on the inefficient signal abundantly obvious to the human eye. These palpable inefficiencies suggest a market with very little "liquidity provision," not the opposite.

The structure of markets, with design and regulation stemming from the days of human trading, could be at fault. Prices must jump in discrete intervals—once 1/8 dollar, now 1 cent. Limit orders must be filled in strict time priority: if order A arrives before order B, order A must be filled completely and B gets nothing. Yet time is continuous. A's order need only arrive a millisecond before B's, and A wins the pot. (Traders report that the ability to quickly cancel limit orders that are in the back of the line is another advantage of very high speed.) You can see an arms race for speed emerge. It's worth spending a lot on computers to speed up trades by a few milliseconds.

If my hunch is correct, it suggests an obvious solution: Suppose that an exchange operated on a discrete clock, as a computer does in order to let signals settle down before processing them. The exchange could run a once-per-second, or even once-per-minute, matching process, with all orders received during the period treated equally. If there are more buy than sell at the crossing price, orders are

[5] The website http://www.nanex.net/FlashCrash/OngoingResearch.html is devoted to weird behavior in high-frequency markets.

filled proportionally. Such an exchange would eliminate extremely high-frequency trading, because there would be no gain or loss from acting faster than a minute.

Would this system be an improvement, to efficiency and liquidity? Would exchanges choose such systems if they were allowed to do so? The Taiwan Stock Exchange already matches limit orders once every 90 seconds (Barber, Lee, Liu, and Odean 2008). Is its performance atrociously worse? These are all good questions! High-frequency trading is a ripe area of research.

Housing, Consumer Credit, and the Size of Regulated Finance

The growth of housing finance and consumer credit raises a different set of issues. It's useful to divide the mortgage business into three parts: mortgage origination, mortgage refinancing, and mortgage-backed securities.

The increase in fees for residential loan origination is easily digested as the response to an increase in demand. The increase in housing demand may indeed not have been "socially optimal" (!). There are plenty of government policies and perhaps a few market dislocations to blame. But it doesn't make much sense to criticize growth in the financial industry for responding to this increase in demand, whatever its source, or for passing along the subsidized credit—which was and remains the government's explicit intention to increase—with the customary fee.

The large fees collected for refinancing mortgages are a bit more puzzling. US mortgages are strangely complicated, predominantly featuring fixed rates, no penalty for prepaying when interest rates fall, limited recourse, and a complex refinancing option. Other countries have gravitated to much simpler contracts. The now-familiar structure of US mortgages emerged after only the Great Depression, when new federal agencies started issuing them. Before the Great Depression, US mortgages lasted only five to ten years and required only the payment of interest. The principal was due at the end of the loan, and was typically refinanced (Green and Wachter 2005, p. 95). Today, the structure of mortgage contracts is pretty much dictated by what the government agencies that dominate the market will buy and guarantee.

These observations suggest that such complex contracts are not a market necessity. However, a glance at my cellphone contract and frequent flyer miles rules suggests to me that price discrimination by needless complexity might be part of the story as well.

Still, collecting fees when interest rates decline or consumers refinance is not conceptually part of GDP. They are state-dependent transfers dictated by the terms of an option contract. And we are unlikely to see a lot of refinancing as interest rates eventually rise.

There was a lot of financial innovation in mortgage-backed securities, some of which notoriously exploded. But here again, whether we spend a bit of GDP filling out forms or paying fees is clearly the least of the social benefit and cost questions. The "shadow banking" system was prone to a textbook systemic run, which happened. This fragility, not the size or fraction of GDP, is the important issue.

A good part of this innovation, such as creating off-balance-sheet, special-purpose vehicles and tailoring securities in order to game credit ratings, was clearly designed to engineer around ill-conceived regulations. That part counts as a regulatory failure needing reform, rather than a market failure needing additional regulation.

Yet much of this financial innovation has the potential to be of large social benefit. Suppose that mortgages were bundled into securities, intermediated by mutual funds whose values float, just like those of equity mutual funds, and held around the world in retirement accounts, pension funds, and our endowments' portfolios, without government guarantees at every step. This would be a terrific financial structure. Though mortgage-backed securities are a bit opaque, they are nowhere near as opaque as the entire balance sheet of, say, Citigroup. Furthermore, such a structure would be immune to runs, bankruptcies, and bailouts, thus requiring minimal regulation. And the fees required to fill out the mortgage-backed security paperwork would surely be less than the bank and regulatory paperwork, regulation, and compliance costs of the current system.

Concluding Remarks

The size and revenues of the finance industry increased because fee income for refinancing, issuing, and securitizing mortgages rose along with the rise in housing transactions and house prices, and because asset-management fee income rose along with a shift to professional management from "roll-your-own" portfolios and a rise in asset values. Compensation to employees with skills in short supply increased. Fee schedules themselves declined a bit. These facts suggest "demand shifted out," not "something big changed in the structure of this industry."

Demand that shifts out can shift back again. Demand for financial services evaporated with the decline in housing and asset values in the 2008 recession and subsequent period of sclerotic growth. Much of the "shadow banking system" has disappeared. For example, asset-backed commercial paper outstanding rose from $600 billion in 2001 to $1.2 trillion in 2007—and now stands at $300 billion. Financial credit market debt outstanding in the flow of funds rose from $8.6 trillion in 2000 to $17.1 trillion in 2008—and now stands at $13.8 trillion. Employment in financial activities rose from 7.7 million in 2000 to 8.4 million in 2007—and is now back to 7.7 million (according to the Bureau of Labor Statistics). Study of "why is finance so big," using data that stops in 2007, may soon take its place alongside studies of "why are Internet stocks so high" in 1999 or studies of "why is there a Great Moderation" in 2006.

An older literature on the size of the financial system, forgotten in the current debate, studies the socially inefficient resources devoted to cash management in the face of positive interest rates, measuring social costs as the area under the money demand curve. Lucas (2000) concluded that finance was about 1 percent of GDP too big by this measure. The fragility of those cash-management schemes can now be added to the list of social costs. Zero interest rates have eliminated these costs for now, and if the Fed continues to pay market interest on reserves, those costs can remain largely eliminated in the future.

The size question for the finance industry going forward, under the Dodd–Frank regulatory structure, is likely to be how many resources are devoted to regulation, regulatory compliance, lobbying to influence those regulations, and the distortions they induce. The social cost question remains how to create a financial system that is not prone to runs, crashes, and bailouts, even if that costs a few percentage points of GDP. Unless sovereign debt bites us first.

Many puzzles remain in the structure of the finance industry. The persistence of high-fee active management chosen by sophisticated institutional investors remains a puzzle. To some extent, as I have outlined, this pattern may reflect insurance provision, that is, the dynamic and multidimensional character of asset-market risk and risk premiums. To some extent, this puzzle also goes hand in hand with the puzzle of why price discovery seems to require so much active trading, and whether and how information trading provides valuable "liquidity." It is possible that there are far too *few* resources devoted to price discovery and market stabilization. In the financial crisis, we surely needed more pools of cash prepared to pounce on fire sales, and more opportunities for negative long-term views to express themselves.

Surveying the current economic literature on these issues, it is certain that we do not very well understand the price-discovery and trading mechanism, nor the economic forces that allowed high-fee active management to survive so long.

Unless we adopt the arrogant view that what we don't understand must be bad, it is clearly far too early to make pronouncements such as "There is likely too much high-cost, active asset management," or "Society would be better off if the cost of this management could be reduced." Such statements are not supported by theory or evidence. Nor is their not-so-subtle implication that resources devoted to greater regulation—by politicians and regulators no less naive than current investors, no less behaviorally-biased, armed with no better understanding than academic economists, and with much larger agency problems and institutional constraints—will improve matters. This proposition amounts to Samuel Johnson's dictum on second marriages, "the triumph of hope over experience."

References

Acharya, Viral V., and Lasse H. Pedersen. 2005. "Asset Pricing with Liquidity Risk." *Journal of Financial Economics* 77(2): 375–410.

Banerjee, Snehal, and Ilan Kremer. 2010. "Disagreement and Learning: Dynamic Patterns of Trade." *Journal of Finance* 65(4): 1269–1302.

Barber, Brad M., Yi-Tsung Lee, Yu-Jane Liu, and Terrance Odean. 2008. "Just How Much Do Individual Investors Lose by Trading?" *Review of Financial Studies* 22(2): 609–632.

Berk, Jonathan B. 2005. "Five Myths of Active Portfolio Management." *Journal of Portfolio Management* 31(3): 27–31.

Berk, Jonathan B., and Richard C. Green. 2004. "Mutual Fund Flows and Performance in Rational Markets." *Journal of Political Economy* 112(6): 1269–95.

Berk, Jonathan B., and Jules H. Van Binsbergen. 2012. "Measuring Managerial Skill in the Mutual Fund Industry." NBER Working Paper 18184.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52(1): 57–82.

Chevalier, Judith, and Glenn Ellison. 1997. "Risk Taking by Mutual Funds as a Response to Incentives." *Journal of Political Economy* 105(6): 1167–1200.

Cochrane, John H. 1991. "Production-Based Asset Pricing and the Link between Stock Returns and Economic Fluctuations." *Journal of Finance* 46(1): 207–234.

Cochrane, John H. 2003. "Stock as Money: Convenience Yield and the Tech-Stock Bubble." Chap. 12 in *Asset Price Bubbles* edited by William C. Hunter, George G. Kaufman, and Michael Pomerleano. Cambridge: MIT Press.

Cochrane, John H. 2011. "Discount Rates." *Journal of Finance* 66(4): 1047–1108.

Cochrane, John H. Forthcoming. "A Mean-Variance Benchmark for Intertemporal Portfolio Theory." *Journal of Finance.*

Fama, Eugene F., and Kenneth R. French. 1996. "Multifactor Explanations of Asset Pricing Anomalies." *Journal of Finance* 51(1): 55–84.

Fama, Eugene F., and Kenneth R. French. 2006. "Dissecting Anomalies." *Journal of Finance* 63(4): 1653–78.

Fama, Eugene F., and Kenneth R. French. 2010. "Luck versus Skill in the Cross-Section of Mutual Fund Returns." *Journal of Finance* 65(4): 1915–47.

Frazzini, Andrea, and Lasse Heje Pedersen. 2011a. "Betting against Beta." Paper available at http://www.econ.yale.edu/~af227.

Frazzini, Andrea, and Lasse Heje Pedersen. 2011b. "Embedded Leverage." Paper available at http://www.econ.yale.edu/~af227.

French, Kenneth R. 2008. "Presidential Address: The Cost of Active Investing." *Journal of Finance* 63(4): 1537–73.

Goetzmann, William N., and Sharon Oster. 2012. "Competition among University Endowments." NBER Working Paper 18173.

Green, Richard K., and Susan M. Wachter. 2005. "The American Mortgage in Historical and International Context." *Journal of Economic Perspectives* 19(4): 93–114.

Greenwood, Robin, and David Scharfstein. 2012. "The Growth of Modern Finance." July, 2012. http://www.people.hbs.edu/dscharfstein/Growth_of_Modern_Finance.pdf.

Grossman, Sanford G., and Joseph E. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70(3): 393–408

Hirshleifer, Jack. 1971. "The Private and Social Value of Information and the Reward to Inventive Activity." *American Economic Review* 61(4): 561–74.

Kim, Oliver, and Robert E. Verrecchia. 1991. "Trading Volume and Price Reactions to Public Announcements." *Journal of Accounting Research* 29(2): 302–21.

Kirilenko, Andrei A., Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2011. "The Flash Crash: The Impact of High Frequency Trading on an Electronic Market." http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1686004.

Kyle, Albert S. 1985. "Continuous Auctions and Insider Trading." *Econometrica* 53(6): 1315–36.

Lamont, Owen A. 2012. "Go Down Fighting: Short Sellers vs. Firms." *Review of Asset Pricing Studies* 2(1): 1–30.

Lamont, Owen A., and Richard H. Thaler. 2003. "Can the Market Add and Subtract? Mispricing in Tech Stock Carve-Outs." *Journal of Political Economy* 111(2): 227–68.

Lucas, Robert E., Jr. 2000. "Inflation and Welfare." *Econometrica* 68(2): 247–74.

Lucca, David O., and Emanuel Moench. 2012. "The Pre-FOMC Announcement Drift." http://www.newyorkfed.org/research/staff_reports/sr512.html.

Merton, Robert C. 1971. "Optimum Consumption and Portfolio Rules in a Continuous Time Model." *Journal of Economic Theory* 3(4): 373–413.

Milgrom, Paul, and Nancy L. Stokey. 1982. "Information, Trade, and Common Knowledge." *Journal of Economic Theory* 26(1): 17–27.

Pensions & Investments. 2011. "Chart: Top Paid CIOs of Tax-Exempt Institutions." November 7. http://www.pionline.com/article/20111107/CHART04/111109905.

Scheinkman, Jose, and Wei Xiong. 2003. "Overconfidence and Speculative Bubbles." *Journal of Political Economy* 111(6): 1183–1219.

Vigna, Paul, and Tom Lauricella. 2012. "Sawtooth Trading Hits Coke, IBM, McDonald's, and Apple Shares." *Wall Street Journal,* July 19. http://blogs.wsj.com/marketbeat/2012/07/19/sawtooth-trading-hits-coke-ibm-mcdonalds-and-apple-shares/.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 51–72

Moore's Law versus Murphy's Law: Algorithmic Trading and Its Discontents[†]

Andrei A. Kirilenko and Andrew W. Lo

Over the past four decades, the remarkable growth of the semiconductor industry as embodied by Moore's Law has had enormous effects on society, influencing everything from household appliances to national defense. The implications of this growth for the financial system has been profound, as well. Computing has become faster, cheaper, and better at automating a variety of tasks, and financial institutions have been able to greatly increase the scale and sophistication of their services. At the same time, population growth combined with the economic complexity of modern society has increased the demand for financial services. After all, most individuals are born into this world without savings, income, housing, food, education, or employment; all of these necessities require financial transactions.

It should come as no surprise then that the financial system exhibits a Moore's Law of its own—from 1929 to 2009 the total market capitalization of the US stock market has doubled every decade. The total trading volume of stocks in the Dow Jones Industrial Average doubled every 7.5 years during this period, but in the most recent decade, the pace has accelerated: now the doubling occurs every 2.9 years, growing almost as fast as the semiconductor industry. But the financial industry

■ *Andrei A. Kirilenko is the Professor of the Practice of Finance, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts. During 2010–2012, Kirilenko served as the Chief Economist of the US Commodity Futures Trading Commission, Washington, DC. Andrew W. Lo is the Charles E. and Susan T. Harris Professor, Director of the Laboratory for Financial Engineering at Sloan School of Management, and a Principal Investigator at the Computer Science and Artificial Intelligence Laboratory, all at the Massachusetts Institute of Technology, Cambridge, Massachusetts. Lo is also Chairman and Chief Investment Strategist, AlphaSimplex Group, LLC, an investment management firm. Their email addresses are ak67@mit.edu and alo@mit.edu.*

[†] To access the disclosure statements, visit http://dx.doi.org/10.1257/jep.27.2.51.

doi=10.1257/jep.27.2.51

differs from the semiconductor industry in at least one important respect: human behavior plays a more significant role in finance. As the great physicist Richard Feynman once said, "Imagine how much harder physics would be if electrons had feelings." While financial technology undoubtedly benefits from Moore's Law, it must also contend with Murphy's Law, "whatever can go wrong will go wrong," as well as its technology-specific corollary, "whatever can go wrong will go wrong faster and bigger when computers are involved."

A case in point is the proliferation of high-frequency trading in financial markets, which has raised questions among regulators, investors, and the media about how this technology-powered innovation might affect market stability. Largely hidden from public view, this relatively esoteric and secretive cottage industry made headlines on May 6, 2010, with the so-called "Flash Crash," when the prices of some of the largest and most actively traded companies in the world crashed and recovered in a matter of minutes. Since then, a number of high-profile technological malfunctions, such as the delayed Facebook initial public offering in March 2012 and an electronic trading error by Knight Capital Group in August 2012 that cost the company $400+ million, have only added fuel to the fire. Algorithmic trading—the use of mathematical models, computers, and telecommunications networks to automate the buying and selling of financial securities—has arrived, and it has created new challenges as well as new opportunities for the financial industry and its regulators.

Algorithmic trading is part of a much broader trend in which computer-based automation has improved efficiency by lowering costs, reducing human error, and increasing productivity. Thanks to the twin forces of competition and innovation, the drive toward "faster, cheaper, and better" is as inexorable as it is profitable, and the financial industry is no stranger to such pressures. However, what has not changed nearly as much over this period is the regulatory framework that is supposed to oversee such technological and financial innovations. For example, the primary set of laws governing the operation of securities exchanges is the Securities Exchange Act of 1934, which was enacted well before the arrival of digital computers, electronic trading, and the Internet. Although this legislation has been amended on many occasions to reflect new financial technologies and institutions, it has become an increasingly cumbersome patchwork quilt of old and new rules based on increasingly outdated principles, instead of an integrated set of modern regulations designed to maintain financial stability, facilitate capital formation, and protect the interests of investors. Moreover, the process by which new regulations are put in place or existing regulations are amended is slow and subject to the vagaries of politics, intense lobbying by the industry, judicial challenges, and shifting public sentiment, all of which may be particularly problematic for an industry as quickly evolving and highly competitive as financial services.

In this paper, we provide a brief survey of algorithmic trading, review the major drivers of its emergence and popularity, and explore some of the challenges and unintended consequences associated with this brave new world. There is no doubt that algorithmic trading has become a permanent and important part of the financial landscape, yielding tremendous cost savings, operating efficiency, and scalability to every financial market it touches. At the same time, the financial system has become much more of a *system* than ever before, with globally interconnected counterparties and privately-owned and -operated infrastructure that facilitates tremendous integration during normal market conditions, but which spreads dislocation rapidly during periods of financial distress. A more systematic and adaptive approach to regulating this system is needed, one that fosters the technological advances of the industry while protecting those who are not as technologically advanced. We conclude by proposing "Financial Regulation 2.0," a set of design principles for regulating the financial system of the Digital Age.

A Brief Survey of Algorithmic Trading

Three developments in the financial industry have greatly facilitated the rise of algorithmic trading over the last two decades. The first is the fact that the financial system is becoming more complex over time, not less. Greater complexity is a consequence of general economic growth and globalization in which the number of market participants, the variety of financial transactions, the levels and distribution of risks, and the sums involved have also grown. And as the financial system becomes more complex, the benefits of more highly developed financial technology become greater and greater and, ultimately, indispensable.

The second development is the set of breakthroughs in the quantitative modeling of financial markets, the "financial technology" pioneered over the past three decades by the giants of financial economics: Black, Cox, Fama, Lintner, Markowitz, Merton, Miller, Modigliani, Ross, Samuelson, Scholes, Sharpe, and others. Their contributions laid the remarkably durable foundations on which modern quantitative financial analysis is built, and algorithmic trading is only one of the many intellectual progeny that they have fathered.

The third development is an almost parallel set of breakthroughs in computer technology, including hardware, software, data collection and organization, and telecommunications, thanks to Moore's Law. The exponential growth in computing power per dollar and the consequences for data storage, data availability, and electronic interconnectivity have irrevocably changed the way financial markets operate.

A deeper understanding of the historical roots of algorithmic trading is especially important for predicting where it is headed and formulating policy and regulatory recommendations that affect it. In this section, we describe five major developments that have fueled its growing popularity: quantitative models in finance, the emergence and proliferation of index funds, arbitrage trading activities, the push for lower costs of intermediation and execution, and the proliferation of high-frequency trading.

Quantitative Finance

The most obvious motivation for algorithmic trading is the impressive sequence of breakthroughs in quantitative finance that began in the 1950s with portfolio optimization theory. In his pioneering PhD thesis, Harry Markowitz (1952) considered how an investor should allocate his wealth over *n* risky securities so as to maximize his expected utility of total wealth. Under some assumptions, he shows that this is equivalent to maximizing the expected value of a quadratic objective function of the portfolio's return which, in turn, yields a mean–variance objective function. The solution to this well-posed optimization problem may be considered the very first algorithmic trading strategy—given an investor's risk tolerance and the means, variances, and covariances of the risky assets, the investor's optimal portfolio is completely determined. Thus, once a portfolio has been established, the algorithmic trading strategy—the number of shares of each security to be bought or sold—is given by the difference between the optimal weights and the current weights. More importantly, portfolio optimization leads to an enormous simplification for investors with mean–variance preferences: all such investors should be indifferent between investing in *n* risky assets and investing in one specific portfolio of these *n* assets, often called the "tangency portfolio" because of the geometry of mean–variance analysis.[1] This powerful idea is often called the "Two-Fund Separation Theorem" because it implies that a riskless bond and a single mutual fund—the tangency portfolio—are the only investment vehicles needed to satisfy the demands of all mean–variance portfolio optimizers, an enormous simplification of the investment problem.

The second relevant milestone in quantitative finance was the development of the Capital Asset Pricing Model (CAPM) by Sharpe (1964), Lintner (1965), and Mossin (1966) in the 1960s, and the intense empirical and econometric investigations it launched in the following two decades. These authors took portfolio optimization as their starting point and derived a remarkably simple yet powerful result: if all investors hold the same tangency portfolio, albeit in different dollar amounts, then this tangency portfolio can only be one portfolio: the portfolio of all assets, with each asset weighted according to its market capitalization. In other words, the tangency portfolio is the total market portfolio. This more-specific form of the Two-Fund Separation Theorem was a critical milestone in both academia and industry, generating several new directions of research as well as providing the foundations for today's trillion-dollar index-fund industry (discussed in the next section).

The third milestone occurred in the 1970s and was entirely statistical and computational. To implement portfolio optimization and the Capital Asset Pricing Model, it was necessary to construct timely estimates of the expected returns and the covariance matrix of all traded equities. This seemed like an impossible task in the 1970s because of the sheer number of securities involved—almost 5,000 stocks on the New York, American, and NASDAQ Stock Exchanges—and the numerical computations involved in estimating all those parameters. For example, a 5,000-by-5,000 covariance matrix contains 12,497,500 unique parameters. Moreover, because the maximum rank of the standard covariance-matrix estimator is simply the number of time series observations used, estimates of this 5,000-by-5,000

[1] The set of mean-variance-optimal portfolios forms a curve when plotted in mean–variance space, and the portfolio that allows mean–variance optimizers to achieve the highest expected return per unit of risk is attained by the portfolio that is tangent to the line connecting the risk-free rate of return to the curve.

matrix will be "singular" (meaning not invertible) for all sample sizes of daily or monthly stock returns less than 5,000. Singularity is particularly problematic for employing Markowitz-type mean–variance optimization algorithms which depend on the inverse of the covariance matrix.

These challenges were met elegantly and decisively in the 1970s by Rosenberg's (1974) linear multifactor risk model in which individual stock returns were assumed to be linearly related to a smaller number K of common "factors." The existence of such a linear relation implies that the total number of unknown covariance-matrix parameters to be estimated is now $nK + K(K + 1)/2 + n$ instead of $n(n - 1)/2$, which increases linearly in n instead of as n^2. In contrast to the 12,497,500 unique parameters in the case of 5,000 stocks, a linear factor model with 50 factors requires only 256,275 parameters—a 50-fold reduction!

Rosenberg took his ideas one step further in 1975 by founding a commercial venture—Barr Rosenberg and Associates, or Barra—that provided clients with timely estimates of covariance matrices for US equities, as well as portfolio optimization software so they could implement Markowitz-style mean-variance-optimal portfolios. It is no exaggeration that Barra's software platform was largely responsible for popularizing algorithmic equity trading—particularly portfolio optimization—among institutional investors and portfolio managers throughout the world. More frequent estimation of optimal portfolios also meant that portfolio managers needed to trade more frequently. As a result, trading volumes began to rise disproportionately faster than the number of newly created securities.

The fourth milestone came in 1973 with the publication of the Black and Scholes (1973) and Merton (1973) articles on the pricing of options and other derivative securities. Although these two seminal articles contained the celebrated Black–Scholes/Merton option-pricing formula—for which Merton and Scholes shared the Nobel prize in economics in 1997—an even more influential idea to come out of this research program was Merton's (1973) insight that under certain conditions, the frequent trading of a small number of long-lived securities can create new investment opportunities that would otherwise be unavailable to investors. These conditions—now known collectively as *dynamic spanning* or *dynamically complete markets*—and the corresponding asset-pricing models on which they are based, have generated a rich literature and a multi-trillion-dollar derivatives industry. The financial services industry has subsequently written hundreds of cookbooks with thousands of recipes describing how to make complex and sometimes exotic dishes such as swaps, caps, collars, swaptions, knock-out and rainbow options, and many others out of simple ingredients—stocks and bonds—by combining them in prescribed quantities and stirring (trading) the mixture frequently to make them as appetizing as possible to investors.

Index Funds

One of the most enduring legacies of Markowitz, Sharpe, Lintner, Tobin, and Mossin is the idea of "passive" investing through index funds. The recipe for an index fund is now well-known: define a collection of securities by some set of easily observable attributes, construct a portfolio of such securities weighted by their market capitalizations, and add and subtract securities from this collection from time to time to ensure that the portfolio continues to accurately reflect the desired attributes.

The original motivation behind fixing the set of securities and value-weighting them was to reduce the amount of trading needed to replicate the index in a cash portfolio. Apart from the occasional index addition and deletion, a value-weighted portfolio need never be rebalanced since the weights automatically adjust proportionally as market valuations fluctuate. These "buy-and-hold" portfolios are attractive not only because they keep trading costs to a minimum, but also because they are simpler to implement from an operational perspective. It is easy to forget the formidable challenges posed by the back-office, accounting, and trade reconciliation processes for even moderate-sized portfolios in the days before personal computers, automated order-generating engines, and electronic trading platforms. A case in point is the precursor to the very first index mutual fund, a $6 million equal-weighted portfolio of 100 New York Stock Exchange (NYSE) equities managed by Wells Fargo Bank for Samsonite's pension fund starting in 1969. An equal-weighted portfolio—a portfolio in which equal dollar amounts are invested in each security—does not stay equally weighted as prices fluctuate, and the process of rebalancing a portfolio of 100 stocks back to equal weighting at the end of each month was such an operational nightmare back then that the strategy was eventually abandoned in favor of a value-weighted portfolio (Bogle 1997). Since then, most investors and managers equate "passive" investing with low-cost, static, value-weighted portfolios (portfolios in which the dollar amount invested in each security is proportional to the total market capitalization of the company issuing that security).

However, with the many technological innovations that have transformed the financial landscape over the last three decades, the meaning of passive investing has changed. A functional definition of passive investing is considerably more general: an investment process is "passive" if it does not require any discretionary human intervention—that is, if it is based on a well-defined and transparent algorithm. Such a definition decouples active investing from active trading; today, a passive investor may be an active trader to minimize transaction costs, manage risks more adroitly, participate in new investment opportunities such as initial public offerings, or respond more quickly to changing objectives and market conditions. Moreover, new investment products such as target-date funds, exchange-traded funds, and strategy indexes such as 130/30, currency carry-trade, hedge-fund replication, and trend-following futures strategies are growing in popularity and acceptance among passive investors despite the active nature of their trading, thanks to the automation facilitated by algorithms. At the same time, the much more active participation of investors has created new technological challenges for the issuers of new financial instruments. We provide an example of this later in this paper when discussing the Facebook and BATS initial public offerings.

Arbitrage Trading

Arbitrage strategies are among the most highly visible applications of algorithmic trading over the past three decades. These strategies are routinely implemented by broker-dealers, hedge funds, and institutional investors with the sole objective of generating profits with lower risk than traditional investments. Arbitrage trading is as old as financial markets, but using algorithms to identify and exploit arbitrage-trading opportunities is a thoroughly modern invention, facilitated by the use of computers, applications of probability and statistics, advances in telecommunications, and the development of electronic markets.

The most common form of algorithmic arbitrage trading is a transaction that attempts to exploit situations where two securities that offer identical cashflows have different market prices. The law of one price implies that such opportunities cannot persist, because traders will quickly construct arbitrage portfolios in which the lower-priced asset is purchased and the higher-priced asset is sold (or shorted) yielding a positive and riskless profit by assumption (because the underlying cashflows of the two securities are assumed to be identical). More generally, an arbitrage strategy involves constructing a portfolio of multiple securities such that the combined cashflows are riskless, and if the cost of constructing such a portfolio is nonzero for reasons other than trading costs, then there exists a version of the arbitrage strategy that generates positive riskless profits, which is a definition of an arbitrage opportunity.

Violations of the law of one price have been routinely exploited in virtually every type of financial market ranging from highly liquid securities such as foreign currencies and exchange-traded futures to highly illiquid assets such as real estate and emerging-market debt. However, in most practical settings, pure arbitrages do not exist because there are subtle differences in securities that cause their prices to differ despite seemingly identical cashflows, like differences in transactions costs, liquidity, or credit risk. The fact that hedge funds like Long-Term Capital Management have suffered severe losses from arbitrage strategies implies that such strategies are not, in fact, pure arbitrages or completely riskless profit opportunities.

However, if the statistical properties of the arbitrage portfolios can be quantified and managed, the risk/reward profiles of these strategies might be very attractive to investors with the appropriate tolerance for risk. These considerations led to the development of a new type of proprietary trading strategy in the 1980s, so-called "statistical arbitrage strategies" in which large portfolios of equities were constructed to maximize expected returns while minimizing volatility. The risks embedded in statistical arbitrage strategies are inherently different from market risk because arbitrage portfolios are, by construction, long and short, and hence they can be profitable during market downturns. This property provides attractive diversification benefits to institutional investors, many of whom have the majority of their assets in traditional long-only portfolios of stocks and bonds. The details of statistical arbitrage strategies are largely unknown because proprietary traders cannot patent such strategies, and thus they employ trade secrecy to protect their intellectual property. However, simple versions of such strategies have been proposed and studied by Lehmann (1990), Lo and MacKinlay (1990), and Khandani and Lo (2007, 2011), and we provide a more detailed exposition of them in the sections that follow.

Apart from the attractive risk/reward profile they offer to investors and portfolio managers, arbitrage strategies play two other critical roles in the financial system: liquidity provision and price discovery. The presence of arbitrageurs almost always increases the amount of trading activity, and larger volume is often interpreted as greater liquidity, meaning that investors often can buy or sell securities more quickly, in larger quantities, and with lower price impact. Moreover, because arbitrage trading exploits temporary mispricings, it tends to improve the informational efficiency of market prices (assuming that the mispricings are genuine). However, if arbitrageurs become too dominant in any given market, they can create systemic instabilities. We provide an example of this in our later discussion of the so-called "Quant Meltdown" in August 2007.

Automated Execution and Market Making

Algorithmic trading is also central to the automation of large buy and sell orders of publicly traded securities such as exchange-traded equities. Because even the most actively traded stocks have downward-sloping demand curves over a short period of time, executing a large "parent" order in a single transaction is typically more costly than breaking up the order into a sequence of smaller "child" orders. The particular method for determining the timing and sizes of these smaller orders is called an "execution strategy," and optimal execution strategies can be derived by specifying an objective function and a statistical model for stock-price dynamics.

For example, Bertsimas and Lo (1998) consider the problem of minimizing the expected cost of acquiring S_o shares of a given stock over T discrete trades. If S_o is a small number, like a "round lot" of 100 shares, then the entire block can be executed in a single trade. However, institutional investors must often trade hundreds of thousands of shares as they rebalance multi-billion-dollar portfolios. By modeling the short-run demand curve for each security to be traded—also known as the "price-impact function"—as well as other state variables driving price dynamics, Bertsimas and Lo (1998) are able to derive the expected-cost-minimizing sequence of trades as a function of those state variables using stochastic dynamic programming. These automated execution algorithms can be computationally quite complex for large portfolios of diverse securities, and are ideally suited for automation because of the accuracy and significant cost savings that they offer, especially when compared to human traders attempting to do this manually. However, under certain market conditions, automated execution of large orders can create significant feedback-loop effects that cascade into systemic events as in the case of the so-called "Flash Crash" of May 6, 2010, which we discuss in the next section.

A closely related activity to automated execution is market making, when an intermediary participates in buying and selling securities to smooth out temporary imbalances in supply and demand because buyers and sellers do not always arrive at the same time. A participant of a trading venue, typically a broker-dealer, can voluntarily apply to register as a designated market maker on a security-by-security basis. To qualify, a potential market maker must satisfy certain net capital requirements and be willing to provide continuous two-sided quotes during trading hours, which means being willing to purchase securities when the public wishes to sell, and to sell securities when the public wishes to buy. Registration does not guarantee profits or customer order flow; it only provides lower trading fees and a designation that can help attract orders from potential customers. Note that participants need not register to function as market makers. Market making is a risky activity because of price fluctuations and adverse selection—prices may suddenly move against market makers and force them to unwind their proprietary positions at a loss. To protect themselves against possible losses, market makers demand compensation, typically in the form of a spread that they charge buyers over sellers known as the "bid–offer spread."

A typical market-making algorithm submits, modifies, and cancels limit orders to buy and sell a security with the objective of regularly capturing the bid–offer spread and liquidity rebates (payments made to participants who provide liquidity to the market), if any, while also continuously managing risky inventory, keeping track of the demand–supply imbalance across multiple trading venues, and calculating the costs of doing business, including trading and access fees, margin requirements, and the cost of capital. As a result, automation of the trading process means that the rewards from market making activities accrue not necessarily to those who register with the exchanges as their designated market makers, but to those with the best connectivity, best algorithms, and best access to customer order flow.

The central issue with respect to algorithmic market making is whether this activity has improved overall market quality, thus allowing investors to raise capital and manage risks more efficiently. To analyze this issue, Hendershott, Jones, and Menkveld (2011) study the introduction of "autoquoting"—the automated transmission of improved terms of trade for larger trade sizes—that was introduced in 2003 on the New York Stock Exchange. Autoquoting did favor algorithmic traders because they could receive valuable information about changes in the order book faster than humans, but did not otherwise alter the advantages and obligations of the NYSE-designated specialists. The authors show that the introduction of autoquoting increased the informativeness of quoted prices, narrowed bid–offer spreads, and reduced the degree of adverse selection associated with trading. At the same time, automation makes technological glitches in the ultracompetitive business of market making extremely costly. We illustrate this point later in the paper with an example of an algorithmic market maker whose fate was sealed minutes after it launched a new trading algorithm.

High-Frequency Trading

A relatively recent innovation in automated financial markets is a blend of technology and hyperactive trading activity known as "high-frequency trading"—a form of automated trading that takes advantage of innovations in computing and telecommunication to consummate millions upon millions of trades per day. High-frequency trading is now estimated to account for 40 to 60 percent of all trading activity across the universe of financial markets, including stocks, derivatives, and liquid foreign currencies (Tabb 2012). However, the number of entities that engage in high-frequency trading is reportedly quite small and what is known about them is not particularly illuminating. Baron, Brogaard, and Kirilenko (2012) examine high-frequency trading in the "E-mini S&P 500 futures contract," an extremely popular futures contract on the Standard & Poor's 500 index that owes its name to the fact that it is electronically traded and in smaller denominations than the traditional S&P 500 index futures contract. Their study finds that high-frequency traders (as designated by their trading activity) earn large, persistent profits while taking very little risk. In contrast to a number of public claims, high-frequency traders do not as a rule engage in the provision of liquidity like traditional market makers. In fact, those that do not provide liquidity are the most profitable and their profits increase with the degree of "aggressive," liquidity-taking activity.

High-frequency trading is a recent innovation in financial intermediation that does not fit neatly into a standard liquidity-provision framework. While the net contribution of high-frequency trading to market dynamics is still not fully understood, their mere presence has already shaken the confidence of traditional market participants in the stability and fairness of the financial market system as a whole. Recent revelations of manipulative trading activity, discussed later in this paper, have only added fuel to the debate about the usefulness of high-frequency trading.

Ghosts in the Machine

As in every other industry that has reduced its costs via automation, the financial services industry has also been transformed by technology. In the modern trading environment, an investor's trading strategy—whether to liquidate a large position, to make markets, or to take advantage of arbitrage opportunities—is typically executed by an automated trading system. Such systems are responsible for the initiation of trading instructions, communication with one or more trading platforms, the processing of market data, and the confirmation of trades. But technology that supersedes human abilities often brings unintended consequences, and algorithmic trading is no exception. A chainsaw allows us to clear brush much faster than a hand saw, but chainsaw accidents are much more severe than handsaw accidents. Similarly, automated trading systems provide enormous economies of scale and scope in managing large portfolios, but trading errors can now accumulate losses at the speed of light before they're discovered and corrected by human oversight. Indeed, the enhanced efficiency, precision, and scalability of algorithms may diminish the effectiveness of those risk controls and systems safeguards that rely on experienced human judgment and are applied at human speeds. While technology has advanced tremendously over the last century, human cognitive abilities have been largely unchanged over the last several millennia. Thus, due to the very success of algorithmic trading, humans have been pushed to the periphery of a much faster, larger, and more complex trading environment.

Moreover, in a competitive trading environment, increased speed of order initiation, communication, and execution become a source of profit opportunities for the fastest market participants. Given these profit opportunities, some market participants, who either trade on their own account or provide execution services to their customers, may choose to engage in a "race to the bottom," forgoing certain risk controls that may slow down order entry and execution. This vicious cycle can lead to a growing misalignment of incentives as greater profits accrue to the fastest market participants with less-comprehensive safeguards, and may become a significant source of risk to the stability and resilience of the entire financial system.

In this section, we review five specific incidents that highlight these new vulnerabilities created or facilitated by algorithmic trading. We consider them in approximate chronological order to underscore the progression of technology and the changing nature of the challenges that financial innovation can bring.

August 2007: Arbitrage Gone Wild

Beginning on Monday, August 6, 2007, and continuing through Thursday, August 9, some of the most successful hedge funds in the industry suffered record losses. The *Wall Street Journal* reported on August 10, 2007: "After the close of trading, Renaissance Technologies Corp., a hedge-fund company with one of the best records in recent years, told investors that a key fund has lost 8.7% so far in August and is down 7.4% in 2007. Another big fund company, Highbridge Capital Management, told investors its Highbridge Statistical Opportunities Fund was down 18% as of the 8th of the month, and was down 16% for the year. The $1.8 billion publicly traded Highbridge Statistical Market Neutral Fund was down 5.2% for the month as of Wednesday . . . Tykhe Capital, LLC—a New York-based quantitative, or computer-driven, hedge-fund firm that manages about $1.8 billion—has suffered losses of about 20% in its largest hedge fund so far this month . . ." (Zuckerman, Hagerty, and Gauthier-Villars 2007). On August 14, the *Wall Street Journal* reported that the Goldman Sachs Global Equity Opportunities Fund "lost more than 30% of its value last week . . ." (Sender, Kelly, and Zuckerman 2007). What made these losses even more extraordinary was the fact that they seemed to be concentrated among quantitatively managed equity market-neutral or "statistical arbitrage" hedge funds, giving rise to the monikers "Quant Meltdown" and "Quant Quake" of 2007.

Because of the secretive nature of hedge funds and proprietary trading firms, no institution suffering such losses was willing to comment publicly on this extraordinary event at the time. To address this lack of transparency, Khandani and Lo (2007) analyzed the Quant Meltdown of August 2007 by simulating the returns of the contrarian trading strategy of Lehmann (1990) and Lo and MacKinlay (1990), and proposed the "Unwind Hypothesis" to explain the empirical facts (see also Goldman Sachs Asset Management 2007; Rothman 2007a, b, c). This hypothesis suggests that the initial losses during the second week of August 2007 were due to the forced liquidation of one or more large equity market-neutral portfolios, primarily to raise cash or reduce leverage, and the subsequent price impact of this massive and sudden unwinding caused other similarly constructed portfolios to experience losses. These losses, in turn, caused other funds to deleverage their portfolios, yielding additional price impact that led to further losses, more deleveraging, and so on. As with Long-Term Capital Management and other fixed-income arbitrage funds in August 1998, the deadly feedback loop of coordinated forced liquidations leading to the deterioration of collateral value took hold during the second week of August 2007, ultimately resulting in the collapse of a number of quantitative equity market-neutral managers, and double-digit losses for many others.

This Unwind Hypothesis underscores the apparent commonality among quantitative equity market-neutral hedge funds and the importance of liquidity in determining market dynamics. In a follow-on study, Khandani and Lo (2011) used transactions data from July to September 2007 to show that the unwinding likely began in July and centered on securities that shared certain common traits such as high or low book-to-market ratios, because such factors were used by many quantitative portfolio managers attempting to exploit the same empirical anomalies.

In retrospect, we now realize that the Quant Meltdown of August 2007 was only one of a series of crises that hit financial markets during the 2007–2008 crisis period. In fact, after the close of trading on August 9, 2007, central banks from around the world engaged in a highly unusual coordinated injection of liquidity in financial markets, not because of equity markets, but because of a so-called "run on repo" when the interbank short-term financing market broke down (Gorton and Metrick 2012). The summer of 2007 ushered in a new financial order in which the "crowded trade" phenomenon—where everyone rushes to the exit doors at the same time—now applied to entire classes of portfolio strategies, not just to a collection of overly popular securities. In much the same way that a passing speedboat can generate a wake with significant consequences for other ships in a crowded harbor, the scaling up and down of portfolios can affect many other portfolios and investors. Algorithmic trading greatly magnifies the impact of these consequences.

May 6, 2010: The Perfect Financial Storm

In the course of 33 minutes starting at approximately 1:32 pm central time, US financial markets experienced one of the most turbulent periods in their history. The Dow Jones Industrial Average experienced its biggest one-day point decline on an intraday basis in its entire history and the stock prices of some of the world's largest companies traded at incomprehensible prices: Accenture traded at a penny a share, while Apple traded at $100,000 per share. Because these dramatic events happened so quickly, the events of May 6, 2010, have become known as the "Flash Crash."

The subsequent investigation by the staffs of the Commodity Futures Trading Commission (CFTC) and Securities and Exchange Commission (SEC) concluded that these events occurred not because of any single organization's failure, but rather as a result of seemingly unrelated activities across different parts of the financial system that fed on each other to generate a perfect financial storm (CFTC/SEC 2010). An automated execution algorithm on autopilot, a game of "hot potato" among high-frequency traders, cross-market arbitrage trading, and a practice by market makers to keep placeholder bid–offer "stub quotes" all conspired to create a breathtaking period of extreme volatility.

Kirilenko, Kyle, Samadi, and Tuzun (2011) analyzed the Flash Crash and found that a rapid automated sale of 75,000 E-mini S&P 500 June 2010 stock index futures contracts (worth about $4.1 billion) over an extremely short time period created a large order imbalance that overwhelmed the small risk-bearing capacity of financial intermediaries—that is, the high-frequency traders and market makers. After buying the E-mini for about 10 minutes, high-frequency traders reached their critical inventory levels and began to unwind their long inventory quickly and aggressively at a key moment when liquidity was sparse, adding to the downward pressure. High-frequency traders rapidly passed contracts back and forth, contributing to the "hot potato" effect that drove up trading volume, exacerbating the volatility.

Meanwhile, cross-market arbitrage trading algorithms rapidly propagated price declines in the E-mini futures market to the markets for stock index exchange-traded funds like the Standard & Poor's Depository Receipts S&P 500, individual stocks, and listed stock options. According to the interviews conducted by the SEC staff, cross-market arbitrage firms "purchased the E-Mini and contemporaneously sold Standard & Poor's Depository Receipts S&P 500, baskets of individual securities, or other equity index products" (CFTC/SEC 2010). As a result, a liquidity event in the futures market triggered by an automated selling program cascaded into a systemic event for the entire US financial market system.

As the periods during which short-term liquidity providers are willing to hold risky inventory shrink to minutes if not seconds, Flash-Crash-type events—extreme short-term volatility combined with a rapid spike in trading volume—can easily be generated by algorithmic trading strategies seeking to quickly exploit temporarily favorable market conditions.

March and May 2012: Pricing Initial Public Offerings in the Digital Age

On Friday, May 18th, 2012, the social networking pioneer, Facebook, had the most highly anticipated initial public offering in recent financial history. With over $18 billion in projected sales, Facebook could easily have listed on the NYSE along with larger blue-chip companies like Exxon and General Electric, so Facebook's choice to list on NASDAQ instead was quite a coup for the technology-savvy exchange. Facebook's debut was ultimately less impressive than most investors had hoped, but its lackluster price performance was overshadowed by an even more disquieting technological problem with its opening. An unforeseen glitch in NASDAQ's system for initial public offerings interacted unexpectedly with trading behavior to delay Facebook's opening by 30 minutes, an eternity in today's hyperactive trading environment.

As the hottest initial public offering of the last ten years, Facebook's opening attracted extraordinary interest from investors and was expected to generate huge order flows, but NASDAQ prided itself on its ability to handle high volumes of trades so capacity was not a concern. NASDAQ's IPO Cross software was reportedly able to compute an opening price from a stock's initial bids and offers in less than 40 microseconds (a human eyeblink lasts 8,000 times as long). However, on the morning of May 18, 2012, interest in Facebook was so heavy that it took NASDAQ's computers up to five milliseconds to calculate its opening trade, about 100 times longer than usual. While this extended calculation was running, NASDAQ's order system allowed investors to change their orders up to the print of the opening trade on the tape. But these few extra milliseconds before the print were more than enough for new orders and cancellations to enter NASDAQ's auction book. These new changes caused NASDAQ's initial public offering software to recalculate the opening trade, during which time even more orders and cancellations entered its book, compounding the problem in an endless circle (Schapiro 2012). As the delay continued, more traders cancelled their previous orders, "in between the raindrops," as NASDAQ's CEO Robert Greifeld rather poetically explained. This glitch created something software engineers call a "race condition," in this case a race between new orders and the print of the opening trade, an infinite loop that required manual intervention to exit, something that hundreds of hours of testing had missed.

Though the initial public offering was scheduled to begin at 11:00 am that morning, delays caused trade opening to occur a half an hour late. As of 10:50 am, traders had not yet received acknowledgements of pre-opening order cancellations or modifications. Even after NASDAQ formally opened the market, many traders still had not received these critical acknowledgements, which created more uncertainty and anxiety (Strasburg, Ackerman, and Lucchetti 2012). By the time the system was reset, NASDAQ's programs were running 19 minutes behind real time. Seventy-five million shares changed hands during Facebook's opening auction, a staggering number, but orders totaling an additional 30 million shares took place during this 19-minute limbo. Problems persisted for hours after opening; many customer orders from both institutional and retail buyers went unfilled for hours or were never filled at all, while other customers ended up buying more shares than they had intended (Strasburg and Bunge 2012; McLaughlin 2012). This incredible gaffe, which some estimates say cost traders $100 million, eclipsed NASDAQ's achievement in getting Facebook's initial public offering, the third largest IPO in US history.

Less than two months before, another initial public offering suffered an even more shocking fate. BATS Global Markets, founded in 2005 as a "Better Alternative Trading System" to NASDAQ and the NYSE, held its initial public offering on March 23, 2012. BATS operates the third-largest stock exchange in the United States; its two electronic markets account for 11–12 percent of all US equity trading volume each day. BATS was among the most technologically advanced firms in its peer group and the envy of the industry. Quite naturally, BATS decided to list its initial public offering on its own exchange. If an organization ever had sufficient "skin in the game" to get it right, it was BATS, and if there were ever a time when getting it right really mattered, it was on March 23, 2012. So when BATS launched its own initial public offering at an opening price of $15.25, no one expected its price to plunge to less than a tenth of a penny in a second and a half due to a software bug affecting stocks with ticker symbols from A to BFZZZ, creating an infinite loop that made these symbols inaccessible on the BATS system (Oran, Spicer, Mikolajczak, and Mollenkamp 2012; Schapiro 2012). The ensuing confusion was so great that BATS suspended trading in its own stock, and ultimately cancelled its initial public offering altogether.

As isolated incidents, both the Facebook glitch and the BATS fiasco can be explained as regrettable software errors that extensive testing failed to catch, despite the best efforts of engineers. But two similar incidents in the space of two months suggest that the problem is more general than a few isolated computer errors. More worrisome is the fact that these glitches are affecting parts of the industry that previously had little to do with technology. After all, initial public offerings have been a staple of modern capitalism since the launch of the Dutch East India Company in 1602. But apparently, launching an initial public offering in a world with microsecond algorithmic trading has become an extremely challenging technical enterprise.

August 2012: Trading Errors at the Speed of Light

On August 1, 2012, a broker-dealer in securities, Knight Capital Group, Inc. experienced what it later called "a technology issue at the open of trading at the NYSE related to a software installation that resulted in Knight sending erroneous orders into the market." These orders and the unintended trades resulted in a rapid accumulation of positions "unrestricted by volume caps" and, between 9:30 am and 10:00 am eastern time, created significant swings in the share prices of almost 150 stocks (McCrank 2012; see also Telegraph 2012; Schapiro 2012). Unable to void most of these trades by classifying them as "erroneous," Knight Capital had no choice but to liquidate them in the open market. This liquidation resulted in a $457.6 million loss for the company, effectively wiping out its capital, causing its stock to lose 70 percent of its value, and forcing it to seek rescuers. After a few nerve-racking days, Knight Capital announced that it had "secured $400 million in financing," allowing it to survive. However, the stock of Knight Capital never really recovered, and in December 2012, the company was acquired by GETCO.

Just 42 days prior to the incident, Knight's chairman and chief executive officer, Mr. Thomas M. Joyce, while testifying before the US House of Representatives Committee on Financial Services, strongly argued in favor of a practice known as *internalization*, in which broker-dealers like Knight are permitted to post prices that are fractions of a penny better than prevailing quotes which are denominated in increments of a penny. For example, if the best bid and offer prices on an organized exchange are $100.01 and $100.02, respectively, internalization would allow Knight to post a bid at $100.011 or an offer at $100.019. Retail brokers can then legally send a retail customer's order (like "buy 500 shares") to Knight rather than to an organized exchange because most markets offer participants "price priority," which means that a buyer can step to the front of the order queue if that buyer is willing to pay a higher price than all other market participants, including the designated market maker. Sometime during the course of the day, often within seconds, the internalizer would find the inventory it owes to the customer by buying 500 shares of the stock at a lower price, say $100.001, from another retail customer or at another trading venue such as a dark pools, another internalizer or an organized exchange. It would then pocket the 1 penny difference between the two prices. Internalizers must use their own capital to fill customers' orders and, due to the Securities and Exchange Commission rule that came out in December 2011 in the wake of the Flash Crash, must have prudent risk management safeguards in place.

The losers from internalization are the organized exchanges that lose order flow and its associated fees to the internalizers. In October 2011, exchanges operated by the NYSE Euronext filed with the Securities and Exchange Commission proposed a rule to establish a "Retail Liquidity Program," a way to attract retail order flow to the New York Stock Exchange by allowing them to execute retail orders at sub-penny prices. Several broker-dealers, including Knight Capital, sent comment letters to the SEC arguing against the Retail Liquidity Program. However, after a prolonged comment period, the SEC concluded that "[t]he vast majority of marketable retail orders are internalized by [over-the-counter] market makers, who typically pay retail brokers for their order flow," while "[e]xchanges and exchange member firms that submit orders and quotations to exchanges cannot compete for marketable retail order flow on the same basis" (SEC 2013). Consequently, on July 3, 2012, the SEC approved the introduction of the Retail Liquidity Program to "promote competition between exchanges and [over-the-counter] market makers." On July 5, 2012, the NYSE Euronext issued a press release stating that the Retail Liquidity Program would be offered on some of its exchanges for one year on a pilot basis starting on August 1, 2012.

On August 2, 2012, in an interview on Bloomberg TV, Knight's CEO Joyce stated: "We put in a new bit of software the night before because we were getting ready to trade the NYSEs Retail Liquidity Program. This has nothing to do with the stock exchange. It had to do with our readiness to trade it. Unfortunately, the software had a fairly major bug in it. It sent into the market a ton of orders, all erroneous, so we ended up with a large error position which we had to sort through the balance of the day. It was a software bug, except it happened to be a very large software bug, as soon as we realized what we had we got it out of the code and it is gone now. The code has been restored. We feel very confident in the current operating environment we've reestablished."

The fall of Knight that began on August 1, 2012, and ended with its firesale acquisition less than six months later was more than just a technological glitch—it was a consequence of the technological arms race that pitted electronic trading platforms against automated broker-dealers in the competition for valuable customer order flow.

September 2012: High-Frequency Manipulation

On September 25, 2012, the Securities and Exchange Commission (2012) issued a cease-and-desist order against Hold Brothers On-Line Investment Services, an electronic broker-dealer who had been involved in manipulative trading activities through offshore high-frequency trading accounts. According to the SEC, from January 2009 to September 2010, these offshore entities engaged in "spoofing" and "layering," high-tech versions of well-known techniques for manipulating prices and cheating investors. "Spoofing" involves intentionally manipulating prices by placing an order to buy or sell a security and then canceling it shortly thereafter, at which point the spoofer consummates a trade in the opposite direction of the canceled order. "Layering" involves placing a sequence of limit orders at successively increasing or decreasing prices to give the appearance of a change in demand and artificially increase or decrease the price that unsuspecting investors are willing to pay; after a trade is consummated at the manipulated price, the layered limit orders are canceled.

The difference between these scams and the more traditional "pump-and-dump" schemes is the speed and electronic means with which they are conducted. For example, the cease-and-desist order from the Securities and Exchange Commission contains the following illustration of the kind of manipulation that went on for nearly two years (SEC 2012, paragraph 25):

> That day, at 11:08:55.152 a.m., the trader placed an order to sell 1,000 GWW shares at $101.34 per share. Prior to the trader placing the order, the inside bid was $101.27 and the inside ask was $101.37. The trader's sell order moved the inside ask to $101.34. From 11:08:55.164 a.m. to 11:08:55.323 a.m., the trader placed eleven orders offering to buy a total of 2,600 GWW shares at successively increasing prices from $101.29 to $101.33. During this time, the inside bid rose from $101.27 to $101.33, and the trader sold all 1,000 shares she offered to sell for $101.34 per share, completing the execution at 11:08:55.333. At 11:08:55.932, less than a second after the trader placed the initial buy order, the trader cancelled all open buy orders. At 11:08:55.991, once the trader had cancelled all of her open buy orders, the inside bid reverted to $101.27 and the inside ask reverted to $101.37.

The most notable fact about this narrative is that all of the manipulative activity took place within 839 milliseconds between 11:08:55 and 11:08:56. It is a physical impossibility for any human trader to have accomplished this manually.

In this case, the guilty parties were caught and fined more than $5.9 million by the Securities and Exchange Commission, the stock exchanges, and the Financial Industry Regulatory Authority, and permanently barred from the securities industry. However, their behavior is unlikely to be an isolated incident, which highlights the challenges facing regulators who need to revamp their surveillance and enforcement practices to be effective in catching the cyber-fraudsters of today.

Financial Regulation 2.0

Although the benefits of automation in financial markets are indisputable, they must be evaluated with two considerations in mind: complexity and human behavior. The software and hardware that control financial markets have become so complex that no individual or group of individuals is capable of conceptualizing all possible interactions that could occur among various components of the financial system. This complexity has created a new class of finance professionals known as "power users," who are highly trained experts with domain-specific technical knowledge of algorithmic trading. But because technological advances have come so quickly, there are not enough power users to go around. Moreover, the advantages that such expertise confers have raised concerns among those who do not have access to such technology that they are being unfairly and systematically exploited. And the growing interconnectedness of financial markets and institutions has created a new form of accident: a systemic event, where the "system" now extends beyond any single organization or market and affects a great number of innocent bystanders. The cautionary tales from the previous section are potent illustrations of this new financial order and provide considerable motivation for the global policy debate on the proper market structure in an automated world.

At the heart of this debate is the question of how "continuous" automated financial markets should be and the costs and benefits to the various stakeholders of transacting at faster and faster speeds. Grossman and Miller (1988) offer a stylized equilibrium framework in which the differences in possible market structures boil down to a tradeoff between 1) the costs to different types of intermediaries for maintaining a continuous presence in a market and 2) the benefits to different types of market participants for being able to execute trades as "immediately" as possible.

Automation of the trading process, including computerized algorithmic trading, has drastically reduced the costs to the intermediaries of maintaining a continuous market presence. In fact, intermediaries with the most efficient trading technology and the lowest regulatory burden realized the largest cost savings. As a result, the supply of immediacy has skyrocketed. At the same time, the frequency of technological malfunctions, price volatility spikes, and spectacular frauds and failures of intermediaries has also increased, while the net benefits of immediacy have accrued disproportionally to those who can better absorb the fixed and marginal costs of participating in automated markets. This has frustrated and disenfranchised a large population of smaller, less technologically advanced market participants who are concerned that regulators are unable to fulfill their mandate to protect investor interests, maintain fair and orderly markets, and promote capital formation.

These concerns have been met with a wide range of proposed policy and regulatory responses: do nothing; impose an outright ban on algorithmic—or at least high-frequency—trading; change the rules regarding who can be a designated intermediary and what responsibilities this designation entails; force all trading on exchanges to occur at fixed discrete intervals of time; or, instead of tinkering with "market plumbing," just introduce a "Tobin tax" on all financial transactions. Each of these proposals contains some merit from the standpoint of at least one set of stakeholders. However, all of the proposals pose difficult tradeoffs.

Doing nothing would allow intermediaries to find more ways to reduce the costs of being continuously present in the market, leading to an even greater supply of immediacy and more efficient trading, but is unlikely to address investors' concerns about fair and orderly markets.

Banning high-frequency trading might yield more fair and orderly markets in the short run—though the usage of "fair" in this context is somewhat strained given that a segment of market participants is being eliminated by fiat—but may also reduce market liquidity, efficiency, and capital formation as automated trading platforms have become increasingly dependent on high-frequency traders.

Changing the definition and requirements of a designated market maker to include high-frequency traders may also lead to more fair and orderly markets since such designations will prevent them from withdrawing from the market when their services are needed most. However, such redesignation would also increase the cost to intermediaries of being present in the market due to higher capital requirements, additional compliance costs for each designated market, and greater legal costs by virtue of being a regulated entity. In the short term, this would reduce the supply of immediacy because some traders may find these costs too high to continue making markets.

Forcing all trades to occur at discrete time intervals would concentrate the supply of immediacy, not unlike the periodic batch auctions of many European stock exchanges in the 1990s. How much immediacy would be demanded by different types of market participants, how much they would be willing to pay for it, and how the costs and benefits of concentrated immediacy would be shared among them are questions that must be answered before the welfare effects of this proposal can be evaluated. However, one indication of consumer preferences is the fact that most batch-auction markets have converted to continuous market-making platforms.

Finally, the Tobin tax—a small transaction tax on all financial transactions—has become a mainstay in the public debate on financial markets. In its most recent reincarnation, a variant of the Tobin tax is set to be implemented on January 1, 2014, by 11 members of the European Union including France, Germany, Italy, and Spain (Mehta 2013). However, 15 other members, including the United Kingdom, are strongly opposed to this measure. While this tax will certainly reduce trading activity across the board, and eliminate high-frequency trading altogether in those tax jurisdictions, it will also reduce market liquidity and impair hedging activity. For example, institutional investors often rely on derivative securities such as options and swaps to hedge risk exposures to fluctuations in stock prices, interest rates, and foreign exchange rates. Intermediaries are willing to take the other side of these transactions only if they can mitigate their own risk exposures by dynamically hedging their positions in the underlying stock, bond, and foreign currency markets. Even a small transactions tax would make such dynamic hedging activity impractical (Heaton and Lo 1995). Moreover, a successful implementation of such a tax requires international coordination, otherwise trading activity and human capital will simply migrate to venues without the tax, as it did in the case of Sweden from 1984 to 1990 (Umlauf 1993; Wrobel 1996).

In fact, all of these proposals are addressing only the symptoms of a much deeper problem: the fact that our financial regulatory framework has become antiquated and obsolete in the face of rapid technological advances that drastically reduced costs to intermediation, but have not correspondingly increased or distributed the benefits of greater immediacy. Minimizing technical and operating errors at the level of individual trading algorithms or automated systems—which should always be encouraged—is not sufficient to minimize the incidence of disruptive market-wide events. In fact, in a competitive environment, "optimal" decisions made by subsystems (for example, at the level of individual trading algorithms or trading firms) may interact with each other in ways that make the entire financial system more prone to systemic disruptions. Therefore, Financial Regulation 2.0 necessarily involves a *systemwide* redesign and ongoing systemwide supervision and regulation.

To bring the current financial regulatory framework into the Digital Age, we propose four basic design principles that we refer to as "Financial Regulation 2.0."

1) Systems-Engineered. Since most financial regulations will eventually be translated into computer code and executed by automated systems, financial regulation should approach automated markets as complex systems composed of multiple software applications, hardware devices, and human personnel, and promote best practices in systems design and complexity management. A number of these practices come from the field of systems engineering and have already been adopted in other industries such as transportation, manufacturing, and nuclear power.

2) Safeguards-Heavy. Financial regulation should recognize that automation and increasingly higher transaction speeds make it nearly impossible for humans to provide effective layers of risk management and nuanced judgment in a live trading environment. Thus, effective risk safeguards need to be consistent with the machine-readable communication protocols, as well as human oversight. Regulators need to encourage safeguards at multiple levels of the system.

3) Transparency-Rich. Financial regulation should aim to make the design and operation of financial products and services more transparent and accessible to automated audits conducted on an ongoing basis by the regulator's own "bots." Ideally, regulation should mandate that versions and modifications of the source code that implement each rule, as well as the data used for testing and validation of the code, are made available to the regulators and potentially the public. Regulators need to change their surveillance and enforcement practices to be more cyber-centric rather than human-centric.

4) Platform-Neutral. Financial regulation should be designed to encourage innovation in technology and finance, and should be neutral with respect to the specifics of how core computing technologies like operating systems, databases, user interfaces, hardware solutions, and software applications work. Doing otherwise would inevitably lock-in outdated practices, ring-fence potentially inefficient ways of doing business, and empower incumbents at the expense of potential new entrants.

Although these principles may seem unrealistic, a recent example of a regulatory initiative consistent with these principles is the set of measures surrounding the creation of "legal entity identifiers"—alphanumeric, machine-readable strings uniquely associated with each separate entity participating in a financial transaction (for example, see the legal-entity-identifier-related publications of the Financial Stability Board at http://www.financialstabilityboard.org/list/fsb_publications/tid_156/index.htm). This initiative is cyber-centric, promotes innovation, imposes system-design principles, increases transparency, enables the creation of additional risk safeguards, and encourages the implementation of risk management processes and workflows that allow human knowledge to complement the computational abilities of machines. This gives us hope that with sufficient motivation, effort, and expertise, Financial Regulation 2.0 will be achievable.

■ *Research support from the MIT Laboratory for Financial Engineering is gratefully acknowledged. We thank David Autor, Matthew Baron, Jayna Cummings, Chang-Tai Hsieh, Ulrike Malmendier, Ann Norman, Mehrdad Samadi, and Timothy Taylor for helpful comments and suggestions. The views and opinions expressed in this article are those of the authors only, and do not necessarily represent the views and opinions of any institution or agency, any of their affiliates or employees, or any of the individuals acknowledged above.*

References

Baron, Matthew, Jonathan Brogaard, and Andrei Kirilenko. 2012. "The Trading Profits of High Frequency Traders." http://conference.nber.org/confer/2012/MMf12/Baron_Brogaard_Kirilenko.pdf.

Bertsimas, Dimitris, and Andrew Lo. 1998. "Optimal Control of Execution Costs." *Journal of Financial Markets* 1(1): 1–50.

Black, Fischer, and Myron Scholes. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81(3): 637–54.

Bogle, John C. 1997. "The First Index Mutual Fund: A History of Vanguard Index Trust and the Vanguard Index Strategy." http://www.vanguard.com/bogle_site/lib/sp19970401.html.

Commodity Futures Trading Commission/Securities and Exchange Commission (CFTC/SEC). 2010. *Preliminary Findings Regarding the Market Events of May 6, 2010.* Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues. May 18. http://www.sec.gov/sec-cftc-prelimreport.pdf.

Denning, Peter J. 2003. "Great Principles of Computing." *Communications of the ACM* 46(11): 15–20.

Goldman Sachs Asset Management. 2007. "The Quant Liquidity Crunch." Goldman Sachs Global Quantitative Equity Group, August. Proprietary document for Goldman Sachs clients; not available to the general public.

Gorton, Gary, and Andrew Metrick. 2012. "Securitized Banking and the Run on Repo." *Journal of Financial Economics* 104(3): 425–51.

Grossman, Sanford J., and Merton H. Miller. 1988. "Liquidity and Market Structure." *Journal of Finance* 43(3): 617–37.

Heaton, John, and Andrew W. Lo. 1995. "Securities Transaction Taxes: What Would Be Their Effects on Financial Markets and Institutions?" In *Securities Transaction Taxes: False Hopes and Unintended Consequences,* edited by Suzanne Hammond, 58–109. Chicago, IL: Catalyst Institute.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld. 2011. "Does Algorithmic Trading Improve Liquidity?" *Journal of Finance* 66(1): 1–33.

Khandani, Amir E., and Andrew W. Lo. 2007. "What Happened to the Quants in August 2007?" *Journal of Investment Management* 5(4): 5–54.

Khandani, Amir E., and Andrew W. Lo. 2011. "What Happened to the Quants in August 2007? Evidence from Factors and Transactions Data." *Journal of Financial Markets* 14(1): 1–46.

Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2011. "The Flash Crash: The Impact of High Frequency Trading on an Electronic Market." http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1686004.

Lehmann, Bruce N. 1990. "Fads, Martingales, and Market Efficiency." *Quarterly Journal of Economics* 105(1): 1–28.

Lintner, John. 1965. "The Valuation of Risky Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Review of Economics and Statistics* 47(1): 13–37.

Lo, Andrew, and Craig MacKinlay. 1990. "When Are Contrarian Profits Due to Stock Market Overreaction?" *Review of Financial Studies* 3(2): 175–205.

Markowitz, Harry. 1952. "Portfolio Selection." *Journal of Finance* 7(1): 77–91.

McCrank, John. 2012. "Knight Glitch Likely to Lead to Regulatory Changes: CEO." Reuters, September 11. http://www.reuters.com/article/2012/09/11/us-knight-ceo-idUSBRE88A1GW20120911.

McLaughlin, Tim. 2012. "Facebook IPO

Glitch Prompts Margin Calls, Headaches." Reuters, May 25. http://www.reuters.com/article/2012/05/25/us-facebook-smallinvestors-idUSBRE84O18P20120525.

Mehta, Nitin. 2013. "Pros and Cons of 'Tobin Tax' Divides EU." *Financial Times*, March 3.

Merton, Robert C. 1973. "Theory of Rational Option Pricing." *Bell Journal of Economics and Management Science* 4(1): 141–83.

Mossin, Jan. 1966. "Equilibrium in a Capital Asset Market." *Econometrica* 34(4): 768–83.

Oran, Olivia, Jonathan Spicer, Chuck Mikolajczak, and Carrick Mollenkamp. 2012. "BATS Exchange Withdraws IPO after Stumbles." Reuters, March 24. http://uk.reuters.com/article/2012/03/24/us-bats-trading-idUKBRE82M0W020120324.

Rosenberg, Barr. 1974. "Extra-Market Components of Covariance in Security Returns." *Journal of Financial and Quantitative Analysis* 9(2): 263–74.

Rothman, Mathew S. 2007a. "Turbulent Times in Quant Land." *U.S. Equity Quantitative Strategies*, August 9. Lehman Brothers Equity Research. http://dealbreaker.com/_old/images/pdf/quant.pdf.

Rothman, Mathew S. 2007b. "View from QuantLand: Where Do We Go Now?" U.S. Equity Quantitative Strategies, Lehman Brothers Research. Proprietary document for Lehman clients only; not available to the general public.

Rothman, Matthew S. 2007c. "Rebalance of Large Cap Quant Portfolio."' U.S. Equity Quantitative Strategies, Lehman Brothers Research. Proprietary document for Lehman clients only; not available to the general public.

Schapiro, Mary. 2012. "Introductory Remarks at SEC's Market Technology Roundtable." October, 2. http://www.sec.gov/news/speech/2012/spch100212mls.htm.

Securities and Exchange Commission (SEC). 2012. "Order Instituting Administrative and Cease-and-Desist Proceedings Pursuant to Sections 15(B) and 21C of the Securities Exchange Act of 1934 and Section 9(B) of the Investment Company Act of 1940, Making Findings, and Imposing Remedial Sanctions and Cease-and-Desist Orders." Administrative Proceeding, File No. 3-15046. September 25.

Securities and Exchange Commission (SEC). 2013. "Order Granting Approval to Proposed Rule Change, as Modified by Amendment No. 1, to Establish the Retail Price Improvement Program on a Pilot Basis until 12 Months from the Date of Implementation." Release No. 34-68937; File No. SR-NASDAQ-2012-129, February 15.

Sender, Henny, Kate Kelly, and Gregory Zuckerman. 2007. "Goldman Wagers on Cash Infusion to Show Resolve." *Wall Street Journal* (Eastern edition). August 14, p. A.1.

Sharpe, William F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19(3): 425–42.

Strasburg, Jenny, Andrew Ackerman, and Aaron Lucchetti. 2012. "Nasdaq CEO Lost Touch Amid Facebook Chaos." *Wall Street Journal* (Eastern Edition). June 11, p. A.1.

Strasburg, Jenny, and Jacob Bunge. 2012. "Social Network's Debut on Nasdaq Disrupted by Technical Glitches, Trader Confusion." *Wall Street Journal* (Eastern Edition). May 19, p. A.2.

Tabb, Larry. 2012. "Written Testimony to the United States Senate Committee on Banking, Housing, and Urban Affairs by Larry Tabb, CEO, TABB Group." September 20. http://www.banking.senate.gov/public/index.cfm?FuseAction=Hearings.Testimony&Hearing_ID=f8a5cef9-291d-4dd3-ad3b-10b55c86d23e&Witness_ID=f520faa2-1cfe-48a5-b373-60bde009d3a3.

Telegraph, The. 2012. "Knight Capital's $440m Trading Loss 'Caused by Disused Software.'" August 14. http://www.telegraph.co.uk/finance/newsbysector/banksandfinance/9475292/Knight-Capitals-440m-trading-loss-caused-by-disused-software.html.

Umlauf, Steven R. 1993. "Transaction Taxes and the Behavior of the Swedish Stock Market." *Journal of Financial Economics* 33(2): 227–40.

Wrobel, Marion G. 1996. "Financial Transactions Taxes: The International Experience and the Lessons for Canada." Background Paper BP-419E, Research Branch, Library of Parliament, Government of Canada.

Zuckerman, Gregory, James Hagerty, and David Gauthier-Villars. 2007. "Impact of Mortgage Crisis Spreads, Dow Tumbles 2.8% as Fallout Intensifies; Moves by Central Banks." *Wall Street Journal* (Eastern Edition). August 10, p. A.1.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 73–96

An International Look at the Growth of Modern Finance[†]

Thomas Philippon and Ariell Reshef

Studies of long-run evolution of the finance industry have largely focused on the United States. These studies reveal three key facts: 1) the share of aggregate income spent on financial intermediation is time varying; 2) the unit cost of financial intermediation is relatively flat; and 3) the pattern of changes in human capital and wages in finance relative to the whole economy exhibits a U-shape over the twentieth century. In this paper, we ask whether these facts hold for a set of other economies with similar levels of development.

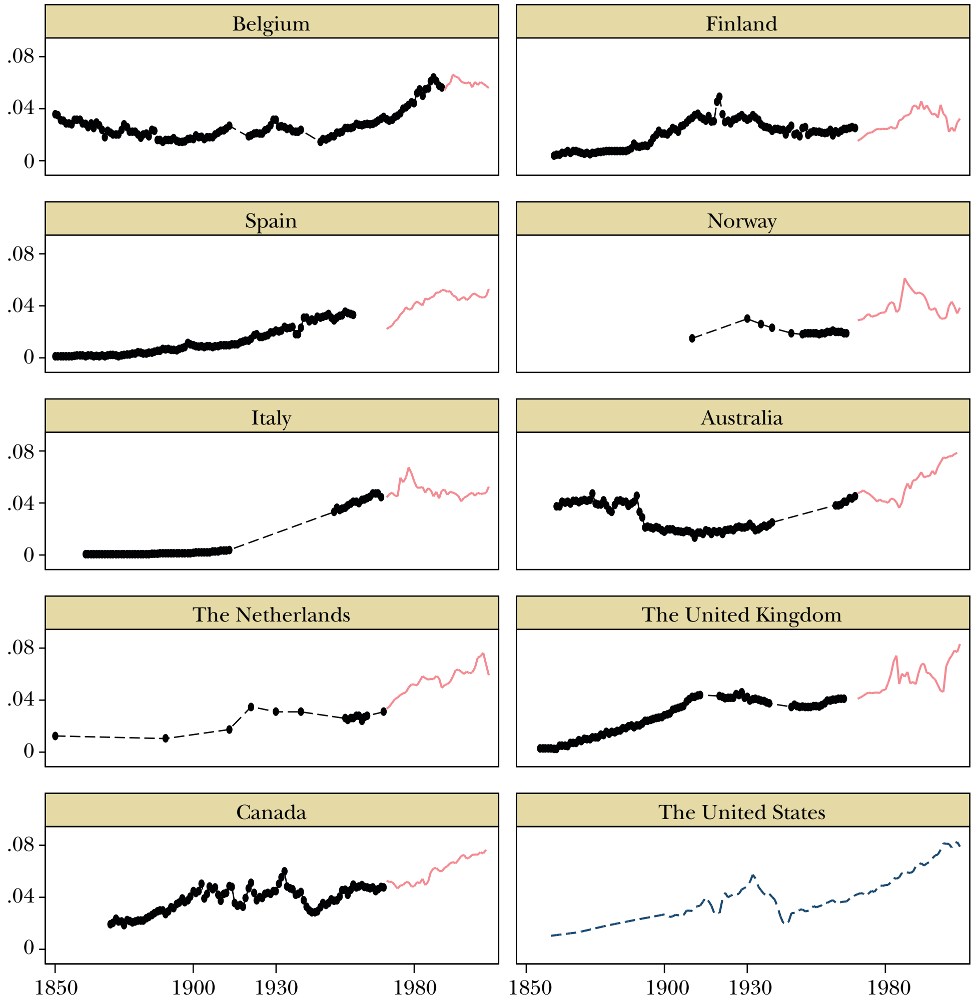
Over the long run, the US financial sector has grown in two waves: The first lasted from (at least) 1860 to the 1930s; and then, following a sharp decline, the second wave starts in 1950 and lasts to the present. The long-run trend of the income share of finance in the United States is similar to that in a number of other now-industrial economies, although—as Figure 1 illustrates—the exact pattern varies by country. A few features in Figure 1 stand out. First, in all of these countries—except Finland, for a brief period—finance's share of income today is significantly higher than it has been during the last 150 years. Second, the overall trend is upward, although periods of decline are evident; in particular, there are sharp drops in Australia after 1888 and in Canada and the United States after 1933 following severe depressions. Third, while the Netherlands, the United

■ *Thomas Philippon is Associate Professor of Finance, Stern School of Business, New York University, New York, New York. He is also a Faculty Research Fellow, National Bureau of Economic Research, Cambridge, Massachusetts, and a Research Affiliate, Centre for Economic Policy Research, London, United Kingdom. Ariell Reshef is Assistant Professor of Economics, University of Virginia, Charlottesville, Virginia. Their email addresses are tphilipp@stern.nyu.edu and ariellr@virginia.edu.*

[†] To access the Appendix, visit
http://dx.doi.org/10.1257/jep.27.2.73.

doi=10.1257/jep.27.2.73

Figure 1
Historical Income Share of the Financial Sector, 1850–2007



Sources: The historic series is mostly from Smits, Woltjer, and Ma (2009) and from various historical statistical sources: Australia in 1861–1939 from Vamplew (1987); Canada in 1870–1926 from Urquhart (1993) and in 1926–1976 from Statistics Canada; Italy in 1958–1968 from Istituto Centrale Di Statistica (various years); The Netherlands in 1921–1969 from Office Statistique des Communautes Europeennes (1966) and den Bakker and de Gijt (1990); Norway in 1910–1960 from the Central Bureau of Statistics of Norway, *Historical Statistics 1968* (1969). Modern data are either from STAN (OECD) or EU KLEMS. Discrepancies between STAN and EU KLEMS data are insignificant. EU KLEMS data are described in O'Mahony and Timmer (2009). The raw historic value added in finance and GDP series for the UK are volume indices; to get the value added share in the UK we assume that the unit cost of financial services divided by the unit cost of GDP (the GDP deflator) is constant from 1970 going backwards. See the online Appendix for complete details.
Notes: Black dots represent historical sources, solid lines represent modern sources. The dashed line for the USA series is from Philippon (2012); this series combines several sources. The historic and modern income share series are the value added of financial intermediation (without real estate) as a share of GDP.

Kingdom, and Canada share the long-run pattern of the rise of finance with the United States, where finance continues to increase after 1980 (and Australia more recently), it seems that in other economies the financial sectors' income share reaches a plateau, and even declines somewhat. Notice also the similarities in the series for Canada and the United States, for the Netherlands and United Kingdom, and for Finland and Norway; these pairs have historically integrated financial sectors. Finally, it is important to understand that these patterns are not explained by the general increase in the income share of services or the decline of agriculture: Figure 1 is qualitatively unchanged when we compute the share of finance in services alone.

What forces can explain the historical growth of the income share of the finance industry as documented in Figure 1? Simple neoclassical models are not likely to provide adequate answers. Explanations that are based on two-sector models with productivity growth differentials—in which there is either low elasticity of substitution in demand and slower productivity growth in finance *à la* Baumol (1967), or elastic demand and faster productivity growth in finance—are also not satisfactory. Philippon (2012) finds that the unit cost of finance relative to other output in the United States is flat (with a slightly higher level from the 1980s and on); this in itself rules out both of the above mechanisms, as the income share of finance varies even when the unit cost does not change. Philippon (2012) argues that a benchmark model predicts a flat share of income for the finance industry, but that changes in industry structure (young firms, capital-intensive projects) or changes in demographics (inequality) should affect the income share of the finance industry.

Another common suggestion is that the growth of the financial sector is linked to globalization, but at a minimum, this relationship is not straightforward. If the relationship was monotone, then the end of the first era of globalization and the collapse of the gold standard in 1914 should have reduced the size of the financial sector. Instead, the growth of finance only slows down in some countries, while it accelerates in several others countries, including Belgium, the Netherlands, Canada, and the United States. The recovery in the size of finance from its mid-twentieth century low and the acceleration of its growth happen before globalization takes off in the 1990s for several countries. And although the Bretton Woods era (1945–71) seems to coincide with no growth in the income share of finance in some countries, in others—Belgium, the United States—it rises (for long-run trends in globalization see Obstfeld and Taylor 2004).

If richer individuals and households have a higher propensity to save, then they may demand more financial services. Thus, we may expect to find higher demand for financial services when inequality is higher. We find some support for this hypothesis in recent times, with significant increases in inequality in the United States, the United Kingdom, and Canada, commensurate with a growing income share for the financial sector after 1980. But inequality in the Netherlands does not increase, and Australia sees only moderate increases in inequality as do most other countries. Also the recent increases in inequality are typically dwarfed by long-run drops in inequality, while finance rises for all countries.[1]

Another hypothesis is that an increase in the degree of specialization can explain the observed patterns. According to this hypothesis, the finance industry performs more tasks that have been done by households (and thus were not previously measured in value added)—like managing savings for retirement—and takes the role of more traditional sources of finance—like shop credit. While such changes are plausibly part of the story, it is difficult to find data to help evaluate how important this force is. For more recent times, Greenwood and Scharfstein (this issue) document an increase in revenue from active management in the United States, but even this cannot explain the bulk of the increase in the US financial sector.

In what follows, we examine some additional aspects of the growth of finance in order to provide some facts with which any theory of this phenomenon should be consistent. We first examine the relationship between the size of the financial sector and income per capita. We find that the income share of the finance industry rises with income in early stages of development, but that relationship does not hold for medium levels of development. Moreover, not all countries in our sample exhibit rising finance shares in more advanced stages of development. We also discuss the relationship between the size of the financial sector and economic growth. We then turn to examine the income share of the finance industry since 1970 in more detail. We also consider skill intensity and wages in finance relative to the whole economy as another potential source of the rise in the income share of the finance industry. We find that demand for skill in finance increases with information and communication technology investments and with financial deregulation, but that wages in finance are only related to the former, not the latter. We then ask whether the cost per unit of financial services has risen in tandem with the income share of finance; we reject this hypothesis. We also discuss potential changes in the quality of financial services that are difficult to observe. In the conclusion, we draw together a number of insights from our discussion and highlight some new questions they raise.

The Size of the Financial Sector and Income

One potential explanation for the growth of finance is that there is greater relative demand for it as income rises (that is, preferences for financial services are nonhomothetic). For example, Buera and Kaboski (2012a) argue that such forces led to the rise of the service sector. As mentioned above, patterns in the growth of finance show it to be over and above the growth of services more broadly,

[1] The inequality data are taken from the World Top Incomes Database, constructed by Facundo Alvaredo, Tony Atkinson, Thomas Piketty, and Emmanuel Saez (website: http://topincomes.g-mond.parisschoolofeconomics.eu/) and from the University of Texas Inequality Project (website: http://utip.gov.utexas.edu/).

so explanations for the rise of the services sector are not sufficient to explain the growth of the financial sector.[2]

We examine the relationship between the income share of the finance industry and average income (real GDP per capita), using data from Maddison (2010). Since income (in logs) progresses with time more-or-less linearly, Figure 1 is also a good representation of the relationship of the income share of the finance industry to income per capita.[3] Almost all countries—Belgium and Australia being the notable exceptions—see the finance industry income share rise at early stages of development. After that, all countries except the United States exhibit a relatively flat share of finance. It is difficult to attribute the common flat part in the middle range of development to disruption due to the period from World War I through World War II because the timing is not consistent across countries and, moreover, incomes continues to rise. While the United States, the United Kingdom, Canada, and the Netherlands see an additional significant rise at higher levels of development, Finland, Spain, Norway, and Italy do not. The pattern for Belgium is different, but we see that at the very highest levels of development, the income share of the finance industry is flat there, too.

We examine the relationship between finance and income in another way, using a proxy for financial sector output. We use data on bank loans to nonfinancial entities: firms in the private sector, government, and households, from Schularick and Taylor (2012) for a sample of 14 now-industrial countries in 1870–2008. The sample of countries is: Australia, Canada, Switzerland, Germany, Denmark, Spain, France, Italy, Japan, the Netherlands, Norway, Sweden, the United Kingdom, and the United States. The proxy for financial output is given by the ratio of these bank loans to GDP. While this is a partial measure of financial output (many other forms of financial intermediation are neglected, as well as insurance), the data have the benefit of being a consistent historical time series. This series is relatively more informative in earlier periods, and for countries that have a relatively more bank-oriented financial system.

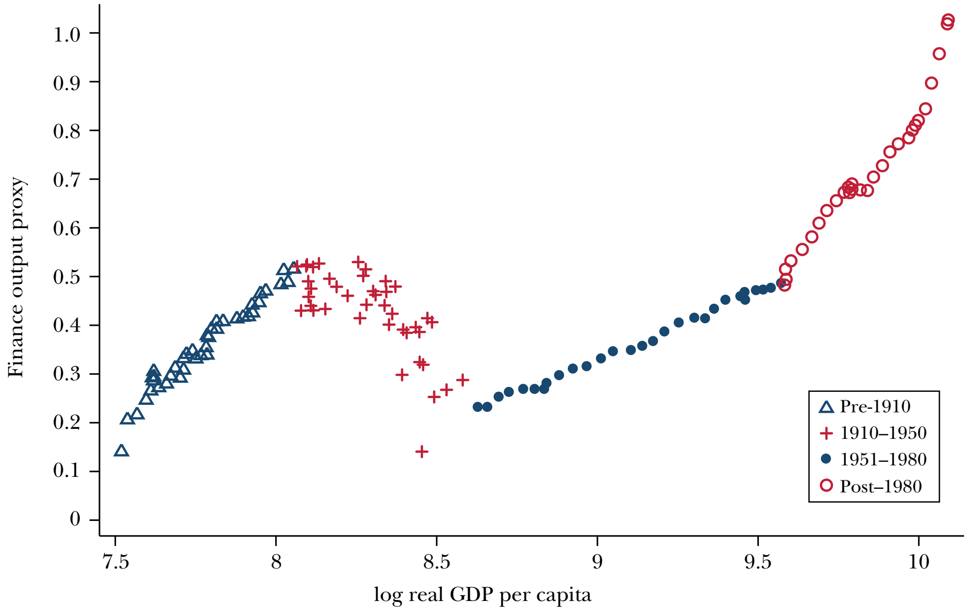
To obtain the average relationship between income and our proxy for financial output in the sample, over time, we fit fixed effects regressions of the type $y_{i,t} = c_i + d_t + \varepsilon_{i,t}$, where y is either log real GDP per capita or bank loans/GDP, c_i capture time-invariant country-specific factors, d_t capture common year-specific factors, and $\varepsilon_{i,t}$ is a projection error. Figure 2 plots the d_t from the regression where y is log real GDP per capita, against d_t from the regression where y is bank loans/GDP.

Four distinct periods are highlighted in Figure 2. Until 1910, financial output and income grow together. The tumultuous period of 1910–1950 exhibits a negative relationship: Income continues to grow, while finance contracts. In the postwar period, after 1950, financial output grows with income. But after 1980

[2] Buera and Kaboski (2012b) argue that scale economies can help explaining increasing sizes of industries and shifts in the composition of the economy. However, Philippon (2012) estimates that financial output is produced at constant returns to scale in the United States.

[3] For more detail, see Figure A2 in the online Appendix available with this paper at http://e-jep.org.

Figure 2
Finance Output and GDP Per Capita



Notes: The figure reports the relationship between the average finance output proxy and average real GDP per capita in a sample of 14 countries over 1870–2008. The finance output proxy is bank loans to nonfinancial entities (firms in the private sector, government, and households), from Schularick and Taylor (2012), divided by GDP. Real GDP per capita (in 1990 prices) is from Maddison (2010). The sample of countries is: Australia, Canada, Switzerland, Germany, Denmark, Spain, France, Italy, Japan, the Netherlands, Norway, Sweden, the United Kingdom, and the United States. Each observation is a year. We fit fixed effects regressions $y_{i,t} = c_i + d_t + \varepsilon_{i,t}$, where y is either log real GDP per capita or bank loans/GDP, c_i are country fixed effects and d_t are year fixed effects. The figure reports the relationship between the year fixed effects from the bank loans/GDP regression with the year fixed effects from the log real GDP per capita regression.

the relationship changes: The proportional change (elasticity) of financial output with respect to income is much higher after 1980 relative to 1951–1980. Alternatively put, relative to the period before 1980, the same proportional change in financial output is related to a smaller rise in income. Statistical analysis confirms that the change between post- and pre-1980 is not only economically large but also statistically significant.[4] Notice that in the later periods, as financial innovations expand the scope of financial intermediation, the proxy of financial output we are using here (bank loans/GDP) increasingly *understates* financial output, especially for countries like the United States, Canada, the United Kingdom, and the Netherlands. Securitization, and the removal of loans (mortgages) off banks'

[4] Restricting attention to the US economy delivers similar results. See Table A1 and Figure A3 in the online Appendix available with this paper at http://e-jep.org.

balance sheets reinforce this tendency to understate. It is therefore even more surprising to see that the financial output proxy, thus measured, increases even more rapidly in later periods relative to income.

Overall, we see that most of the rise in living standards after 1870 was obtained with less income spent on finance and less financial output than what is observed after 1980; and the relationship between financial output and income has changed after 1980.

It is also worthwhile noting that in this sample both the income share of finance and our proxy for financial output are not correlated with *growth* in GDP per capita; if anything, there is a small negative correlation after 1950. We do not suggest that finance is not important for growth; sustaining income growth over such a long period may very well be related to the fact that finance has been able to grow, or remain at substantial levels. Indeed, in broad cross sections of countries, finance is positively related to growth; see Rousseau and Sylla (2003) and Levine (2005). But in this sample, the secular rise of financial output does not seem to deliver *faster* growth. Several theories predict a positive relationship between expenditure on the financial sector's screening or monitoring services and growth—for example, Greenwood and Jovanovic (1990) and Greenwood, Sanchez, and Wang (2010), respectively—but this is not the case in this sample.[5]

Laeven, Levine, and Michalopoulos (2012) develop a theory in which the technology for screening new projects becomes less efficient for newer innovations (which are typically more complex and less easily understood); thus, growth ceases without financial innovation. In their model, the income share of finance is constant. But if newer screening technology becomes proportionately more costly to operate (not a feature of their model), then a constant growth rate may be consistent with a growing income share of the finance industry, at least for a while.

Recent Cross-Country Patterns of the Growth of Finance

Although many high-income countries have seen a rise of the financial sector over the long run, in recent times the experience of the US financial sector has been distinctive in a number of ways. In this section, we describe and discuss these differences using data from the European Union KLEMS dataset in 1970–2006; we restrict the sample to countries that report data on most variables of interest from the early 1970s. The sample of countries is: Austria, Belgium, Canada, Denmark, Finland, France, Germany, Japan, the Netherlands, Sweden, the United Kingdom, and the United States. The data were downloaded from http://www.euklems.net/; see O'Mahony and Timmer (2009) for a summary of the methodology and construction of this database.

[5] Other prominent papers relating finance to growth include Bencivenga and Smith (1991), Levine (1991), King and Levine (1993), Obstfeld (1994), and Aghion, Howitt, and Mayer-Foulkes (2005). These papers investigate different mechanisms by which the financial sector can enhance growth.

Figure 3 reports the income share of the finance industry, defined as above as value added in finance divided by total value added (that is, GDP). The countries in Panel A exhibit consistently increasing income shares of finance after 1970. These countries share the recent trend with the United States, and they all end the sample with a share greater than 6 percent of GDP. Overall, the US financial sector starts among the lowest in terms of income share and ends up among the highest. The increase of finance's income share in the United States is second only to that of the Netherlands.

We juxtapose the increasing trends in Panel A with those of the countries in Panel B, which exhibit relatively flat (Denmark) or mixed trends. Within this group there is considerable variation: for example, the income share of Belgium's financial sector increases by 3 percentage points and then declines slightly; France and Sweden see a sharp increase followed by a fall almost to initial levels, and Germany sees a weak increase. These financial sectors of Panel B countries all end the period with a share smaller than 6 percent of GDP. The different trends within this group, and relative to countries in Panel A, show that recently the growth of finance is not a uniform phenomenon.

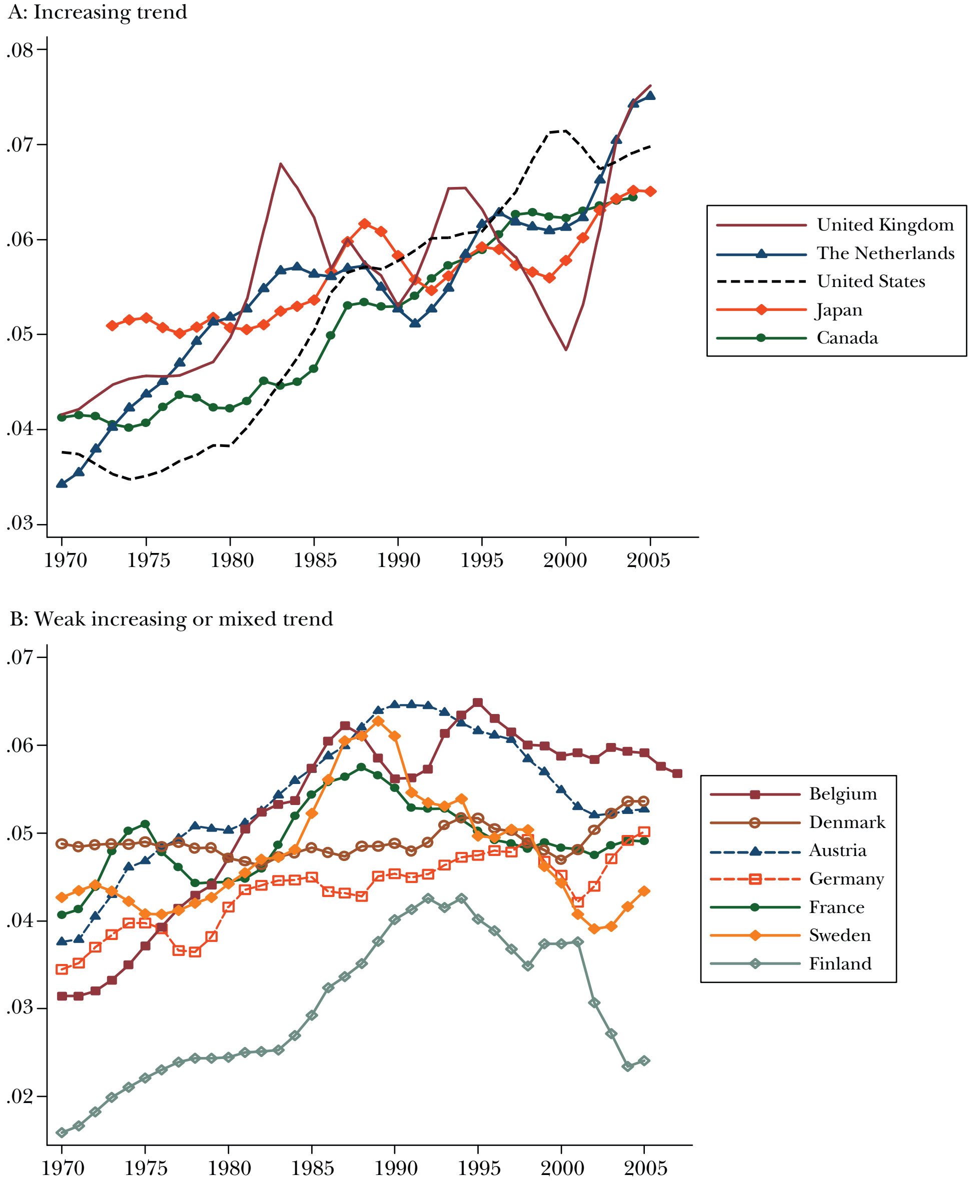
We next turn to describing wages in finance relative to the whole economy, that is, the finance industry relative wage. Average wages in finance are given by the ratio of labor compensation in finance to (full-time equivalent) employment in finance. The relative wage of finance is given by dividing average wages in finance by average wages in the whole economy, similarly computed. Labor compensation includes wages, salaries and supplements, employers contributions to social programs, tips, and—importantly for our purposes—bonuses and executive compensation. However, labor compensation does not include income from the exercise of stock options, or the share of proprietors' income that is accrued as compensation for labor services of owners of businesses. For example, this measure misses the income of hedge fund partners (but not that of their employees) that accrues to their labor services. Disentangling hedge fund partners' "labor income" from proprietors' capital income is not possible given the available sources.

Figure 4 reports the relative wage in the finance industry (the average wage in finance relative to the average wage in the economy as a whole). Panel A reports countries with an increasing relative wage in finance. We add France to this group, which exhibits a similar trend for relative wages in finance after an initial, sharp decline. It is noteworthy that the United States experiences one of the greatest increases in this sample, matched only by the Netherlands. But this trend for a higher relative wage in finance is not shared with all countries, as reported in Panel B. Other countries experience mixed trends in relative wages in finance, most notably the United Kingdom.

Skilled workers are paid more than unskilled workers, so we ask whether different patterns of skill intensities in finance relative to the whole economy—across countries and time—can explain the patterns in Figure 4. Skilled workers are defined consistently in the data as holding at least a college or university degree. We examine the relative skill intensity in finance, defined as the share of skilled workers

Figure 3

Value Added Shares of Finance in GDP



A: Increasing trend

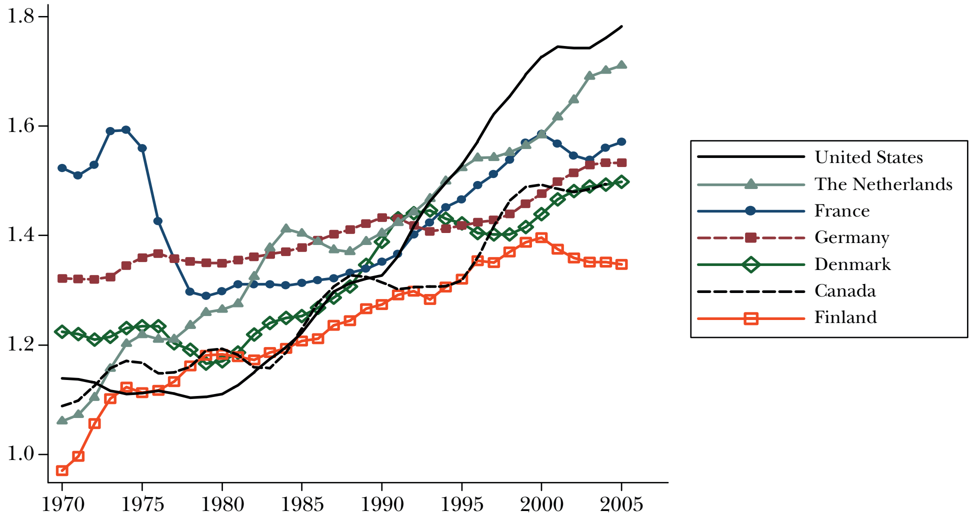
B: Weak increasing or mixed trend

Source: Authors' calculations using data from EU KLEMS.

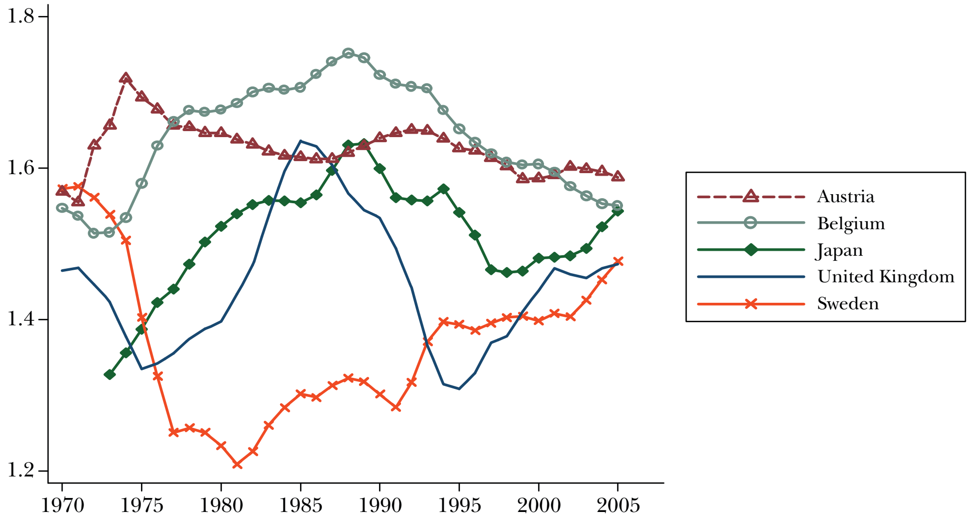
Notes: The figures report the share of finance in GDP. Series are three-year moving averages. Panel A groups countries that exhibit a strong increasing trend. Panel B groups countries that exhibit either a weak upward or mixed trend.

Figure 4
Relative Wage in Finance

A: Increasing trend

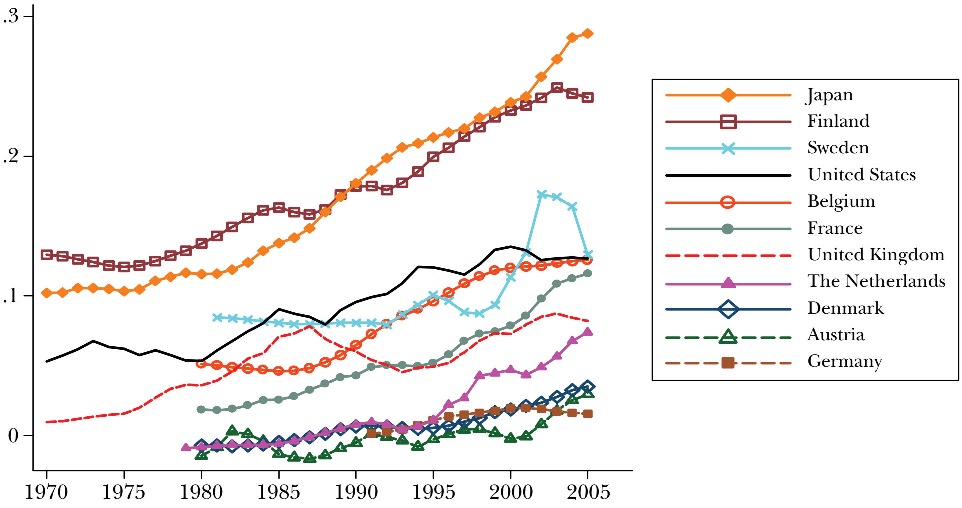


B: Mixed trend



Source: Authors' calculations using data from EU KLEMS. Series are three-year moving averages.
Notes: The figures report the average wage in finance relative to the average wage in the whole economy. Average wages are computed by dividing labor compensation by full-time equivalent employment. Panel A groups countries that exhibit an increasing trend (except for France in the beginning of the sample). Panel B groups countries that exhibit either a mixed or decreasing trend.

Figure 5
Relative Skill Intensity in Finance



Source: Authors' calculations using data from EU KLEMS.
Notes: Relative skill is defined as the share of high-skilled workers' (full-time equivalent) employment in finance minus the corresponding share in the whole economy. Skilled workers in all countries are comparable and attain at least a college or university degree. Data for Canada are not available from the EU KLEMS. Series are three-year moving averages.

in employment (measured in terms of full-time equivalent worker) in the financial sector minus the same share in the whole economy. Thus, an upward-sloping line shows that the employment share of skilled workers in finance is rising faster than the overall relative supply of skill.[6]

While the share of jobs held by skilled workers is rising across all economies in our sample (not shown), Figure 5 shows that finance becomes relatively more skill intensive compared to the overall supply of skilled labor in all countries. We also see wide variation in the relative skill intensity in finance, which points to country-specific factors. Within this variation, the United States tends to be higher than most countries—but Finland and Japan exhibit an even higher relative skill intensity in finance. The increase in skill intensity cannot explain finance wages in Figure 4 because relative skill intensity in finance is increasing for all countries in the sample while we see mixed patterns in Figure 4. While skill intensity in the US financial sector increases relative to the whole economy, it does not increase more than the average country. As we show in Philippon and Reshef (2012), faster growth in the cost of skilled labor (returns to skill), together with the increase in relative skill intensity in finance in the United States explains little of the growth of the relative wage in finance.

[6] We obtain a very similar figure when we use the relative wage bill share for skilled workers in finance as an alternative measure of skill intensity.

We also consider wages of skilled workers (defined as above) in finance relative to wages of skilled workers in the whole economy. Panel A of Figure 6 reports countries with consistently increasing relative skilled wages in finance. Panel B exhibits countries with mixed trends. Overall, we see increasing relative skilled wages in finance: skilled workers in finance gain over skilled workers elsewhere in all but two countries, Austria and Belgium, where skilled relative wages in finance are relatively high to begin with and then decline. Once again, the change for the US economy is the largest. Using several methodologies, in Philippon and Reshef (2012) we show that the increase in relative wages in finance is not primarily driven by compositional changes within the group of skilled workers. Given the similarities with Figure 4, differences in skilled relative wages in finance versus the whole economy can help explain at least part of the general rise in overall relative wages in finance. In the next section, we examine two determinants of the increase in relative wages and skill intensities in finance: technology and financial regulation.

Finance Wages and Demand for Skill

While high wages are now common in finance, this has not always been the case, as can be seen in Figure 4 and Figure 6. In Philippon and Reshef (2012), we document the historical pattern of finance wages relative to the nonfarm private sector over 1909–2006 for several types of workers and comparison groups. We find a U-shape over the sample period for average wages, skilled wages, and executive compensation in finance, using a variety of methods. These findings are in line with Goldin and Katz (2008), who document a large increase in the wage premium for Harvard undergraduates who choose a career in finance since 1970. Kaplan and Rauh (2010) and Bakija, Cole, and Heim (2012) study earnings of individuals with very high incomes, with a particular emphasis on the financial sector. Similarly, finance has become more skill intensive, as documented in Figure 5. Oyer (2008) argues that income differences attract MBAs to finance, rather than consulting or marketing. This change is reflected in the skill intensity of finance.

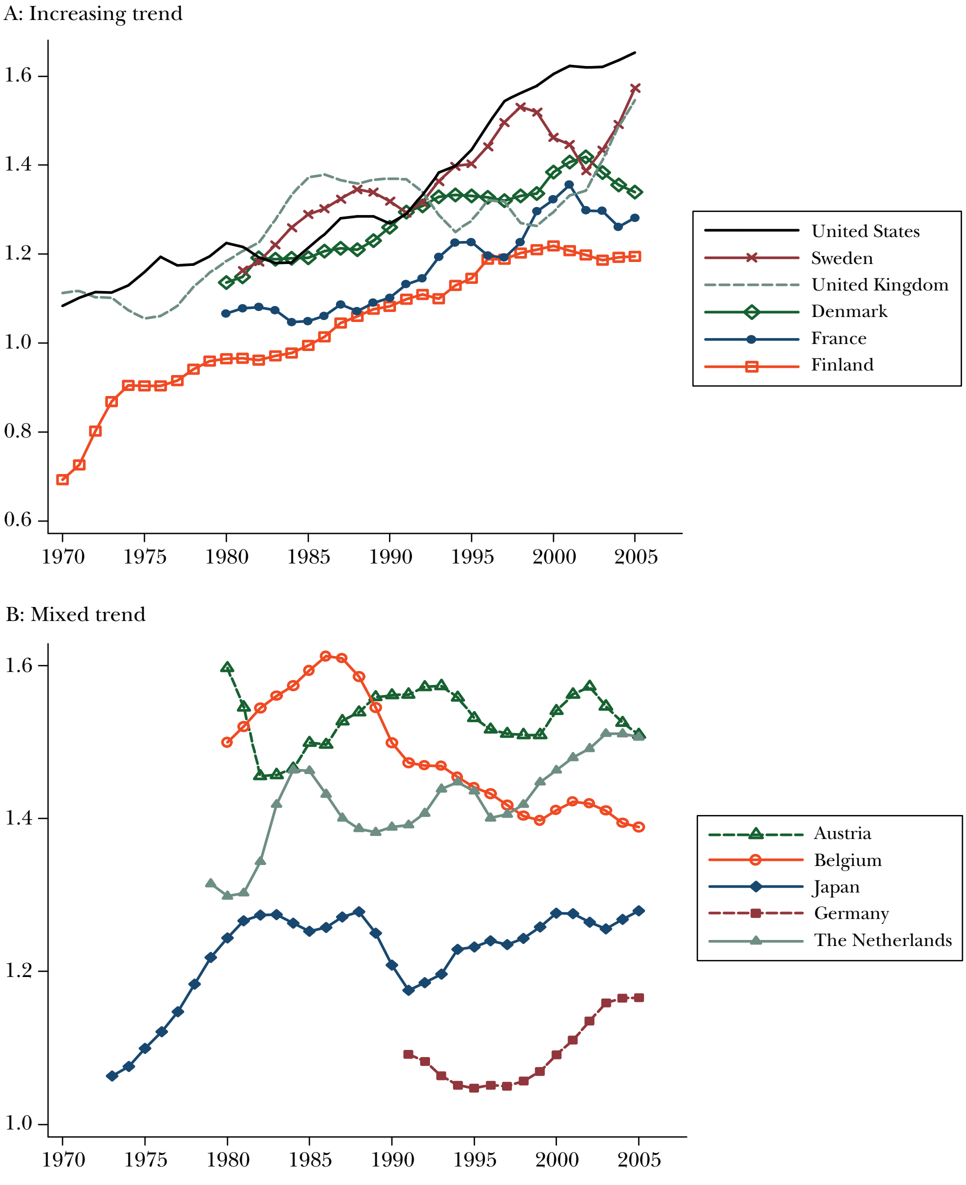
A long literature points to the fact that information and communication technology increase demand for highly educated workers; for example, see Autor, Katz, and Krueger (1998). And as we argue in Philippon and Reshef (2012), financial deregulation differentially increases demand for skill in finance in the United States. Moreover, these two factors can also affect wages. We examine these hypotheses briefly below in an international context. In ongoing work (Boustanifar, Grant, Philippon, and Reshef 2012), we study systematically several other potential driving factors behind demand for skill and wages in finance. Here we report some preliminary findings.

Financial Regulation

Tight financial regulation limits the range of permissible activities and it forces standard transparent reporting, which in turn restricts the creativity of skilled workers and limits the complexity of their operations. In addition, standardization and

Figure 6

Relative Wage of Skilled Labor in Finance



A: Increasing trend

B: Mixed trend

Source: Authors' calculations using data from EU KLEMS.

Notes: The figures report the average wage of skilled workers in finance relative to the average wage of skilled workers in the whole economy. Average wages are computed by dividing labor compensation by full-time equivalent employment. High-skilled workers in all countries are comparable and attain at least a college or university degree. Data for Canada are not available from EU KLEMS. Panel A groups countries that exhibit an increasing trend. Panel B groups countries that exhibit a mixed trend, or roughly no trend since 1980.

limiting complexity reduces the need to use wage contracts with high-power incentives. Indeed, in Philippon and Reshef (2012), we conclude that financial regulation is the main determinant of both demand for skill and wages in the US financial sector, along with other factors including technology, nonfinancial corporate activity, and financial globalization, which play a secondary role. Does financial deregulation correlate well with wages and demand for skill in our cross-country sample?

To try to answer this question we use data from Abiad, Detragiache, and Tressel (2008), who study financial reform (which is not necessarily deregulation) along seven dimensions in 1973–2005: reduction in credit controls, removal of interest rate controls, removal of entry barriers, privatization, capital account liberalization, securities market development, and introduction of prudential regulation and supervision. These measures do not take into account organizational and activity restrictions that are important for the financial landscape, particularly for the United States: bank branching and separation of investment banking from retail banking. Major changes occurred in these important aspects of the regulatory environment in the United States and are taken into account in the index we constructed in Philippon and Reshef (2012) but not in the Abiad, Detragiache, Tressel (2008) data.

We construct an index of financial deregulation that aggregates seven dimensions of financial reform.[7] A clear pattern emerges. Starting in the 1970s, the level of financial regulation is relatively heterogenous across countries: Austria, Sweden, and France have relatively high levels of financial regulation, while Canada, the Netherlands, and Germany have relatively low levels. However, over time all countries move toward deregulation and generally converge to a more lightly regulated regime.

With some exceptions, countries that deregulate more also experience larger increases in relative skill intensity in finance. The exceptions are Austria and Denmark, which are among the countries that deregulate their financial sector most aggressively but do not experience large increases in relative skill intensity. Other countries line up more closely.

The relationship between deregulation and relative wages in finance is less clear. For example, according to our index, the United States, the Netherlands, and Canada start the sample with relatively light regulation and therefore in the context of this comparison do not deregulate much. But these countries experience larger increases in relative wages in finance, both on average and for skilled workers. Starting from relatively tight regulation, Austria and Belgium deregulate aggressively, but their financial sectors do not exhibit increases in relative wages.[8]

[7] See the online Appendix available with this paper at http://e-jep.org for complete description and Appendix Figure A4 for the evolution of the index for all countries in the sample. A detailed description of the changes in each dimension of financial regulation over the sample are reported in Appendix Table A2.

[8] An alternative source of data on bank regulation is from Barth, Caprio, and Levine (2008), who document a multitude of dimensions of bank regulation in 1999 and 2007. Despite the shorter period and its focus on banking alone, this dataset has invaluable detail on the scope of bank activities and organization of the industry, which is in line with our view on how regulation affects demand for skilled labor and the

Technology

Workers in finance need to collect, process, and analyze information, so it is no surprise that the financial sector was an early adopter of information and communication technology.[9] It is widely accepted that information technology is particularly complementary to complex tasks (more specifically, nonroutine cognitive tasks) and that it substitutes for routine tasks (Autor, Levy, and Murnane 2003). Educated (skilled) workers tend to perform complex tasks, so relative demand for such workers increases with investment in information technology. Moreover, if there is heterogeneity among educated workers in the degree to which they are productive using information and communication technology, we may see skilled wages increase more in industries that invest more in information and communication technology.

We use data on the share of information and communication technology (ICT) capital in total capital compensation from the European Union KLEMS dataset, using constant 1995 prices. This is a measure of the intensity of ICT capital *use*, which takes into account both quantities and prices (rather than quantities alone or value of capital installed). For the United States, we use data from the Bureau of Economic Analysis (Fixed Assets Tables). Data for Canada is not available from the EU KLEMS, so we do not include Canada here.

Figure 7 shows the difference between the intensity of information and communication technology in the financial sector and its intensity in the whole economy. In most countries—with the United States the notable exception—finance has increased its ICT intensity much more than in the whole economy. The surprising result for the United States is driven by the fact that as a whole the United States is among the most intensive economies in using information and communication technology whereas its financial sector is not particularly intensive in its use of information and communication technology relative to financial sectors elsewhere.

Regression Analysis

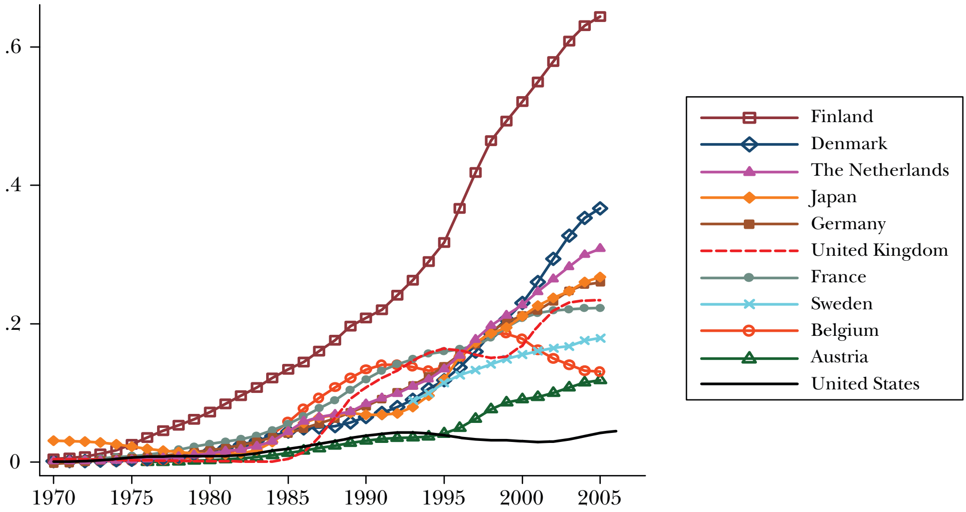
To what extent can financial deregulation and investment in information and communication technology explain various characteristics of the financial sector in this cross-country data? We expect differential positive effects on demand for skilled workers resulting from complementarity between these two variables. We also expect differential effects on the wages of skilled labor if there is need for higher-quality skilled workers to perform more data analysis and to be more creative.

wages they command. Changes in regulation according to this measure are not strongly correlated with changes in regulation in Abiad, Detragiache, and Tressel (2008) in the relevant period. We acknowledge that both of these regulation indices are limited either in scope or in time coverage. Here we only test the explanatory power of financial deregulation based on Abiad, Detragiache, and Tressel (2008) due to its longer sample.

[9] Yates (2000) reports evidence of early information and communication technology adoption during the previous information revolution, starting at the end of the 19th century. Although most of the evidence is for management in manufacturing, some examples exist for insurance.

Figure 7

Relative ICT (Information and Communication Technology) Capital Share in Finance



Source: Authors' calculations using data from EU KLEMS.
Notes: The figure reports the difference between the ICT (information and communication technology) capital share in finance and the ICT share in the whole economy, using constant prices in 1995. Data for Canada are not available from the EU KLEMS. Data for the US are from the Bureau of Economic Analysis, Fixed Assets Tables. Series are three-year moving averages.

Table 1 offers some illustrative regressions. In these regressions, we use three dependent variables: relative skill intensity in finance (see Figure 5); the relative wage of finance (see Figure 4); and the relative wage of skilled labor in finance (see Figure 6). The first variable captures demand for skill, the second overall compensation, while the third captures the differential wages of skilled workers in finance.[10]

All regressions include country fixed effects to account for systematic differences across countries. In even columns, we add year fixed effects to account for common trends. We standardize all the variables in the regressions over the entire sample, so the coefficients can be interpreted as the effect of one standard deviation change in the regressor on the regressand, also in terms of standard deviations (beta coefficients). The regressors are lagged by one year to allow for delayed effects, although results using longer lags or no lags are similar. We drop the United States from these regressions since we find the deregulation index woefully inadequate to describe the changes in regulatory environment in the US economy.

In column 1 in Table 1, we see that relative skill intensity in finance is positively associated with both deregulation and information and communication

[10] See Table A3 in the online Appendix available with this paper at http://e-jep.org for descriptive statistics for all variables.

Table 1
Determinants of Skill Intensity and Wages in Finance

	Dependent variables:					
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Relative skill intensity		*Relative wage*		*Relative skilled wage*	
Financial deregulation, $t-1$	0.199*** (0.027)	0.123*** (0.041)	0.066 (0.042)	−0.074 (0.061)	0.091** (0.040)	−0.069 (0.062)
Relative ICT share, $t-1$	0.301*** (0.026)	0.102** (0.041)	0.287*** (0.042)	0.268*** (0.074)	0.275*** (0.038)	0.235*** (0.061)
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	No	Yes	No	Yes	No	Yes
Observations	254	254	297	297	254	254
R^2, within	0.67	0.74	0.27	0.34	0.35	0.47
Number of countries	10	10	10	10	10	10

Source: Authors.
Notes: In these regressions, we use three dependent variables: relative skill intensity in finance; the relative wage of finance; and the relative wage of skilled labor in finance. We standardize all the variables in the regressions over the entire sample, so the coefficients can be interpreted as the effect of one standard deviation change in the regressor on the regressand, also in terms of standard deviations (beta coefficients). The regressors are lagged by one year. We drop the United States from these regressions. ***, and ** indicate levels of significance of 1 percent and 5 percent.

technology; this result is robust to including year fixed effects (column 2). Countries that deregulate more and increase the intensity of investment in information and communication technology see demand for skill rise more than average; this is in line with our results in Philippon and Reshef (2012). The size and statistical significance of the year fixed effects increases over time (not shown), indicating that there is, in addition, a common trend.[11]

We now turn to relative wages. In columns 3 and 4 we see that higher relative wages in finance are associated with information and communications technology, but not with deregulation. Once again, the size and statistical significance of the year fixed effects increase over time (not reported here). Results for relative wages of skilled labor are similar (columns 5 and 6): intensity of information and communications technology is a robust predictor of wages, but deregulation is not. One potential explanation for this is that the measure of deregulation used here does not capture essential dimensions that are important for wages. Another issue is that variation in income taxes influences wages but is omitted from the analysis here.

In all regressions that include year effects, their size and statistical significance increase over time. What may be accounting for the common trends in demand for skill and wages in finance? In Philippon and Reshef (2012), we find that financial

[11] Results using an alternative measure for the demand for skill, namely the wage bill share of skilled workers, are very similar. See Table A4 in the online Appendix available with this paper at http://e-jep.org.

(and trade) globalization does not affect relative skill intensity in finance in the United States. However we do find that it helps explain relative wages and in fact reduces significantly the explanatory power of deregulation in our historical wage regressions. We leave it for future research to determine whether this conjecture holds in the international sample as well. We investigate this point systematically in Boustanifar, Grant, Philippon, and Reshef (2012).

We conclude this section by noting that deregulation and information and communication technology may be associated with the overall relative increase in labor costs in finance, which contributes to the size of the sector, but there is also scope for common global trends that are not country specific.

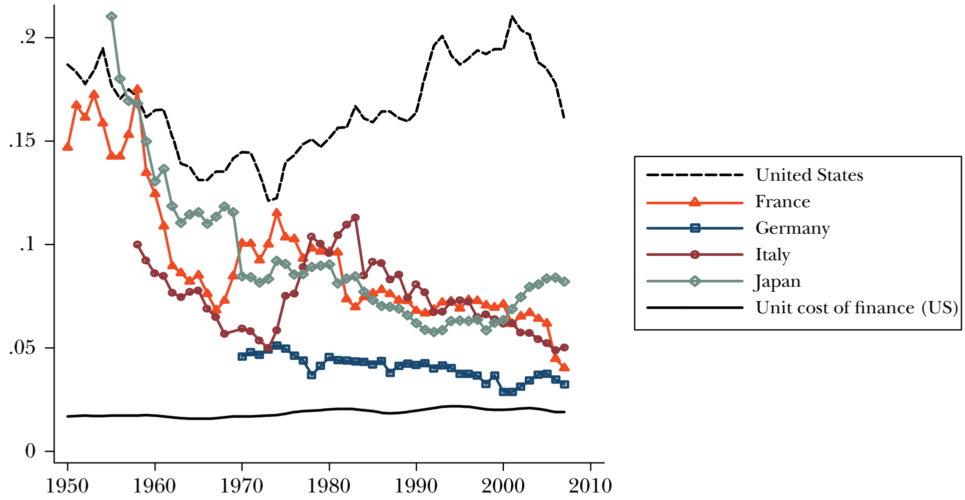
Costs versus Output

Has the rise in financial sector value added in the United States been matched by an increase in the cost per unit of financial services produced? At a conceptual level, this poses the difficult problem of measuring a "unit" of financial services, and adjusting for changes in composition and quality. Philippon (2012) reports a painstaking effort to measure correctly the unit cost of financial intermediation. Executing such a measure for a broad set of countries is a formidable task, which we hope future research will tackle. Here we provide a much cruder measure: We simply divide value added in finance by the outstanding value of bank loans to nonfinancial entities (firms in the private sector, government, and households) from Schularick and Taylor (2012). In addition to the United States, we only do this for four other countries: France, Germany, Italy, and Japan. We restrict attention to these countries because they all have financial sectors that are relatively heavily reliant on banks.

Figure 8 reports the cost ratio of finance value added divided by bank loans, together with the quality-adjusted unit cost measure for the United States from Philippon (2012). The measure of finance value added divided by bank loans is much higher than the unit cost measure. This is a manifestation of the fact that bank loans do not encompass all financial outputs. For the United States, the cost ratio does not trend in the sample, which is consistent with the relatively flat unit cost. For the other countries it falls. We observe qualitatively similar trends when we look at the ratio of value added in banking alone relative to loans (not reported here). Thus, at least using this crude measure, we conclude that the rise of the income share of finance is not driven by an increase in cost per unit of intermediation.

Next, we ask whether changes in the quality of financial services can help explain the recent rise of the income share of finance in the United States relative to other countries. If higher quality comes at a higher cost, then the puzzle is solved. For example, the proliferation of derivatives markets could in theory have benefitted the economy by improving the informativeness of stock prices. But Bai, Philippon, and Savov (2011) find that the predictive power of US stock prices is stable over the last 50 years. And Hadas (2011) argues that commodities' prices have become *less* informative.

Figure 8
Finance Value Added Divided by Bank Loans



Sources: Bank loans are from Schularick and Taylor (2012). Finance value added is from EU KLEMS or STAN (OECD); Italy in 1958–1968 from Istituto Centrale Di Statistica; Japan in 1955–1969 from the Economic and Social Research Institute, Cabinet Office, Government of Japan.
Notes: The figure reports the ratio of finance value added divided by bank loans to nonfinancial entities (firms in the private sector, government and households) for various countries. It also reports "Unit cost of finance (US)," a quality-adjusted unit cost of finance measure for the United States from Philippon (2012).

An alternative approach is to look for signs that, by some measure, US financial markets are performing in a way that allocates capital more effectively. If this comes at a higher cost, then the puzzle is solved. Better-functioning financial markets could in theory help households improve the diversification of their risk, but there is no strong evidence for an increase in consumer risk sharing, let alone evidence that this has happened to a greater extent in the United States. In fact, Aguiar and Bils (2011) show that consumption inequality has closely tracked income inequality over the period 1980–2007. Alternatively, better-functioning financial markets could improve the allocation of capital across firms. This outcome is difficult to measure, but Hsieh and Klenow (2009) look at the dispersion of marginal productivity across US manufacturing firms and estimate the potential gains in total factor productivity from removing allocative inefficiencies in these firms. They find potential gains of 36 percent in 1977, 31 percent in 1987, and 43 percent in 1997. This suggests that the allocation of capital across US manufacturing firms has deteriorated, because the potential gain from removing allocative inefficiencies has increased from 1977 to 1997. Using similar methodology, Osotimehin (2012) finds no trend in potential gains in total factor productivity in French manufacturing over 1991–2006. These findings are at odds with improvements in allocation of capital and risk sharing. However, if there is more innovation in the United States and more young firms, then intermediation can be more expensive because it is difficult to screen and monitor such firms, as suggested by Philippon (2012) and Laeven, Levine, and Michalopoulos (2012).

Yet another possible explanation for the increase in the cost of financial intermediation is the increased concentration in the US banking sector from 1980 and on. The number of US commercial banks insured by the Federal Deposit Insurance Corporation hovered around 14,000 for most of the twentieth century, but started dropping more-or-less continuously after 1984, until it reached 6,300 in 2011. Similarly, the number of FDIC-insured saving institutions dropped continuously from 3,400 in 1984 to 1,067 in 2011. Commensurately, Haldane (2010) shows that the total assets of top-three US banks as a percent of total commercial banking sector assets shows no trend until 1990, after which it rises from 10 to 40 percent in 2007. Although Haldane (2010) also shows that similar trends prevail in the United Kingdom, it is still possible that market power in the US banking industry has increased more than elsewhere.

Finally, Greenwood and Scharfstein (this issue) provide an interesting analysis by looking into the black box of the finance industry in the United States. They find that much of the growth of finance is accounted for by an increase in investments under active management, which command relatively high—albeit not increasing—fees. This has been driven by an increase in households' participation in the stock market. Greenwood and Scharfstein argue that the growth in active management may benefit households by improving diversification; and that by lowering the cost of capital, this benefits particularly young entrepreneurial firms. But this answer begs the question: Why did active management grow so much in the United States? And has this happened elsewhere? These are interesting questions for future research to answer.

Conclusions

A well-functioning financial sector facilitates information transmission, risk sharing, and allocation of capital, which are key components for the success of capitalist economies. Thus, the rise of the financial sector is sometimes defended by arguing that a more developed financial sector encourages economic growth. Indeed, in broad cross sections of countries, a larger financial sector is positively correlated with economic growth (for example, Rousseau and Sylla 2003; Levine 2005).

But it is quite difficult to make a clear-cut case that at the margin reached in high-income economies, the expanding financial sector increases the rate of economic growth. The long-run patterns of the rise of the financial sector since the nineteenth century, shown in Figure 1, do not have any obvious correlation with trends in growth rates within countries.

Moreover, Figures 1 and 2 demonstrate that the relationship between the size of the financial sector and income is complex, and that most of the rise in living standards from 1870 was obtained with less financial output and a smaller share of income spent on finance than what is observed after 1980. It also seems that at the current height of development, the relationship between financial output and income per capita may have changed.

There may very well be third factors driving both finance and income: For example, Acemoglu and Robinson (2012) argue that the institutional foundations of prosperity were laid out by the middle of the nineteenth century in many of today's high-income countries (with roots long before that). This type of change can simultaneously cause growth of income, industrialization, and financial development. At a minimum, the secular rise in the financial sector does not seem to deliver *faster* growth. But if finding more growth opportunities becomes ever harder with development, then a larger financial output and a larger share of income may be needed to sustain growth in the sample of now-industrialized countries that we investigate.[12]

Of course, any analysis of the interrelationship between the growth of the financial sector and economic growth in recent decades must also take into account the global recession that began in 2007 and the stagnant growth that has followed. The growth of finance is normally commensurate with growth in credit, but sometimes credit runs out of check. Jordá, Schularick, and Taylor (2011) find that recessions that coincide with excessive credit are deeper and longer, both for normal recessions and financial crisis recessions; and Schularick and Taylor (2012) find that more credit increases the likelihood of a financial crisis. Haldane (2010) estimates the net present value of the most recent crisis between one and five times annual world GDP.

Assessing whether there is "too much" finance—as Arcand, Berkes, and Panizza (2012) and Cecchetti and Kharroubi (2012) argue—must take account of not only diminishing benefits, but also costs, and of counterfactual scenarios in which the growth of finance is inhibited. Whether the social benefits outweigh the costs of the growth of finance is still an open question. Measuring the net social benefits of the growth of finance is a difficult task, which we do not take up here. Instead, this paper discusses some of the determinants of the growth of finance, and asks whether the size of the sector is commensurate with supply of bank credit. While it is difficult to believe that the growth of finance has not come with some benefits—either a wider reach or an increase in quality of services—our findings show that this conclusion is not straightforward, especially for the subset of economies with large and growing financial sectors. Researchers are still in the process of building a model that adequately explains the rise of the financial sector. Based on the time-series and cross-country evidence in this paper, we would argue that any such model needs to fit several facts.

First, the financial sector share of income grows over time. But even within high-income countries, finance reaches very different sizes and represents very different

[12] This idea is akin to Milton Friedman's thermostat analogy (Friedman 2003): Keeping growth constant may require varying degrees of finance, and lately we may be in need of much more of the stuff to keep on at the same growth rate.

shares of the economy. In particular, the US financial sector experiences the largest rise in the share of its financial sector. This phenomenon should be understood separately from the general rise in the share of services across countries.

Second, there is no particular correlation between the size of the financial sector and economic growth in time series data. Moreover, the correlation between financial output and per capita income varies considerably over the last 130 years. While there is a positive relationship between credit and income in the period after 1950, this relationship changes considerably after 1980 when income grows more slowly relative to credit.

Third, wages in finance—average and skilled—have grown relative to wages in the economy as a whole for many countries. Some countries exhibit mixed trends, but in those countries, finance wages are relatively high to begin with.

Fourth, financial services have become relatively more skill-intensive since 1970, and financial deregulation and investment in information and communication technology play a role in explaining this. In addition, there is scope for common global factors, such as increased competition between financial centers to help explain these trends.

Fifth, the rise of finance is not likely to be explained by a rise in the unit cost of financial services.

Our discussion is complementary to Greenwood and Scharfstein's paper in this issue, which provides an illuminating and insightful analysis of the black box of finance. They attribute a sizable portion of the growth of finance in the United States to the increase in active asset management and to an extension of household credit (mostly mortgages). They argue that the growth of active management in the United States is a benefit that came at the cost of management fees; and that the growth of household credit is a benefit that came at the cost of financial stability. These activities are related to higher fees, and are likely related to more skilled labor, which may require higher compensation.

As we build a deeper understanding of what drives growth in the financial sector, both over time within national economies and in cross-country comparisons, we will be in a better position to evaluate in a more rigorous way whether finance is too big, or too expensive, from a social point of view. But the available evidence at present suggests that at the very high end of financial development, rapidly diminishing social returns may have set in.

■ *We thank William Johnson, Hernan Moscoso Boedo, Sophie Osotimehin, Alan Taylor, and Eric Young for helpful discussions. Hamid Boustanifar, Everett Grant, and Alice Li provided excellent research assistance. This paper has greatly benefitted from the comments of the Managing Editor and the other editors who reviewed the paper.*

References

Abiad, Abdul, Enrica Detragiache, and Thierry Tressel. 2008. "A New Database of Financial Reforms." International Monetary Fund Working Paper 08/266.

Acemoglu, Daron, and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* New York: Crown.

Aghion, Philippe, Peter Howitt, and David Mayer-Foulkes. 2005. "The Effect of Financial Development on Convergence: Theory and Evidence." *Quarterly Journal of Economics* 120(1): 173–222.

Aguiar, Mark A., and Mark Bils. 2011. "Has Consumption Inequality Mirrored Income Inequality?" NBER Working paper 16807.

Arcand, Jean-Louis, Enrico Berkes, and Ugo Panizza. 2012. "Too Much Finance?" IMF Working Paper No. 12/161.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger. 1998. "Computing Inequality: Have Computers Changed the Labor Market?" *Quarterly Journal of Economics* 113(4): 1169–1214.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118(4): 1279–1333.

Bai, Jennie., Thomas Philippon, and Alexi Savov. 2011. "Have Financial Markets Become More Informative?" http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2023889.

Bakija, Jon, Adam Cole, and Bradley T. Heim. 2012. "Jobs and Income Growth of Top Earners and the Causes of Changing Income Inequality: Evidence from U.S. Tax Return Data." http://home.comcast.net/~bradheim/files/BakijaColeHeimJobsIncomeGrowthTopEarners.pdf.

Barth, James R., Gerard Caprio Jr., and Ross Levine. 2001. "The Regulation and Supervision of Banks around the World: A New Database." World Bank Policy Research Working Paper 2588.

Barth, James R., Gerard Caprio Jr., and Ross Levine. 2008. "Bank Regulations Are Changing: For Better or Worse?" World Bank Policy Research Working Paper 4646.

Baumol, William J. 1967. "Macroeconomics of Unbalanced Growth: The Anatomy of the Urban Crisis." *American Economic Review* 57(3): 415–26.

Bencivenga, Valerie R., and Bruce D. Smith. 1991. "Financial Intermediation and Endogenous Growth." *Review of Economic Studies* 58(2): 195–209.

Boustanifar, Hamid, Everett Grant, Thomas Philippon, and Ariell Reshef. 2012. "Wages and Human Capital in Finance: International Evidence, 1970–2007." Working Paper, University of Virginia.

Buera, Francisco J., and Joseph P. Kaboski. 2012a. "The Rise of the Service Economy." *American Economic Review* 102(6): 2540–69.

Buera, Francisco J., and Joseph P. Kaboski. 2012b. "Scale and the Origins of Structural Change." *Journal of Economic Theory* 147(2): 684–712.

Cecchetti, Stephen G., and Enisse Kharroubi. 2012. "Reassessing the Impact of Finance on Growth." Bank of International Settlements, conference draft.

Central Bureau of Statistics of Norway. 1969. *Historical Statistics 1968.* (In English and Norwegian.) Olso, Norway.

den Bakker, Gert P., and Jan de Gijt. 1990. "Who Came Off Worst: Structural Change of Dutch Value Added and Employment during the Interwar Period." Central Bureau of Statistics, The Netherlands, National Accounts Research Division Paper NA-040.

EU KLEMS. N.d. A dataset. http://www.euklems.net/.

Friedman, Milton. 2003. "The Fed's Thermostat." *Wall Street Journal,* August 19.

Goldin, Claudia, and Lawrence F. Katz. 2008. "Transitions: Career and Family Life Cycles of the Educational Elite." *American Economic Review* 98(2): 363–69.

Greenwood, Jeremy, and Boyan Jovanovic. 1990. "Financial Development, Growth, and the Distribution of Income." *Journal of Political Economy* 98(5, Part 1): 1076–1107.

Greenwood, Jeremy, Juan M. Sanchez, and Cheng Wang. 2010. "Financing Development: The Role of Information Costs." *American Economic Review* 100(4): 1875–91.

Hadas, Edward. 2011. "Commodity Prices are Failing New Zealand Test." *Financial Times,* August 8.

Haldane, Andrew G. 2010. "The $100 Billion Question." Comments given at the Institute of Regulation and Risk, Hong Kong, March 30. http://www.bankofengland.co.uk/publications/Documents/speeches/2010/speech433.pdf.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *Quarterly Journal of Economics* 124(4): 1403–48.

Instituto Centrale di Statistica. Various years. *Annuario Statistico Italiano.* Roma.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2011. "When Credit Bites Back: Leverage, Business Cycles, and Crises." NBER Working Paper 17621.

Kaplan, Steven N., and Joshua Rauh. 2010. "Wall Street and Main Street: What Contributes to the Rise in the Highest Incomes?" *Review of Financial Studies* 23(3): 1004–1050.

King, Robert G., and Ross Levine. 1993. "Finance and Growth: Schumpeter Might Be Right." *Quarterly Journal of Economics* 108(3): 717–37.

Laeven, Luc, Ross Levine, and Stelios Michalopoulos. 2012. "Financial Innovation and Endogenous Growth." http://faculty.haas.berkeley.edu/ross_levine/Papers/financial_innovation_9_10.pdf.

Levine, Ross. 1991. "Stock Markets, Growth, and Tax Policy." *Journal of Finance* 46(4): 1445–65.

Levine, Ross. 2005. "Finance and Growth: Theory and Evidence." In *Handbook of Economic Growth* edited by P. Aghion, and S. N. Durlauf, vol. 1A, pp. 865–934. Elsevier.

Maddison, Angus. 2010. "Historical Statistics of the World Economy: 1–2008 AD." Dataset available at the Groningen Growth and Development Centre. http://www.ggdc.net/maddison/Maddison.htm.

Obstfeld, Maurice. 1994. "Risk-Taking, Global Diversification, and Growth." *American Economic Review* 84(5): 1310–29.

Obstfeld, Maurice, and Alan M. Taylor. 2004. *Global Capital Markets: Integration, Crisis and Growth.* Cambridge University Press.

OECD STAN. N.d. STAN STructural ANalysis Database. http://www.oecd.org/industry/ind/stanstructuralanalysisdatabase.htm.

Office Statistique des Communautes Europeennes. 1966. *Comptes Nationaux 1955–1965.* EUROSTAT, Bruxelles.

O'Mahony, Mary, and Marcel P. Timmer. 2009. "Output, Input and Productivity Measures at the Industry Level: The EU KLEMS Database." *Economic Journal* 119(538): F374–F403.

Osotimehin, Sophie. 2012. "Aggregate Productivity and the Allocation of Resources over the Business Cycle." Available at: https://sites.google.com/site/sosotimehin/papers.

Oyer, Paul. 2008. "The Making of an Investment Banker: Stock Market Shocks, Career Choice, and Lifetime Income." *Journal of Finance* 63(6): 2601–28.

Philippon, Thomas. 2012. "Has the U.S. Finance Industry Become Less Efficient? On the Theory and Measurement of Financial Intermediation." NBER Working Paper 18077.

Philippon, Thomas, and A. Reshef. 2012. "Wages and Human Capital in the U.S. Financial Industry: 1909–2006." *Quarterly Journal of Economics* 127(4): 1551–1609.

Rousseau, Peter L., and Richard Sylla. 2003. "Financial Systems, Economic Growth, and Globalization." Chapter 13 in *Globalization in Historical Perspective*, edited by M. D. Bordo, A. M. Taylor, and J. G. Williamson. University of Chicago Press.

Schularick, Moritz, and Alan M. Taylor. 2012. "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles and Financial Crises, 1870–2008." *American Economic Review* 102(2): 1029–61.

Smits, J.-P., P. J. Woltjer, and D. Ma. 2009. "A Dataset on Comparative Historical National Accounts, ca. 1870–1950: A Time-Series Perspective." Groningen Growth and Development Centre Research Memorandum GD-107, University of Groningen. http://ggdc.eldoc.ub.rug.nl/root/WorkPap/2009/GD-107/.

Urquhart, M. C. 1993. *Gross National Product, Canada, 1870–1926: The Derivation of the Estimates.* McGill-Queen's University Press.

Vamplew, Wray. 1987. *Australians, Historical Statistics.* Volume 9 of *Australians.* Australia: Fairfax, Syme & Weldon Associates.

Yates, JoAnne. 2000. "Business Use of Information and Technology during the Industrial Age." Chap. 5 in *A Nation Transformed by Information: How Information Has Shaped the United States from Colonial Times to the Present*, edited by A. D. Chandler Jr., and J. W. Cortada. Oxford University Press.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 97–108

Asset Management Fees and the Growth of Finance

Burton G. Malkiel

From 1980 to 2006, the financial services sector of the United States economy grew from 4.9 percent to 8.3 percent of GDP. A substantial share of that increase was comprised of increases in the fees paid for asset management. This paper examines the significant increase in asset management fees charged to both individual and institutional investors. Despite the economies of scale that should be realizable in the asset management business, the asset-weighted expense ratios charged to both individual and institutional investors have actually risen over time. If we exclude index funds (an innovation that has made market returns available even to small investors at close to zero expense), fees have risen substantially as a percentage of assets managed.

One could argue that the increase in fees charged by actively managed funds could prove to be socially useful, if it reflected increasing returns for investors from active management or if it was necessary to improve the efficiency of the market for investors who availed themselves of low-cost passive (index) funds. But neither of these arguments can be supported by the data. Actively managed funds of publicly traded securities have consistently underperformed index funds, and the amount of the underperformance is well approximated by the difference in the fees charged by the two types of funds. Moreover, it appears that there was no change in the efficiency of the market from 1980 to 2011. Arbitrage opportunities to obtain excess risk-adjusted returns do not appear to have been available at any time during the early part of the period. Passive portfolios that bought and held all the stocks in a broad-based market index substantially outperformed the average active manager throughout the entire period. Thus, the increase in fees is likely to represent a

■ *Burton G. Malkiel is Chemical Bank Chairman's Professor of Economics, Emeritus, Princeton University, Princeton, New Jersey, and Chief Investment Officer for Wealthfront, a software-based financial advisory firm.*

http://dx.doi.org/10.1257/jep.27.2.97. doi=10.1257/jep.27.2.97

deadweight loss for investors. Indeed, perhaps the greatest inefficiency in the stock market is in "the market" for investment advice.

Economies of Scale in Asset Management

There should be substantial economies of scale in asset management. It is no more costly to place an order for 20,000 shares of a particular stock than it is to order 10,000 shares. Brokerage commissions (which are usually set in a flat dollar amount per transaction, at least within broad ranges of transaction size) are likely to be similar for each purchase ticket, as are the "custodial fees" paid to the bank that holds the securities that are owned. The same annual report and similar filings to the Securities and Exchange Commission are required whether the investment fund has $100 million in assets or $500 million. The due diligence required for the investment manager is no different for a large mutual fund than it is for a small one. Modern technology has fully automated such tasks as dividend collection, tax reporting, and client statements.

To be sure, an active investment manager of a small company (so-called "small-cap") fund may find that somewhat more effort will be required than for the management of large-cap funds. This is so because diversification and liquidity requirements will constrain the fund manager from holding too large a proportion of any one company's outstanding stock—which is a problem far less likely to arise for a fund investing in large ("large-cap") companies. Thus, the managers of small-cap funds are likely to be required to hold and follow a larger number of securities and to be far more concerned about the liquidity of their holdings. Nevertheless, the fund's infrastructure will not change. There will be no substantial additional expense in a small-cap fund for general market analysis, industry analysis, accounting, general oversight, or reporting requirements. Even if additional securities analysts need to be hired for a larger fund, expenses are likely to increase by only a small proportion of any increase in assets managed.

Academic research has documented substantial economies of scale in mutual fund administration. Latzko (1999) estimated a cost function for 2,610 mutual funds and concluded that the average cost curve for the typical mutual fund is downward sloping over the entire range of fund assets. Dyck and Pomorski (2011) documented substantial positive scale economies for asset managers of (defined benefit) pension plans. Coats and Hubbard (2007) do not dispute the existence of considerable economies of scale in the mutual fund industry, but argue that substantial competition exists in the industry. They argue that barriers to entry are low and new entry into the industry is common. What is undeniable, however, is that the fees paid to investment managers have increased substantially over time.

In 1980, the entire equity mutual fund industry managed less than $26 billion of assets. In 2010 the equity assets of the mutual fund industry totaled almost $3.5 trillion: thus, the total value of equity assets held by the mutual fund industry rose by a multiple of 135 times from 1980 to 2010. Surely, there had to be enormous economies of scale that could have been passed on to consumers, resulting in a lower cost of management as a percentage of total assets. But we will see below that

Table 1
Asset-Weighted Expense Ratios for Domestic Equity Funds
(in basis points)

	Including index funds	*Excluding index funds and ETFs**	*Share of equity mutual funds actively managed*
1980			
Expense ratios (basis points)	66.0	66.1	
Total assets (billions)	$25.81	$25.71	99.7%
1990			
Expense ratios (basis points)	83.3	85.0	
Total assets (billions)	$136.11	$131.69	96.8%
2000			
Expense ratios (basis points)	83.8	94.9	
Total assets (billions)	$2,158.50	$1,817.48	84.2%
2010			
Expense ratios (basis points)	69.2	90.9	
Total assets (billions)	$3,488.35	$2,473.59	70.9%

Source: Author using data from Lipper Analytic Services.
Note: Table 1 shows expense ratios (in basis points) for all equity mutual funds reporting to Lipper Analytic Services, as well as total assets (in billions of dollars).
**ETFs* are exchange-traded funds.

the scale economies in asset management appear to have been entirely captured by the asset managers. The same finding appears to hold for asset managers who cater to institutional investors.

Fees Paid to Mutual Fund Managers

Substantial fixed costs are involved in the formation and management of a mutual fund company. Executives of the fund need to be hired, including those responsible for portfolio management and marketing. A legal capability needs to be established to handle compliance and reporting requirements. If the fund is to be actively managed, security analysts must be employed. But as the assets of the fund grow, the fixed-cost infrastructure of the fund should comprise a smaller percentage of the fund's total assets. Fund management expenses should fall as a percent of fund assets.

Table 1 shows expense ratios for all equity mutual funds reporting to Lipper Analytic Services. Reading down the first column, which includes the universe of all funds, we see that expense ratios have been roughly flat over time. The annual expense ratio was 66.0 basis points (a basis point is 1/100 of 1 percent) in 1980 and 69.2 basis points in 2010. But the total assets of equity mutual funds increased by more than 135 times. Thus, the total expenses paid to equity mutual fund managers increased from $170.8 million to $24,143 billion—an increase of over 141 times. Holders of public mutual funds have made enormous contributions to the gross revenues flowing to the asset management industry. In the presence of widely recognized substantial economies of scale entailed in the asset-management business, we can conclude that the benefits of scale economies have largely been directed to asset managers rather than accruing to the benefit of fund shareholders.

However, one innovation in the asset management business—the index fund and its exchange-traded counterpart—has allowed the individual investor to benefit from scale economies. The first equity "index fund" (meaning, a fund that simply buys and holds all the funds in some, usually broad, stock-market index) was established by the Vanguard Group of Investment Companies in the late 1970s. While competition in the actively managed segment of the mutual fund market has primarily taken the form of product differentiation, the generic index fund part of the market has experienced vigorous price competition. In this indexed segment of the asset management industry, price competition has been fierce. Exchange-traded funds that track either the Standard and Poor's 500 Stock Index (an index that comprises about 75 percent of all listed stocks) or the Wilshire 5,000 Total Stock-Market Index are available to individual investors at expense ratios of 5 basis points or less. The third column of Table 1 indicates that the share of fund assets represented by low-cost index funds has grown substantially since 1980. The index mutual funds now comprise nearly one-third of the total mutual fund assets. The remainder consists of fund assets that are "actively managed" by investment management companies.

Column 2 of Table 1 presents the expense ratios of these actively managed equity mutual funds. These data show no evidence that scale economies have benefited shareholders in actively managed mutual funds. Expense ratios paid by the shareholders of actively managed funds have increased substantially from about 66 basis points in 1980 to over 90 basis points in 2010. While competition has driven down the expense ratios of index funds and exchange-traded funds, which trade like uniform commodities, competition has not lowered fees for the differentiated active funds.

Of course, when stated as a percentage of *assets,* fees do look low—close to 1 percent of assets for individuals. But a reasonable alternative way of appraising these fees is to compare them with the returns managers produce—in which case the fees no longer look "low." If overall stock-market returns average, say, 7 percent a year, then those same fees of 1 percentage point are actually about 14 percent of stock-market returns for individuals. If, instead, one measures fees as a percentage of the dividends distributed to mutual fund shareholders, mutual fund fees take up well over 50 percent of dividend distributions. But even these recalculations may substantially understate the *real cost* of active investment management. A more reasonable way to assess the benefits of active management is to measure fees as a percentage of the "excess" returns produced by active managers over the returns available from low-cost index funds; and these excess returns, as we will discuss in the last section of this paper, seem nonexistent. Finally, we should note that the fee numbers in Table 1 are asset-weighted. To the extent that mutual fund customers have switched from high-cost funds to low-cost ones, the data tend to make overall industry expense ratios look more moderate than they are.[1]

[1] The Securities and Exchange Commission has mandated more transparency with respect to fees, and mutual fund prospectuses are now required to contain fee information, stated in dollar amounts. Perhaps what might be more revealing would be a requirement to state those fees in terms of the percentage of the fund's long-run returns that have been consumed by fees.

Table 2
Average Fees Paid to Fund Managers for Institutional Investors
(in basis points, asset weighted)

A: Active domestic equity managers for corporate funds, publics funds, and endowments

	1996	*1999*	*2002*	*2005*	*2008*	*2011*
Corporate funds	52.9	54.4	54.2	54.9	53.5	55.0
Public funds	38.7	39.7	42.0	49.3	46.6	48.0
Endowments	51.3	51.3	59.9	59.1	64.4	64.0
Total	**46.8**	**46.6**	**52.4**	**54.1**	**54.7**	**55.0**

B: Active fixed-income managers for corporate funds, public funds, and endowments

	1996	*1999*	*2002*	*2005*	*2008*	*2011*
Corporate funds	32.6	34.3	27.5	28.0	29.7	30.0
Public funds	26.2	25.6	23.2	25.2	25.7	26.0
Endowments	29.6	30.4	27.1	29.0	34.7	36.0
Total	**29.0**	**29.1**	**26.3**	**27.3**	**30.0**	**30.1**

Source: Author using data from Greenwich Associates.

Before leaving this discussion of mutual fund fees, we need to acknowledge the arguments of the mutual-fund industry trade group, the Investment Company Institute, commonly known as the ICI. In a 2010 research report, the ICI has argued that the expense ratios of mutual funds have declined since 1990. What the ICI includes in their calculation of fund fees are so-called "sales costs" or "load fees." It is true that sales charges (for funds that do charge them) have declined over time (although many actively managed funds are so-called "no load" funds that have zero sales charges). According to the ICI, annualized sales loads have dropped from 0.99 percent of assets in 1940 to 0.13 percent of assets in 2009. This calculation is disputed by Bogle (2010b). Even if accurate, however, the reduction of sales charges simply reflects the drop in trading costs that has characterized the financial services industry. Brokerage commissions have declined as well. But the far larger and more important metric is the annual investment management expense fees charged by the asset management industry. As is shown in the data above, these fees have grown substantially.

Asset-management fees have also increased for institutional investors. While the level of institutional fees is lower than that for individual investors, the data in Table 2A show that expense ratios charged large institutional investors for active management of equity funds have increased from about 47 basis points to 55 basis points from 1996 to 2011. Table 2A shows that equity management expense ratios charged to corporate funds, public funds, and endowment funds have all increased over the past 15 years. Table 2B shows similar data for fixed-income managers (that is, managers who specialize in debt rather than equity). Expense ratios as a percentage of assets have been roughly flat. But because total fixed-income assets have increased over the 15-year period, total fees paid to fixed-income managers have increased significantly. We can conclude that asset-management fees for both institutional and individual investors have increased

Table 3
Percentage of US Equity Funds Outperformed by Benchmarks

		Percent outperformed	
Fund category	*Benchmark index*	*2011*	*5 years through 2011*
All domestic equity	S&P 1500	84%	62%
All large cap funds	S&P 500	81%	62%
All mid-cap funds	S&P Mid-Cap 400	67%	80%
All small-cap funds	S&P Small-Cap 600	86%	73%
Global funds	S&P Global 1200	69%	63%
International funds	S&P 700	69%	78%
Emerging market funds	S&P/IFCI composite	54%	83%

Source: Standard & Poor's and CRSP Survivor Bias-Free US Mutual Fund Data Base.
Note: Table 3 presents percentage of US equity funds that were outperformed by various benchmark indexes over the five-year period ending December 31, 2011.

substantially over time. This increase in asset-management fees has played an important role in the growth of the financial services industry since 1980.

Is the Increase in Asset-Management Fees Justified by the Value Added to Investors?

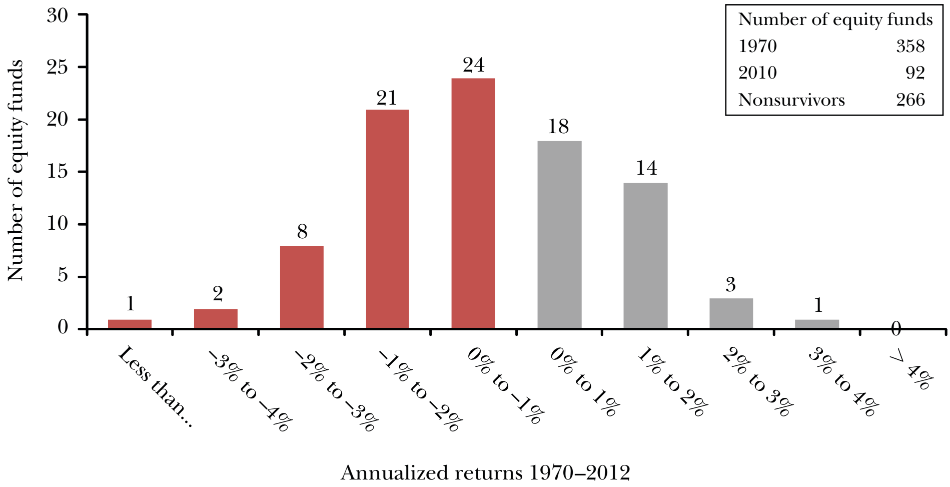
Whatever the costs charged to the owners of actively managed mutual funds, they could be more than justified if such funds produced superior returns for investors. After all, investors would happily pay annual fees of 1 percent of asset value to fund managers if active management produced gross returns that were 2 percent higher than passive index funds before the imposition of fees. Thus, the appropriate way to judge the economic benefits of expense ratios is to examine the relative returns of active and passive funds net of the fees charged. Fortunately, the complete records of both actively and passively managed mutual funds are available.

The data consistently provide overwhelming support for low-cost indexing as an optimal strategy for individual investors. 2011 was a particularly good year for indexing, because 84 percent of large capitalization fund managers were outperformed by the large-cap Standard and Poor's 500 Index. In addition, 82 percent of bond fund managers were outperformed by the Barclays U.S. Aggregate Bond Index. Similar numbers were recorded for managers of European stocks, emerging market equities, and small-cap managers. Over longer periods of time, about two-thirds of active managers are outperformed by the benchmark indexes, and the one-third that may outperform the passive index in one period are generally not the same as in the next period. In Malkiel (2011), I showed that there is little persistence in superior performance; indeed, whatever persistence there is in mutual fund returns reflects the fact that very high-cost funds do tend to exhibit somewhat consistent negative relative returns.

Table 3 presents percentages of US equity funds that were outperformed by various benchmark indexes over the five-year period ending December 31, 2011.

Figure 1

Returns of Surviving Funds: Mutual Funds 1970 to 2012, Compared with S&P Returns



Source: Author using data from Lipper Analytic Services.

Among actively managed funds, it was the small- and mid-cap funds (involving small- and medium-sized companies) and emerging markets funds and international funds that were even more likely to be outperformed by their benchmarks. While active fund managers often argue that markets are less efficient for smaller firms and for equities in emerging markets, whatever advantages may exist for active management in these sectors of the equity market appear to be outweighed by the higher fees charged relative to large-cap domestic equity management.

Figure 1 presents an analysis of the returns provided to investors over more than a 40-year period since 1970. In 1970, there were 358 equity mutual funds. (Today, thousands of active funds are marketed to the public.) Of the original group, 92 funds have survived. Hence, these data are compromised by survivorship bias. We can be confident that the 266 funds that did not survive had poorer records than did the surviving funds! Funds with especially poor records in a mutual fund complex are often merged into other funds with better past records. Yet even examining a dataset affected by substantial survivorship bias, the possibility of outperforming a broad-market index is extraordinarily small. One can count on the fingers of one hand the number of equity mutual funds that have beaten the market by two percentage points or more. My point is not that it is literally impossible to beat the market, but rather that investors who turn to active asset managers in an attempt to do so are far more likely to find themselves in the negative part of the distribution, rather than enjoying superior performance.

Table 4 presents detailed data on active fixed-income or bond portfolio management. Comparing Tables 3 and 4, we see that it is even less likely for active

Table 4
Percentage of Fixed Income Funds Outperformed by Benchmarks, Five Years through 2011

Fund category	*Comparison index*	*Percent outpeformed*
Government long	Barclays Long Government	94%
Government intermediate	Barclays Intermediate Government	67%
Government short	Barclays 1–3 Year Government	67%
Investment-grade long funds	Barclays Long Government/Credit	92%
Investment-grade intermediate funds	Barclays Intermediate Government/Credit	61%
Investment-grade short funds	Barclays 1–3 Year Government/Credit	94%
High-yield funds	Barclays High Yield	96%
Mortgage-backed securities funds	Barclays Mortgage-Backed Securities	75%
Global income funds	Barclays Global Aggregate	72%

Source: Standard & Poor's.

management of fixed income portfolios to produce excess returns over the returns from passive indexes. Even for high-yield bonds, where good credit analyses might be expected to produce excess returns, the percentage of managers outperforming their benchmark indexes is extremely small. Again, in the very areas where active management is often recommended—in this case, high yield bonds—the results are particularly dismal. The higher fees charged by such managers completely overwhelm whatever benefits they might produce.

It might be argued that even if active management has not produced excess returns for investors, the increase in fees supported socially useful arbitrage activities, which made the market more efficient. But there is no evidence that our markets were less efficient before the increase in fees. In a less-efficient market, managed funds would show better returns than unmanaged funds. But, according to Jensen (1968, 1975), even before 1980, active managers did not outperform their benchmarks. My own work (1995) comparing the returns of active managers versus passive index funds during the 1970s and 1980s showed no evidence that opportunities to earn excess returns existed before 1990. So the higher fees do not seem necessary to increase efficiency in the US equity and bond markets, as these markets showed no unexploited inefficiencies even before the increase in fees.

The Costs of Active Management

Despite the considerable economies of scale that exist in the active money management business, the annual fees charged to both individual and institutional investors have been either flat or rising over the past three decades. To be sure, the sales charges or load fees imposed on the purchases of most mutual funds have been

Table 5
Average Returns, Active Funds, versus Index
(20 years through 12/31/2011)

Large-Cap Equity Funds Average	7.18	Small-Cap Equity Funds Average	5.50*	Fixed Income Funds	5.69
S&P 500 Index	7.81	MSCI US Small-Cap 1750	6.98*	Barclays US Aggregate Bond Index	6.50
S&P 500 Index Advantage	0.64	MSCI US Small-Cap 1750 Advantage	1.48*	Barclays US Aggregate Bond Index Advantage	0.82

Source: Author using data from Lipper Analytic Services and Vanguard.
*Ten years of data to 12/31/2011.

lowered over the same period—just as brokerage commission costs of other types have declined. But ongoing asset-management fees have not reflected the scale economies that have been realized as the industry has grown. This increase in asset-management fees has contributed to the increase in the share of GDP accounted for by the financial services industry. At the same time, the financial innovations of index funds and exchange-traded funds have provided instruments that allow individual investors to obtain the returns offered by the stock and bond markets as a whole at virtually zero cost.

One could argue that the costs of active management can be justified by the benefits of promoting price discovery and market efficiency. But there is no evidence that the stock and bond markets were any more efficient in 2011 than they were in 1980. Here I use the term "efficiency" to reflect a lack of arbitrage opportunities that would enable active investment managers to beat the market after adjusting for risk. Active portfolio management has failed to generate excess returns rather consistently from 1980 to the present. Thus, the extra costs of active management do not benefit either investors or society as a whole.

We can estimate the costs borne by investors by comparing the average returns from actively managed mutual funds with low-cost index mutual funds or exchange-traded funds that track various market benchmarks. Most equity mutual funds invest in large capitalization stocks for which the appropriate benchmark is the Standard & Poor's 500 Stock Index. Table 5 presents the comparison. Over the past 20 years, it appears that investors paid 0.64 percent of the aggregate value of the total market capitalization in the (futile) search for superior returns. French (2008) made a similar comparison over the 1980–2006 period and found a 67 basis point advantage for passive investing. Table 5 shows an even larger advantage for fixed-income funds. The table also shows a 148 basis point advantage of passive over active management in small-cap funds, where the market is sometimes claimed to be less efficient. The larger gap reflects both the much higher management fees charged by small-cap managers and the increased costs of portfolio turnover with less-liquid smaller companies.

Why Do Excessive Fees Persist?

How can we explain the puzzle of why investors continue to pay excessive fees for financial services of such questionable value? Explanations that are unambiguously convincing may well be unachievable, as is the case for many of the puzzles in finance. But I would suggest that the following considerations play at least some role in increasing our understanding of what seems to be inexplicable consumer behavior.

Many consumers of financial services may judge the effectiveness or quality of investment advice by the price charged by the purveyor of the service. While the aspirin in a brand name like Bayer and in a generic product are identical, there are at least some other products where consumers correctly judge that the expensive, branded product is of higher quality than the lower-cost alternative. Kleenex is usually of higher quality than generic facial tissues. Q-tips are often superior to less-expensive cotton swabs. Thus, many consumers may view a branded, actively managed mutual fund to be superior to a generic index fund. For many consumers, the demand curve for mutual funds (over a certain range) may be positively sloped.

Advertising by the fund industry is geared to promote the idea that investing is very complicated, that "experts" are required to help, and that actively managed funds are really worth the high prices that are charged. Critics such as Bogle (2010a) have suggested that the fund industry is principally a marketing industry and advertisements are often misleading. Fund performance is often advertised as "outstanding," but the fine print reveals that this is true only for a carefully selected and limited time and against a carefully selected peer group or benchmark.

Overconfidence is also likely to play an important role in explaining investor behavior. Many investors may truly believe that they can select the best stocks and the best investment managers.

The fact that professional investors appear willing to pay excessive fees to their investment managers seems particularly puzzling. To be sure, the fees paid by institutions are lower than those paid by individuals. But institutional investors are usually highly sophisticated, and it is hard to believe that they naively accept earning inadequate returns while paying high management fees. Three factors may play at least a partial role in explaining this conundrum. First, institutional investors are particularly prone to suffer from overconfidence. Kahneman and Riepe (1998) and Kahneman (2011) have suggested that institutional investors may represent unique examples of overconfidence and hubris. They may truly believe that they will eventually earn excess returns despite historical evidence to the contrary. Second, it is important to point out that the most sophisticated institutions do not pay the average fees noted in Table 3. Investors such as Yale's David Swensen, author of what has been called "the endowment model" (2000), could easily negotiate lower fees since any asset manager would be delighted to have Yale University as a high-profile client.[2] Finally, we should note that more professional investors do

[2] One characteristic of the investing policies of universities and foundations is that much, if not most, of their endowments are considered permanent. Other institutional investors, such as pension funds, face a set of liabilities with fixed horizons. Universities have the advantage of considering that they face an

index their investments in publicly traded securities than is the case for individual investors. Professional investors index about one-third of their holdings of publicly traded securities.

The growth of indexing raises an interesting question. If every investor indexed, who would ensure that new information is rapidly incorporated into market prices? Surely one advantage of having an industry of active investment managers is that price discovery is enhanced and security prices are more likely to reflect accurately the underlying conditions of different companies. Thus, there is clearly some socially useful role for active management. What is less clear is whether we need nearly as much active money management as exists. My own guess is that there is far more professional market activity than is needed to ensure that we have an optimal amount of price discovery. Moreover, I can think of no reasonable argument that would suggest that the substantial rise in fees documented above was necessary to enhance the efficiency of market prices.

Concluding Comments

Our discussion of asset management fees reveals a paradox in its implications for the efficiency of markets. Clearly, one needs some active management to ensure that information is properly reflected in securities prices. Those professionals who act to exploit any differential—however small—between price and estimated value deserve to be compensated for their efforts. But it appears that the number of active managers and the costs they impose far exceed what is required to make our stock markets reasonably efficient, in the sense that no clear arbitrage opportunities remain unexploited. Worldwide, vast numbers of highly trained independent experts are expressing estimates of value each day. Outperforming the consensus of hundreds of thousands of professionals at the world's major financial institutions is next to impossible, as it has been for decades.

What has changed in the last few decades, however, is the financial innovation of the index fund and its cousin, the exchange-traded fund. Today, market-matching returns are now available to all investors at low "commodity" prices, on the order of 5 basis points (0.05 percent of assets) or less. Indeed, discount brokers exist (worldwide) who execute orders for exchange-traded funds at zero commissions.

Investors should consider fees charged by active managers not as a percentage of total returns, but as a percentage of the risk-adjusted incremental returns above the market. Thus, the fees charged by active portfolio managers should not be considered as 1 percent of assets or even 10 to 20 percent of total returns. Fees expressed as a percentage of the incremental returns earned by active managers are likely to exceed 100 percent. And since active managers often turn over their

infinite horizon. Thus, universities and foundations can easily invest in illiquid assets such as real estate, private equity, and so on. These markets are generally less efficient than the markets for publicly traded securities. Active management is quite appropriate in these markets, and these asset classes are also likely to earn illiquidity premiums for their investors.

portfolios about once a year, taxable individual investors will be subject to short-term capital gains taxes as well.

Of course the mutual fund industry as well as institutional asset managers, who thrive on high-fee actively managed funds management, are always trumpeting the benefits of switching into funds or managers with the best recent performance. For example, advertisements often suggest that individuals will be better off switching into funds with four- or five-star Morningstar ratings, despite Morningstar's acknowledgment that simply ranking funds by expense ratio provides a better predictor of future returns. In fact, Morningstar (see Kimmel 2012) studied the behavior of mutual fund investors from 2000 through 2011 and found that investors lost billions through their return-chasing behavior. Had they simply bought and held a broad-based index fund, they would have improved their return by almost 2 percentage points per year. The major inefficiency in financial markets today involves the market for investment advice, and poses the question of why investors continue to pay fees for asset management services that are so high. It is hard to think of any other service that is priced at such a high proportion of value.

■ *I am deeply indebted to John Bogle, Michael Nolan, and Charles Ellis for help in the preparation of this paper. I am enormously grateful for extremely useful comments on the first draft of this paper from David Autor, Chang-Tai Hseih, Timothy Taylor, and Ulrike Malmendier. [Malkiel is a consultant to Vanguard and a director at Vanguard, Europe.]*

References

Bogle, John C. 2010a. *Common Sense on Mutual Funds: Fully Updated 10th Anniversary Edition.* Wiley.

Bogle, John C. 2010b. *Don't Count on It! Reflections on Investment Illusions, Capitalism, "Mutual" Funds, Indexing, Entrepreneurship, Idealism, and Heroes.* Wiley.

Coates, John C., and R. Glenn Hubbard. 2007. "Competition in the Mutual Fund Industry: Evidence and Implications for Policy." Harvard Law and Economics Discussion Paper 592. http://ssrn.com/abstract=1005426 or http://dx.doi.org/10.2139/ssrn.1005426.

Dyck, I. J. Alexander, and Lukasz Pomorski. 2011. "Is Bigger Better? Size and Performance in Pension Plan Management." Rotman School of Management Working Paper 1690724. http://dx.doi.org/10.2139/ssrn.1690724.

French, Kenneth R. 2008. "Presidential Address: The Cost of Active Investing." *Journal of Finance* 63(4): 1537–73.

Jensen, Michael C. 1968. "The Performance of Mutual Funds in the Period 1945–1964." *Journal of Finance* 23(2): 389–416.

Jenson, Michael C. 1975. "Is Financial Analysis Useless?" Proceedings of a Seminar on the Efficient Market Hypothesis, Financial Analysts Research Foundation. (Reprinted in *Handbook of Financial Economics*, edited by J. L. Bicksler, North-Holland, 1980.)

Kahneman, Daniel. 2011. *Thinking Fast and Slow.* New York: Farrar, Straus and Giroux.

Kahneman, Daniel, and Mark W. Riepe. 1998. "Aspects of Investor Psychology." *Journal of Portfolio Management* 24(4): 52–65.

Kimmel, Russ. 2012. "How Expense Ratios and Star Ratings Predict Success." http://news.morningstar.com/articlenet/article.aspx?id=347327.

Latzko, David A. 1999. "Economies of Scale in Mutual Fund Administration." *Journal of Financial Research* 22(3): 331–39.

Malkiel, Burton G. 1973 [2011]. *A Random Walk Down Wall Street.* W. W. Norton.

Malkiel, Burton G. 1995. "Returns from Investing in Equity Mutual Funds 1971–1991." *Journal of Finance* 50(2): 549–72.

Swensen, David F. 2000. *Pioneering Portfolio Management: An Unconventional Approach to Institutional Investment.* New York: The Free Press.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 109–132

Investing in Preschool Programs[†]

Greg J. Duncan and Katherine Magnuson

At the beginning of kindergarten, the math and reading achievement gaps between children in the bottom and top income quintiles amount to more than a full standard deviation. Early childhood education programs provide child care services and may facilitate the labor market careers of parents, but their greatest potential value is as a human capital investment in young children, particularly children from economically disadvantaged families (Heckman 2006). After all, both human and animal studies highlight the critical importance of experiences in the earliest years of life for establishing the brain architecture that will shape future cognitive, social, and emotional development, as well as physical and mental health (Sapolsky 2004; Knudsen, Heckman, Cameron, and Shonkoff 2006). Moreover, research on the malleability (plasticity) of cognitive abilities finds these skills to be highly responsive to environmental enrichment during the early childhood period (Nelson and Sheridan 2011). Perhaps early childhood education programs can be designed to provide the kinds of enrichment that low-income children most need to do well in school and succeed in the labor market.

We summarize the available evidence on the extent to which expenditures on early childhood education programs constitute worthy social investments in the human capital of children. We begin with a short overview of existing early childhood education programs, and then summarize results from a substantial body of methodologically sound evaluations of the impacts of early childhood education. We find that the evidence supports few unqualified conclusions. Many early childhood

■ *Greg J. Duncan is Distinguished Professor in the School of Education, University of California, Irvine. Katherine Magnuson is Associate Professor of Social Work, University of Wisconsin—Madison. Their email addresses are gduncan@uci.edu and kmagnuson@wisc.edu.*

[†] To access the Appendix and disclosure statements, visit http://dx.doi.org/10.1257/jep.27.2.109.

doi=10.1257/jep.27.2.109

education programs appear to boost cognitive ability and early school achievement in the short run. However, most of them show smaller impacts than those generated by the best-known programs, and their cognitive impacts largely disappear within a few years. Despite this fade-out, long-run follow-ups from a handful of well-known programs show lasting positive effects on such outcomes as greater educational attainment, higher earnings, and lower rates of crime. Since findings regarding short and longer-run impacts on "noncognitive" outcomes are mixed, it is uncertain what skills, behaviors, or developmental processes are particularly important in producing these longer-run impacts.

Our review also describes different models of human development used by social scientists, examines heterogeneous results across groups, and tries to identify the ingredients of early childhood education programs that are most likely to improve the performance of these programs. We use the terms "early childhood education" and "preschool" interchangeably to denote the subset of programs that provide group-based care in a center setting and offer some kind of developmental and educational focus. This definition is intentionally broad, as historical distinctions between early education and other kinds of center-based child care programs have blurred. Many early education programs now claim the dual goals of supporting working families and providing enriched learning environments to children, while many child care centers also foster early learning and development (Adams and Rohacek 2002).

Existing Preschool Programs

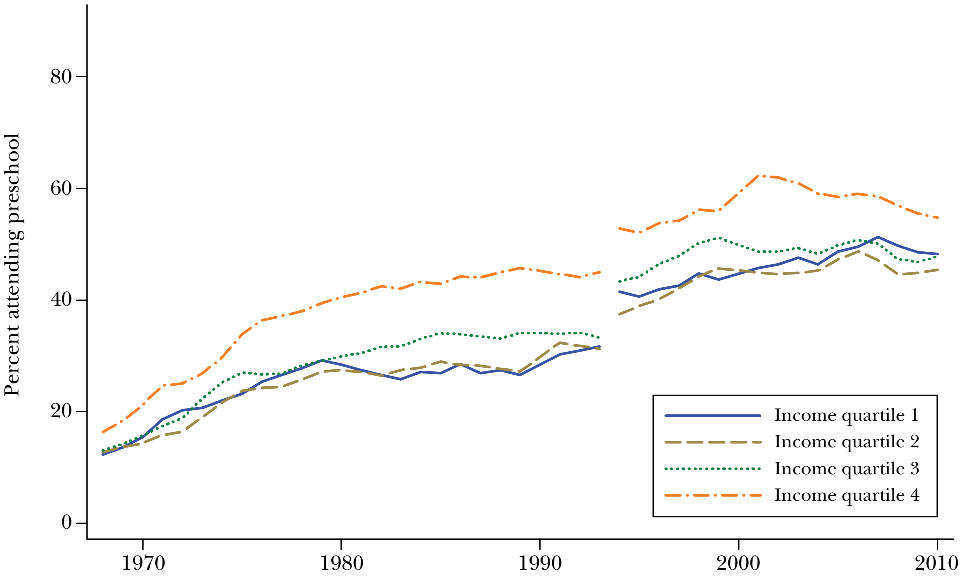
Most children enrolled in early childhood education attend private programs, some nonprofit and others for-profit. In 2011, the average cost of full-time, center-based care for a four-year old ranged from $3,900 in Mississippi to just over $14,000 in the District of Columbia (National Association of Child Care Resource and Referral Agencies 2012). Given the high cost of care, it is unsurprising that enrollment rates of children residing in families with incomes in the bottom half of the income distribution are persistently 10–20 percentage points lower than for children in the highest quarter. Figure 1, based on the data from the October Supplement to the Current Population Survey, shows this enrollment gap by income level. The figure also shows a steady rise in enrollment in early childhood education programs among three- and four-year-olds over the past 40-some years. This increase is broad-based, across income groups and for the children of both employed and nonemployed mothers.

States and the federal government have sought to increase the participation of low-income children in early childhood education programs in a number of ways: through Head Start, pre-kindergarten programs, and means-tested child care assistance programs that can be used to pay for center-based care.[1] Overall, both federal and state investments in these programs increased substantially in real terms

[1] The federal government also provides some financial assistance to families seeking child care via the Child and Dependent Care Tax Credit as well as exclusions from income for benefits under dependent

Figure 1

Percent of Three- and Four-year-olds Enrolled in Preschool by Family Income Quartile



Source: Authors using data from the October Current Population Survey.
Notes: Data represent three-year moving averages. Parents report on whether the child attends "regular school." The line break in 1994 corresponds to the addition of a question prompt, which defined regular school as including "nursery school, kindergarten or elementary school . . ." See Magnuson, Meyers, and Waldfogel (2007) for further discussion of how the Current Population Survey compares with other sources of data on preschool enrollment.

through the early 2000s, but in more recent years funding has not grown substantially (Barnett, Carolan, Fitzgerald, and Squires 2011; Magnuson and Shager 2010; Schulman and Blank 2012).

Head Start, the federal government's largest compensatory preschool program, is designed to enhance children's social and cognitive development by providing a comprehensive set of educational, health, nutritional, and other social services. In 2005, virtually all Head Start programs were center-based and half offered full-day (six hours or more) services, five days a week (Hamm 2006). Most children enrolled in Head Start in 2009 were three (36 percent) or four years old (51 percent). In 2010, the federal Head Start appropriation of about $7.2 billion was distributed to 1,591 local private and public nonprofit grantees serving 904,153 children. Some states supplement federal funds to increase access to Head Start programs; for details, see the Head Start website at http://eclkc.ohs.acf.hhs.gov/hslc/mr

care assistance programs; however, few low-income families benefit from these programs (Forry and Sorenson 2006; Magnuson, Meyers, and Waldfogel 2007).

/factsheets/fHeadStartProgr.htm. Local grantees are required to provide at least 20 percent matching funds. All this brings program costs to around $9,000 per child per year (Ludwig and Phillips 2007).

Pre-kindergarten programs are funded primarily by states or local school districts. In 2011, 39 states and the District of Columbia spent about $5.5 billion on pre-kindergarten initiatives that collectively served approximately 28 percent of the nation's four-year-olds and 4 percent of three-year-olds (for details, see Barnett, Carolan, Fitzgerald, and Squires 2011). Most pre-kindergarten programs target low-income children (31 state programs have income eligibility requirements), and most offer health, vision, and hearing screenings as well as at least one other form of support service. One-half of state pre-kindergarten programs require teachers to have training in early child development and nearly one-third require BA degrees. Typically, states use a mixed service delivery system that provides programming in local elementary schools as well as community-based settings.

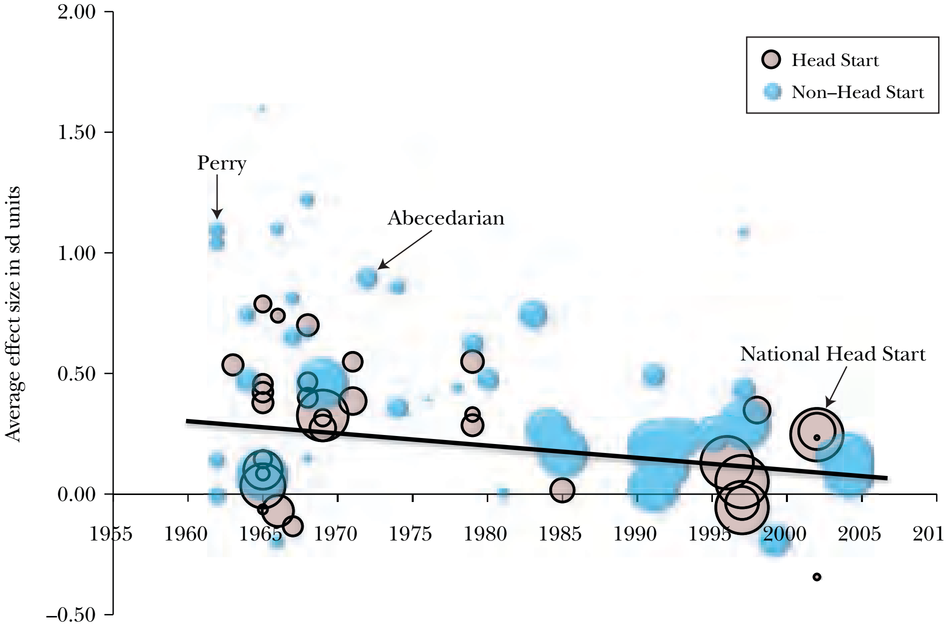
With expenditures in 2010 amounting to approximately $9.5 billion, federal and state-funded *means-tested child care subsidies* can be used for various types of child care, including center-based care, family day care, and other forms of informal care, and they cover a wide age range of children (birth through age 12). Their primary goal has continued to be supporting working families rather than educating young children, although increased spending on subsidies has been linked to higher rates of preschool attendance among young children (Magnuson, Meyers, and Waldfogel 2007). Because parents' preferences and needs for child care may not always align well with what is provided by preschool programs, and because child care subsidy spells are often quite short (Ha, Magnuson, and Ybarra 2012), these subsidies are best viewed as an indirect way to promote early childhood education for three- and four-year-olds.

Empirical Studies of the Effectiveness of Early Childhood Education

Empirical studies of the effects of investments in early childhood education on children's human capital encompass a range of methodologies and a wide variety of programs. We focus on evaluations of preschool programs conducted over the course of the last half-century that are based on strong experimental or quasi-experimental methods and provide impact estimates for cognitive or achievement-related outcomes.[2] Despite the hundreds of evaluation studies of early childhood education programs that have been published over the past 50 years,

[2] A full list of these studies appears in the online appendix available with this paper at http://ejep.org. As described there, programs selected for our analysis had both treatment and control/comparison groups, included at least 10 participants in each condition, incurred less than 50 percent attrition, and measured children's cognitive development close to the end of their "treatment" programs. Studies had to have used random assignment or one of the following quasi-experimental designs: change models, fixed effects modes, regression discontinuity, difference in difference, propensity score matching, interrupted time series, instrumental variables, and some other types of matching. Studies that used

Figure 2
Average Impact of Early Child Care Programs at End of Treatment
(standard deviation units)



Source: Authors.
Notes: Figure 2 shows the distribution of 84 program-average treatment effect sizes for cognitive and achievement outcomes, measured at the end of each program's treatment period, by the calendar year in which the program began. Reflecting their approximate contributions to weighted results, "bubble" sizes are proportional to the inverse of the squared standard error of the estimated program impact. There is a weighted regression line of effect size by calendar year.

a handful of programs have figured especially prominently in policy discussions: in particular, Perry Preschool, the Abecedarian[3] program, Head Start, and more recently some state and local pre-kindergarten programs.

Meta-Analysis

Figure 2 shows the distribution of 84 program-average treatment effect sizes for cognitive and achievement outcomes, measured at the end of each program's treatment period, by the calendar year in which the program began. Reflecting their approximate contributions to weighted results, "bubble" sizes are proportional to the inverse of the squared standard error of the estimated program impact. The figure differentiates between evaluations of Head Start and other early childhood

quasi-experimental designs must have had pre- and post-test information on the outcome or established baseline equivalence of groups on demographic characteristics determined by a joint test.

[3] "Abecedarian" can mean one who is learning the alphabet.

education programs and also includes a weighted regression line of effect size by calendar year.

Taken as a whole, the simple average effects size for early childhood education on cognitive and achievement scores was .35 standard deviations at the end of the treatment periods, an amount equal to nearly half of race differences in the kindergarten achievement gap (Duncan and Magnuson 2011). However, as can be seen from Figure 2, average effect sizes vary substantially and studies with the largest effect sizes tended to have the fewest subjects. When weighted by the inverse of the squared standard errors of the estimates, the average drops to .21 standard deviations.

All of the 84 programs that generated the effect size data shown in Figure 2 met minimum standards for quality of research methods. However, some of the programs lasted for only a couple of summer months, while others ran for as long as five years. Some of the evaluations used random assignment while others relied on less-rigorous quasi-experimental methods. Almost all focused on children from low-income families, but they varied in the racial and ethnic composition of treatment groups.

One might assume that these differences would account for much of the effect-size variability observed in Figure 2. However, that is not always the case. Weighted average effect sizes were insignificantly different between evaluations that did (.25 standard deviations) and did not (.19 standard deviations) use random assignment; and between those that were (.31 standard deviations) and were not (.18 standard deviations) published in peer-review journals. The effect sizes of programs designed by researchers (.39 standard deviations) were significantly larger than programs not designed by researchers (.18 standard deviations).

Programs beginning before 1980 produced significantly larger effect sizes (.33 standard deviations) than those that began later (.16 standard deviations). Declining effect sizes over time are disappointing, as we might hope that lessons from prior evaluations and advances in the science of child development would have led to an increase in program effects over time. However, the likely reason for the decline is that counterfactual conditions for children in the control groups in these studies have improved substantially. We have already seen in Figure 1 how much more likely low-income children are to be attending some form of center-based care now relative to 40 years ago. This matters because, though center-based care programs have varying degrees of educational focus, most research suggests that center-based care is associated with better cognitive and achievement outcomes for preschool age children (NICHD Early Childcare Research Network and Duncan 2003).

Even more impressive are gains in the likely quality of the home environment provided by low-income mothers, as indexed by their completed schooling. In 1970, some 71 percent of preschool age children in the bottom 20 percent of the income distribution had mothers who lacked a high school degree, while only 5 percent of the mothers had attended at least some postsecondary schooling (based on authors' calculation of the October Current Population Survey data). By 2000, the corresponding percentage of children with mothers who did not have a high school degree had dropped by nearly half (to 37 percent), while the percentage with mothers who had completed some postsecondary schooling increased five-fold (to just over 25 percent). Today, therefore, children from low-income households are likely to be benefiting from much higher-quality home environments than their counterparts four decades ago. Both higher-quality home environments and increases in other forms of center-based child care raise the bar for impact estimates coming from early childhood education programs.

Two particularly salient features of early childhood education programs are duration and starting age. Abundant literature suggests that the number of years spent in K–12 or postsecondary education is linked to labor market success (Card 1999). Thus, it seems plausible to expect that longer exposure to early childhood education environments before school entry should boost later academic achievement as well. But while simple associations indicate that longer participation in a preschool program generates larger treatment effects, models with a full set of controls for program and evaluation quality yield only small and statistically insignificant associations (+.04 standard deviations per additional year) between program duration and magnitudes of impacts (Leak, Duncan, Li, Magnuson, Schindler, and Yoshikawa 2012). The absence of larger effects for longer-duration programs may be due to the failure of such programs to use curricula and activities that capitalize fully on the skills gained in the early years of program participation.

As for starting age, neuroscience evidence on the plasticity of cognitive and language abilities suggests that these skills are highly amenable to environmental enrichment during the early childhood period. Starting in infancy, responsive caregiving and language-rich interactions are associated with better developmental outcomes, and more specifically stronger early language development (Tamis-LeMonda, Bornstein, and Baumwell 2001). Based on such findings, we might expect to find an "earlier is better" pattern of effects for early childhood education programs that provide such high-quality interactions for children. Evidence from the best-known early-life preschool programs is mixed: programs such as Early Head Start produce very small impacts on cognitive development (Love et al. 2003), whereas others, like the Abecedarian program, show much larger impacts (Ramey and Campbell 1984). Analysis of the meta-analytic database shows that, taken as a whole, effect sizes were neither larger nor smaller for children who started programs at younger ages (Leak, Duncan, Li, Magnuson, Schindler, and Yoshikawa 2012). This suggests that other modes of early childhood investments—for example, home visitation for high-risk, first-time mothers (Olds, Sadler, and Kitzman 2007) or developmental screenings and interventions for children living in families with documented domestic violence—may be more-effective ways of building children's capacities during the very early years of life.

Model Program Impacts: Perry Preschool and Abecedarian

As shown in Figure 2, average end-of-treatment effect sizes for the Perry Preschool and Abecedarian programs are several times larger than the weighted mean effect size for all studies in the meta-analytic database that met our inclusion criteria. A key reason for the prominence of these two studies and a few others is that long-term follow-ups show strikingly positive impacts in adulthood and impressive benefit–cost ratios.

Perry provided one or two years of part-day educational services and weekly home visits to 58 low-income, low-IQ, African American children aged three and four in Ypsilanti, Michigan, during the 1960s. The curriculum was geared to the children's age and capabilities, emphasizing child-initiated learning activities. Staff encouraged children to engage in play activities that would promote their problem-solving skills as well as their intellectual, social, and physical development. Program staff made weekly one- to two-hour afternoon visits to each family. The center's child-to-teacher ratio was low; each of four teachers served only 20–25 children every year. Per-pupil costs amounted to about $20,000 per child (in 2011 dollars). While Perry's large impacts on IQ at the point of school entry had all but disappeared by third grade (Schweinhart, Montie, Xiang, Barnett, Belfield, and Nores 2005), the program produced lasting improvements through age 40 on employment rates and substantially reduced the likelihood that participants had been arrested. Heckman, Moon, Pinto, Savelyev, and Yavitz (2010) estimate that the program generated about $152,000 in benefits over the life course, boosting individuals' earnings, reducing use of welfare programs, and, most importantly for the benefit calculation, reducing criminal activity. These financial benefits produced a social rate of return between 7 and 10 percent.

The Abecedarian program, which served 57 low-income, mostly African American families from Chapel Hill, North Carolina, provided even more-intensive services than Perry Preschool. Beginning in 1972, children assigned to the Abecedarian "treatment" received year-round, full-time center-based care for five years, starting in the child's first year of life. The Abecedarian preschool program included transportation, individualized educational activities that changed as the children grew older, and low child–teacher ratios of 3:1 for the youngest children and up to 6:1 for older children. Abecedarian teachers followed a curriculum that focused on language development and explained to teachers the importance of each task as well as how to teach it. High-quality health care, additional social services, and nutritional supplements were also provided to participating families (Ramey and Campbell 1979; Campbell, Ramey, Pungello, Sparkling, and Miller-Johnson 2002).

At two years of age, the control-group children in the Abecedarian program had IQ scores that averaged about one standard deviation below the mean, as would be expected for children from very economically disadvantaged backgrounds (Ramey, Campbell, Burchinal, Skinner, Gardner, and Ramey 2000). By the time the children reached age five, however, their IQ scores were close to the national average, and 10 points higher than scores of comparable children who did not participate in the program. Similarly large effects were observed for achievement on verbal and quantitative tests (Ramey and Campbell 1984). Nearly 15 years later, the program's effect on IQ scores at age 21 (.38 standard deviations) was still substantial but smaller than at age five. Children in the Abecedarian program entered college at 2.5 times the rate of children in the control group, and the intervention also reduced rates of teen parenthood and marijuana use by nearly half, although it did not lead to statistically significant reductions in criminal activity. Expressed in 2011 dollars, the costs associated with Abecedarian's five-year duration totaled about $80,000 per child, and the program is estimated to have produced $160,000 in net present benefits for its participants and their parents (Barnett and Masse 2007; Currie 2001).

It is difficult to extract policy lessons from these two initiatives for early childhood education programs that states or the federal government might offer today. Both programs were designed and evaluated by researchers and each served only several dozen children—conditions that scaled-up programs cannot match. Moreover, as we have pointed out above, counterfactual conditions three decades ago were likely of a comparatively low quality. The average number of years of maternal education completed was about 10 years for both the Perry and Abecedarian preschool treatment groups, reflecting the low levels of parental education among low-income families at that time.

Head Start Impacts

Large-scale policy lessons might be gleaned more reliably from studies of Head Start, since that program now provides services to almost a million three- and four-year-olds. Early quasi-experimental evaluations of Head Start found significant short-term gains in participants' achievement test scores, but as with Perry and Abecedarian, these achievement gains appeared to fade over time (Cicirelli 1969; McKey, Condelli, Ganson, Barrett, McConkey, and Plantz 1985). Despite methodological critiques of these early studies (McGroder 1990), a random-assignment national study of Head Start was not undertaken for another 30 years.

Begun in 2002, the Head Start Impact Study (HSIS) used wait-list lotteries to assign children to the opportunity to enroll in a Head Start program. Results indicated that after one academic year in the program, four-year-olds who had the opportunity to enroll in Head Start gained significantly more in six language and literacy areas than control-group children who lost the enrollment lotteries, with these intent-to-treat effects (effects for the group of children who had the opportunity to enroll) ranging from .09 to .31 standard deviations (US Department of Health and Human Services 2005). In contrast, there were few program impacts on math skills or on children's attention, anti-social, or mental health problems. The official report of the Head Start Impact Study (US Department of Health and Human Services 2005) provides estimates of differences between (parents of) children offered and children not offered a chance to get into the Head Start center with the waitlist lottery. Some children offered the chance didn't take it, and some children not offered a slot ended up in other Head Start centers. Ludwig and Phillips (2007) make the proper "treatment on the treated" estimate in light of this noncompliance, and the resulting effect sizes were roughly 50 percent larger than intent-to-treat effect sizes. By the end of first grade, both achievement levels and behavioral ratings of treatment group children were essentially similar to achievement levels of control-group children (US Department of Health and Human Services 2010).

Why might Head Start's initial achievement impacts disappear so quickly? All children learn, but they learn at different rates. If the test scores of Head Start and comparison-group children converge during elementary school, then the treatment group's preschool gains must be offset later by larger gains in the control group. Why this happens is not entirely clear; most arguments focus on the quality of subsequent schools that children attend. If little learning occurs in low-quality schools, then early advantages imparted by programs such as Head Start might be lost. In this case, preschool does not "immunize" against the adverse effects of subsequent low-quality schooling (Currie and Thomas 2000; Lee and Loeb 1995).

Currie and Thomas (2000) showed that Head Start impacts fade out more rapidly for African-American children than for white children; in examining why, they show that African-American children in Head Start attend lower-quality schools, as measured by students' average test scores, relative to the schools attended by African-American children who did not attend Head Start. In contrast, for white children, average school quality did not differ by Head Start participation status. Similarly, Zhai, Raver, and Jones (2012) find that the benefits to children of an intervention designed to enhance the developmental quality of Head Start programs persisted into kindergarten only for those children who attended relatively higher-quality elementary schools, again measured by student test scores.

An alternative explanation of achievement-impact fadeout is that kindergarten teachers might be particularly effective at teaching children with low levels of skills. In this case, it may be that the classroom is not of generally low quality, but instructional efforts may favor children at the lower end of the skill distribution, which would include larger concentrations of children who had not participated in early childhood education. Indirect evidence supporting this hypothesis is provided in the work of Engel, Claessens, and Finch (forthcoming), who find that kindergarten teachers spend the most time on very basic math instruction (like learning numbers) despite the fact that the vast majority of kindergarteners have already acquired such skills. If this explanation holds, the effects of early childhood education programs are most likely to persist in subsequent schooling environments in which learning gains are equally distributed across children with high and low levels of initial skills.[4]

As with Perry and Abecedarian program findings, quickly declining test score impacts for recent cohorts of Head Start children appear to be at odds with the long-term impacts on important young adult outcomes found in analyses of older Head Start cohorts. Some of the older-cohort studies use strong quasi-experimental methods and find quite striking long-run program impacts. One of the most recent and comprehensive is Deming's (2009) sibling-based fixed-effect analysis, which found that, compared with siblings who did not attend Head Start or other preschool programs, children who attended Head Start in the 1980s and early 1990s

[4] A third explanation would be that program impacts do not persist because early elementary instruction is most beneficial to children who enter school with high levels of initial skills and that Head Start program impacts are not sufficiently large to get children to a point at which they will benefit from such instruction. There does not seem to be good evidence to support this conjecture.

were over 8 percentage points more likely to graduate from high school. Deming's more-general composite of positive early adult outcomes—including high school graduation, college attendance, idleness, crime, teen parenthood, and health status—shows an estimated impact of .23 standard deviations.

Ludwig and Miller's (2007) regression discontinuity study of Head Start attendees in the late 1960s found that successful efforts to increase the likelihood that poor counties would establish Head Start programs by providing federal grant-writing assistance led to gains of 3–4 percentage points in high school graduation rates and postsecondary schooling in the 1990 census data relative to counties with very similar levels of poverty that were not offered such assistance, although such effects were attenuated by 2000. Taken together, these studies suggest that despite the decline in program impacts on achievement test scores as children progress through elementary school, there may be measurable and important effects of Head Start on children's life chances.

Pre-Kindergarten Programs

Some rigorous evaluations of pre-kindergarten programs were completed too recently to have been included in the database used to produce Figure 2. Most of these studies use regression discontinuity designs based on strict birthday cutoffs. Test-based assessments are given to children who just started attending pre-kindergarten and those who just completed it. The tests of children who just completed the program are compared with those about to attend. Children whose parents are not interested in enrolling them in the program are not part of either group. For this reason (and a few others), these designs are not directly comparable to either intent-to-treat or treatment-on-the-treated estimates from experimental studies (Lipsey, Weiland, Yoshikawa, Wilson, and Hofer 2011; Gibbs, Ludwig, and Miller 2011). The most comprehensive overview is Wong, Cook, Barnett, and Jung (2008), which examines five state pre-kindergarten programs and finds short-run effects on achievement test scores that are somewhat larger than those estimated in the National Head Start Impact Study, although the size of the impacts varies considerably across states and types of test (weighted average intent-to-treat impacts range from .17 for vocabulary to .68 for "print awareness").

The highly regarded Tulsa pre-kindergarten program has also been carefully evaluated. A birthday cutoff-based regression discontinuity evaluation of the program found large and significant effects on children's achievement, with effect sizes ranging from .38 to .79 (Gormley, Gayer, Phillips, and Dawson 2005). Adjusting for differences in children's backgrounds (using propensity score matching methods), the researchers found that the Tulsa pre-kindergarten program reduced attendees' timidity and improved their attentiveness. The program did not appear to affect disobedience, apathy, aggression, learning task problems, or problems interacting with peers or teachers (Gormley, Phillips, Newmark, Welti, and Adelstein 2011).

The only longer-run follow-up study conducted to date of pre-kindergarten program uses propensity matching and administrative data on third grade test scores. Hill, Gormley, and Adelstein (2012) estimating program impacts for two cohorts. They find no lasting discernible achievement impacts for the first cohort by third grade. For the second cohort there is evidence of persisting math impacts (.18 standard deviations), perhaps reflecting an increased emphasis on math instruction, including the introduction of new curricula, during elementary school. The lack of longer-run evaluations of pre-kindergarten programs suggests that drawing strong policy conclusions about their effectiveness is unwarranted, as other programs have likewise demonstrated early promising results that faded over the first few years of school.

The Puzzle: Academic Fade-Out, but Long-Term Benefits

Most early childhood education studies that have tracked children beyond the end of the program treatment find that effects on test scores fade over time. An analysis of cognitive and achievement outcomes in our meta-analytic database, which includes model programs such as Perry Preschool as well as Head Start and many other programs, shows an estimated decrease in program impact effect sizes of about .03 standard deviations per year. With end-of-treatment effect sizes averaging around .30 standard deviations, this implies that positive effects persist for roughly 10 years (Leak et al. 2011; see also Aos, Lieb, Mayfield, Miller, and Pennucci 2004; Camilli, Vargas, Ryan, and Barnett 2010). This finding raises a puzzle: How do we reconcile the fade-out of preschool program impacts on test scores during elementary school with the evidence showing that such programs nonetheless have beneficial impacts on a broad set of later-life outcomes like high school graduation rates, teen parenthood, and criminality?

One obvious possible explanation is that preschool programs may affect something other than basic achievement and cognitive test scores, and perhaps these other program impacts, unlike achievement and cognitive impacts, persist over time. In turn, this raises the question of exactly how early childhood education programs affect various aspects of development, including cognitive skills, personality traits like conscientiousness, and the behavior categories like attentiveness or antisocial behavior that are often emphasized by development psychologists. The literature on the effects of preschool has drawn on several different models of human development.

In one prominent example, Cunha and Heckman (2007) posit a cumulative model of the production of human capital that allows for the possibility of differing childhood investment stages as well as roles for the past effects and future development of both cognitive and socio-emotional skills. In this model, children have endowments at birth of cognitive potential and temperament that reflect a combination of genetic and prenatal environmental influences. The Cunha and Heckman model highlights the interactive nature of skill building and investments from families, preschools and schools, and other agents. It suggests that human capital accumulation results from "self-productivity"—skills developed in earlier stages bolster the development of skills in later stages—as well as the dynamic complementary that results when skills acquired prior to a given investment increase the productivity of that investment. These two principles are combined in the hypothesis that "skill begets skill."

Several aspects of this model are relevant for preschool investment policy. If focused on the preschool period, the Cunha and Heckman (2007) model implies that school readiness is a product of the child's cognitive and socio-emotional skills upon entry into the preschool period, plus preschool-period investments from parents and possibly from an early childhood education program. The hypothesis of dynamic complementarity implies that the effects of parental and early childhood education investments on child outcomes will be largest for children who enter the preschool period with the highest levels of cognitive and socio-emotional skills.

Predictions emerging from the models of human capital development proposed in the developmental psychology literature are different. These models, too, focus on how individuals' endowments interact with environmental experiences, and suggest that both individual capacities and experience shape development (Blair and Raver 2012). However, they diverge from the Cunha and Heckman (2007) model by distinguishing how environments and different types of investments (for example, parent and early-childhood-education investments) interact to shape development. Developmental models say that certain kinds of programs may be most productive for higher-skilled children while others are geared towards helping bring up the skills of low-skill children and don't match well to the needs of higher-skill children. For example, Ramey and Ramey's "compensatory model" (1998) posits that preschool investments can function as a substitute for enriched home environments. Thus, children whose skill development may be compromised by economic disadvantage or low-quality home environments are predicted to benefit more from early childhood education programs than more-advantaged children. This hypothesis provided the rationale for the initial and continued funding for programs such as Head Start and Early Head Start, which target children from disadvantaged backgrounds.

If early childhood education programs seek to build children's early skills to generate lasting changes in adults' human capital, which skills should they target? Economists tend to lump IQ and achievement into a "cognitive" category and everything else into a "noncognitive" category, but this distinction is unhelpful for a variety of reasons. First, "cognitive" skills are a heterogeneous mixture of "achievement" and more-basic cognitive capacities. Although scores on tests of cognitive ability and achievement tend to be highly correlated, there is an important conceptual difference between them. "Achievement" commonly refers to concrete academic skills such as literacy and numeracy that develop in response to parenting, schooling, and other human capital investments, including early childhood education, whereas IQ or general cognitive ability is considered to be a relatively more-stable trait. Second, learning skills such as the ability to sustain attention when performing tasks, plan ahead, and control emotions in the face of provocation involve many of the same elements of brain circuitry as learning concrete skills, and are therefore inherently "cognitive." Third and most important, different branches of psychology typically categorized noncognitive skills in very different ways. Conceptualizing and measuring distinct components of "noncognitive" skills is a vital first step in understanding why early childhood education and other human capital inventions have an effect.

Most personality psychologists have centered their work on the "big five" personality traits, which are derived from factor analyses of observer- and self-reports of behaviors and include conscientiousness, openness, agreeableness, emotional stability, and extraversion—plus general cognitive ability. Education research consistently shows that conscientiousness best correlates with overall attainment and achievement (Almlund, Duckworth, Heckman, and Kautz 2011). Although these traits have traditionally been viewed as relatively stable across the lifespan, some evidence indicates that they can change in response to life experiences and interventions (for example, Roberts, Walton, and Viechtbauer 2006; Almlund et al. 2011).

Developmental psychologists view children's skills and behaviors as determined by the interplay between their innate abilities, their dispositions, and the quality of their early experiences—which may include early childhood education (Committee on Integrating the Science of Early Childhood Development, 2000). They classify skills and behaviors in a number of ways, and some of their categories correspond to the "big five" personality traits. For example, our own recent review classified important competencies into four groups: achievement, attention, "externalizing behavior" problems, and mental health (Duncan and Magnuson 2011). Attention refers to the ability to control impulses and focus on tasks (for example, Raver 2004). "Externalizing behavior" refers to a cluster of related behaviors including antisocial behavior, conduct disorders, and more-general aggression (Campbell, Shaw, and Gilliom 2000). Mental health constructs include anxiety and depression as well as somatic complaints and withdrawn behavior (Bongers, Koot, van der Ende, and Verhulst 2003). All of these skills and behaviors might respond to investments in early childhood education.

Testing and comparing how these theories of human development apply in the context of early childhood education is difficult, because despite arguments that early childhood education programs are likely to generate broad impacts on children's behavior and social competence (Zigler and Trickett 1978), most preschool studies do not measure many of these kinds of outcomes at program completion. Some studies have included measures of problem behavior, typically ratings of children's antisocial or aggressive behaviors, with mixed results. Perry significantly reduced problem behavior, especially among boys, and the examination by Heckman, Moon, Pinto, Savelyev, and Yavitz (2010) of Perry's long-run effects finds that these behavior impacts explain a substantial proportion of the program's effects on boys' crime and employment outcomes. However, both early cognitive and behavioral impacts explain program impacts on girls' later outcomes. Moreover, for both genders a substantial share of the program impacts on adult outcomes is not explained by any of the observed early program impacts.

Other programs provide little evidence of program impacts on children's behavior. Deming's (2009) analysis of Head Start found no short-run effects of Head

Start on parental reports of children's behavior problems. Haskins (1985) reported that the Abecedarian program had the unexpected effect of increasing teacher reports of children's aggressiveness in the early school years, although these effects appeared to fade with time. Of course, these studies are vulnerable to the criticism that they did not measure a broader set of relevant skills, including student's attention or other aspects of their behavior and mental health.

Overall, reconciling disparate patterns of impacts in the short and longer term is a key challenge for anyone hoping to extract policy lessons about the effectiveness of early childhood education programs. Accomplishing this task will require a proven model of human development that incorporates various cognitive, personality, and behavioral dimensions and can predict what kinds of children stand to benefit most from early childhood education investments.

Within-Program Heterogeneity

Although policymakers appropriately care most about the average impacts of early childhood education programs, a number of lessons can be learned from looking at the distribution of treatment effects of given programs. For example, such heterogeneity might make it possible to identify groups that could particularly benefit from preschool programs. Data on treatment-effect heterogeneity may also boost our understanding of human capital development processes if they identify groups that particularly benefit from the preschool setting.

Consider evidence from the Infant Health and Development Program (IHDP), shown in Figure 3. Beginning shortly after a child's birth, the IHDP offered a package of services that included a full-day, cognitively enriching curriculum for children between ages one and three, modeled after the Abecedarian program. Nearly 1,000 children in eight sites across the country were randomly assigned to the IHDP treatment or to a control group that received no early childhood education services but some health services (Gross, Spiker, and Haynes 1997). To be eligible for the program, infants had to have weighed less than 2,500 grams (5.5 pounds) at birth, but eligibility was not restricted by family income, race, or ethnicity.

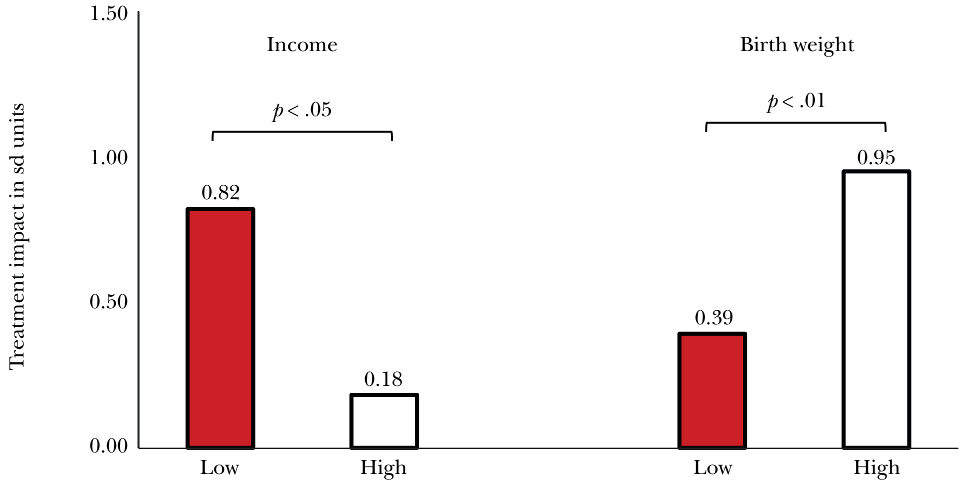
For the economically disadvantaged children in the sample—those with family income below 180 percent of the poverty line in their first year of life—participation in the Infant Health and Development Program (IHDP) produced large impacts on cognitive development. Specifically, children in the treatment group outscored their control-group counterparts by .82 standard deviations on the Stanford–Binet IQ mental subscale by age three.[5] For children in higher-income families, the IHDP's program impact was much smaller, only .18 standard deviations. Thus, if "disadvantage" is defined by family income, IHDP treatment impacts heavily favored

[5] This estimate comes from Duncan and Sojourner (forthcoming) and is based on weights designed to match the demographic characteristics of the Infant Health and Development Program sample to those of all US births.

Figure 3

Impacts of the Infant Health and Development Program on Age-3 IQ, by Income and Birth Weight

(standard deviation units)



Source: Authors.

Notes: The figure shows the impact, in standard deviation units, of the Infant Health and Development Program treatment on Age-3 IQ, for lower- and higher-income children and for lower- and higher-birth-weight children in the program. All models also condition on child gender, birth weight, gestational age at birth, neonatal health index, and site indicators.

disadvantaged infants. However, an alternative definition of "disadvantage" can lead to a different conclusion. Children disadvantaged by being born with a "very low" birth weight (less than 1500 grams or 3.3 pounds) benefited significantly less from the IHDP intervention than "advantaged" heavier babies in this low-birth-weight sample.

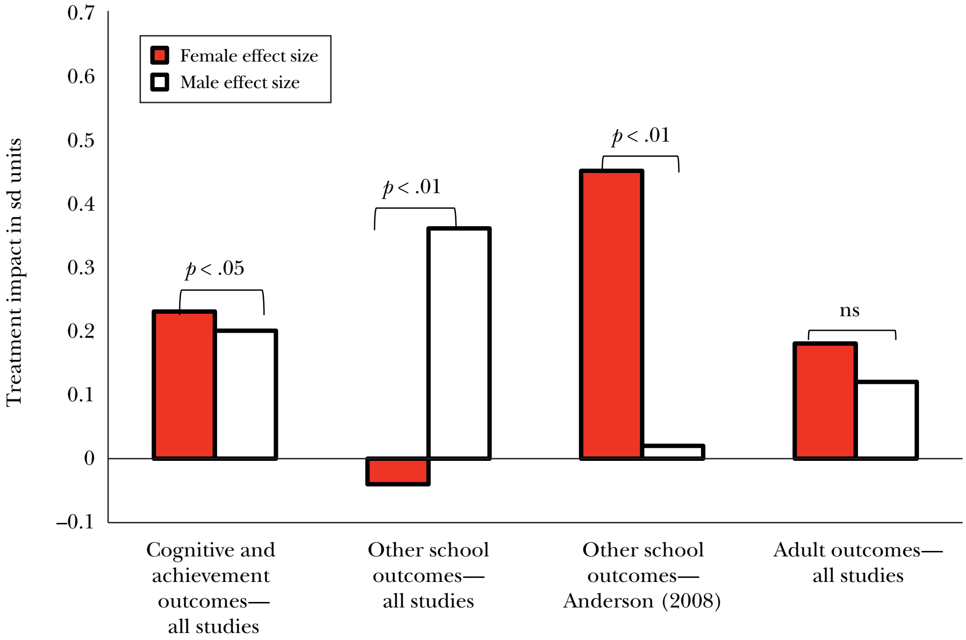
It is not difficult to generate possible explanations for these patterns. For example, the income results are consistent with theories positing that the focus of the Infant Health and Development Program on enriched early learning compensates or substitutes for lower levels of parental investment and academic stimulation in low-income families. The differences by birth weight are consistent with the "skill begets skill" perspective. Potential gains for very low birth weight babies' cognitive development may be constrained by neurological challenges that the program was unable to address. In other words, the match between what the program provided and children's individual differences may explain why some disadvantaged groups show larger effects, but not others.

A systematic accounting of heterogeneity in the effects of preschool programs is a complicated undertaking. For example, Anderson's evaluations of three researcher-designed early childhood education programs—Perry, Abecedarian, and the Early Training Project—described in Anderson (2008), showed much larger benefits for girls than boys. Turning to our meta-analytic database, we found

Figure 4

Gender Differences in Early Childhood Education Impacts

(standard deviation units)



Source: Adapted from Kelchen, Magnuson, Duncan, Schindler, Shager, and Yosikawa (2012), figure 2.
Note: This figure looks at outcomes by gender for the three programs evaluated in Anderson (2008)—Perry, Abecedarian, and the Early Training Project—and for a group of 22 programs that included the three programs evaluated in Anderson (2008) plus 19 other programs that estimated program impacts by gender.

19 other programs that estimated program impacts by gender. Evaluations of these programs do not show consistently larger effects for girls. The first bar in Figure 4 (which is adapted from Kelchen, Magnuson, Duncan, Schindler, Shager, and Yoshikawa 2012) shows that on cognitive and achievement outcomes, the average effect across all 22 studies is slightly larger for females. However, the second bar shows that when a broad set of school outcomes are considered, including special education, grade retention, and other aspects of general school adjustment, boys appear to benefit much *more* from these programs than girls. Looking just at the three programs in Anderson (2008) (the third bar of Figure 4), the "other school outcomes" variable strongly favors females, so the difference in findings is generated by the inclusion of results from a broader set of studies. For the adult outcomes across all studies (fourth bar), females are favored, but the difference is not significant.

Even when studies determine that a particular program has been a success on average, overall, the positive outcomes differ across programs and populations. For example, Perry Preschool and Head Start significantly reduced criminal activity, but

Abecedarian did not. Garces, Thomas, and Currie (2002) found that Head Start increased educational attainment for whites, but not for blacks, and led to reductions in crime for blacks but not whites.

There is much more to be learned about heterogeneity in the effects of preschool programs, although efforts to identify differential effects can be hampered by small sample sizes and limited baseline information, especially in the older studies. The program and population specificity of program impacts argues against a single explanation for how preschool programs improve long-run outcomes. Greater attention should be given to understanding both who benefits the most from particular programs and why.

The Search for Active Program Ingredients

Research on early childhood education has focused greater attention on evaluating particular programs than on identifying the particular ingredients in these programs that produce significant improvements in children's learning and behavior. The research problem here is difficult. For example, some scholars have focused on structural aspects of early childhood education environments, such as class size and teacher education, yet these features of programs are likely to affect children only indirectly, by influencing their experiences within classrooms. Perhaps not surprisingly, associations between these features of classrooms and preschoolers' learning are inconsistent and weak (Mashburn et al. 2008).[6]

Much harder to measure than class size or teacher education, but potentially more important for children's actual experiences in early childhood education programs, is what developmental psychologists have referred to as "process quality"—the quality of classroom interactions, including the amount of instructional and emotional support children receive. Associations between these aspects of process quality and children's outcomes are more consistently positive, if still modest (Burchinal, Kainz, and Cai forthcoming). As attention has shifted to improving classroom interactions, two aspects of program design emerge as policy levers that may, together, improve program effectiveness: curriculum and related professional development. To cite one example, best practices for mathematics instruction explicitly incorporate foundational math conceptual learning within everyday activities and provide activities that support a developmental progression of mathematical learning (Clements and Samara 2011). Despite the identification of best practices and the availability of curricula that provide lesson plans, research consistently finds that the instructional quality of most preschool classrooms is poor (Justice, Mashburn, Hamre, and Pianta 2008).

[6] None of these studies is based on random assignment of children to different preschool class sizes, nor do any conduct long-run follow-ups. Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) find noteworthy longer-run impacts of assignment to smaller kindergarten-to-grade-3 classrooms in the Project Star data.

It appears that an effective strategy is to combine a proven curriculum that offers well-designed lesson plans and activities, based on an understanding of children's trajectories of learning within specific content areas, with strong professional development to target improvement in specific instructional practices. Several random-assignment studies of curricular innovations in early childhood education programs have shown substantial effects on children's learning in math and literacy, and these curricula are currently found in some effective preschool programs. The What Works Clearinghouse provides up-to-date information on rigorous evaluations of early childhood education curricula (at http://ies.ed.gov/ncee/wwc/).

The Boston pre-kindergarten system provides a scaled-up model of how this might work. System leaders developed a curriculum from proven literacy, math, and social skills interventions. The academic components focused on concept development, the use of multiple methods and materials to promote children's learning, and a variety of activities to encourage analysis, reasoning, and problem-solving (Weiland and Yoshikawa forthcoming). Extensive professional development training and on-going coaching ensured that teachers understood the curriculum and were able to implement it effectively in their classrooms. A regression-discontinuity evaluation showed relatively large impacts on vocabulary, math, and reading (effect sizes ranging from .45 to .62 standard deviations) as well as smaller, but still noteworthy effects on working memory and inhibitory control (effect sizes ranging from .21 to .28 standard deviations; Weiland and Yoshikawa forthcoming).

Conclusions

Theories and evidence across the social sciences argue that early childhood may be a promising period for effective educational investments, particularly for disadvantaged children. Early cognitive and socio-emotional skills are sensitive to environmental inputs, and building skills early in life may produce lasting effects. Most evaluations of early education programs show that such programs improve children's school readiness, specifically their pre-academic skills, although the distribution of impact estimates is extremely wide, and gains on achievement tests typically fade over time. Some studies of children who attended preschool 20 or more years ago find that early childhood education programs also have lasting effects on children's later life chances, improving educational attainment and earnings and, in some cases, reducing criminal activity. High-quality early childhood education programs thus have the potential to generate benefits well in excess of costs. Despite general agreement about these aspects of early childhood education studies, important questions about the wisdom of large-scale investments in early childhood education remain unanswered.

First, we need to know much more about how early childhood education works: that is, the connections between program components and particular child outcomes. Because program impacts on cognitive ability and achievement often fade within a few years of the end of the programs, these skills do not appear to be driving longer-run effects. Data constraints have made it difficult to identify the other skills, behaviors, or developmental processes that lead to such positive outcomes in early adulthood, but efforts to better identify and measure likely pathways are critical for improving our understanding of human capital accumulation and judging whether policy and programmatic efforts are worthwhile investments. It also important to think about what programs (or parts of programs) might be scaled up in a cost-effective manner.

Second, we need a better understanding of the pattern of these program effects over time. This is likely to require new data collection efforts because administrative data about participation in these programs, demographic background, and scores on various tests are unlikely to provide necessary information on the full range of attention, behavior, and mental health measures.

Finally, we need a more complete understanding of which skills, or constellation of skills, are likely to produce improved outcomes later in life. This requires not only an understanding of how programs affect later skills, but also a better grasp of how skills, behavior, attention, and mental health in childhood build human capital and other labor market outcomes in adulthood.

Given the potential payoff from early education and the importance of early skills in forecasting later school and labor market success, supporting low-income children's participation in high-quality early childhood investment may well constitute a wise investment. The potential for profitable investments exists at both margins—enrolling low-income children who are not currently attending a preschool program as well as improving the quality of existing programs—although we know more about the former than the latter (Duncan, Ludwig, and Magnuson 2010). What may be more important in the long term than any specific programmatic change is a change in how research is conducted in this area. Rather than looking merely at average short-run outcomes of early childhood education programs based on a limited number of achievement tests, researchers should focus on the heterogeneity of outcomes across groups, conduct long-term follow-up, and examine a wide range of outcome variables that would illuminate the program ingredients and developmental processes that make some of these programs so successful.

■ *This paper draws extensively from our collaborations with Hirokazu Yoshikawa, Holly Schindler, and Jack Shonkoff. We are grateful to the following funders of the National Forum on Early Childhood Policy and Programs: the Birth to Five Policy Alliance, the Buffett Early Childhood Fund, Casey Family Programs, the McCormick Tribune Foundation, the Norlien Foundation, Harvard University, and an anonymous donor. We would also like to thank the Institute of Education Sciences (R305A110035) for supporting this research, to Abt Associates, Inc. and the National Institute for Early Education Research for making their data available to us, and to Weilin Li for helpful research assistance. The authors are also grateful for support from the NICHD-supported Irvine Network on Interventions in Development (HD065704 P01).*

References

Adams, Gina, and Monica Rohacek. 2002. "More Than a Work Support? Issues around Integrating Child Development Goals into the Child Care Subsidy System." *Early Childhood Research Quarterly* 17(4): 418–40.

Almlund, Mathilde, Angela L. Duckworth, James J. Heckman, and Tim Kautz. 2011. "Personality Psychology and Economics." IZA Discussion Paper 5500.

Anderson, Michael. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association* 103(484): 1481–95.

Aos, Steve, Roxanne Lieb, Jim Mayfield, Marna Miller, and Annie Pennucci. 2004. "Benefits and Costs of Prevention and Early Intervention Programs for Youth." September 12. Washington State Institute for Public Policy. http://www.wsipp.wa.gov/rptfiles/04-07-3901.pdf.

Barnett, W. Steven, Megan E. Carolan, Jen Fitzgerald, and James H. Squires. 2011. *The State of Preschool 2011: State Preschool Yearbook.* National Institute for Early Education Research, Rutgers.

Barnett, W. Steven, and Leonard N. Masse. 2007. "Comparative Cost–Benefit Analysis of the Abecedarian Program and Its Policy Implications." *Economics of Education Review* 26(1): 113–25.

Besharov, Douglas J., and Caeli A. Higney. 2007. "Head Start: Mend It, Don't Expand It (Yet)." *Journal of Policy Analysis and Management* 26(3): 678–681.

Blair, Clancy, and C. Cybele Raver. 2012. "Child Development in the Context of Adversity: Experiential Canalization of Brain and Behavior." *American Psychologist* 67(4): 309–318.

Bongers, Ilja L., Hans M. Koot, Jan van der Ende, and Frank C. Verhulst. 2003. "The Normative Development of Child and Adolescent Problem Behavior." *Journal of Abnormal Psychology* 112 (5): 179–92.

Burchinal, Margaret, Kirsten Kainz, and Karen Cai. Forthcoming. "How Well Do Our Measures of Quality Predict Child Outcomes? A Meta-Analysis and Coordinated Analysis of Data from Large-Scale Studies of Early Childhood Settings." In *Reasons to Take Stock and Strengthen our Measures of Quality*, edited by M. Zaslow. Baltimore, MD: Brooks Publishing.

Camilli, Gregory, Sadako Vargas, Sharon Ryan, and W. Steven Barnett. 2010. "Meta-Analysis of the Effects of Early Education Interventions on Cognitive and Social Development." *Teachers College Record* 112 (3): 579–620.

Campbell, Frances A., Craig T. Ramey, Elizabeth Pungello, Joseph Sparkling, and Shari Miller-Johnson. 2002. "Early Childhood Education: Young Adult Outcomes from the Abecedarian Project." *Applied Developmental Science* 6(1): 42–57.

Campbell, Susan B., Daniel S. Shaw, and Miles Gilliom. 2000. "Early Externalizing Behavior Problems: Toddlers and Preschoolers at Risk for Later Adjustment." *Development and Psychopathology* 12(3): 467–88.

Card, David E. 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, Vol. 3, edited by O. Ashenfelter, and D. Card, 1801–63. Amsterdam: North-Holland.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane W. Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593–1660.

Cicirelli, Victor G. 1969. *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development.* Report presented to the Office of Economic Opportunity pursuant to contract B89-4536 (report No. PB 184 328). Westinghouse Learning Corporation. Washington, DC: National Bureau of Standards, Institute for Applied Technology.

Clements, Douglas, and Julie Samara. 2011. "Early Childhood Mathematics Intervention." *Science* 333(6045): 968–70.

Committee on Integrating the Science of Early Childhood Development. 2000. *From Neurons to Neighborhoods: The Science of Early Childhood Development*, edited by Jack Shonkoff and Deborah Phillips. Washington, DC: National Academy Press.

Cunha, Flavio, and James J. Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97(2): 31–47.

Currie, Janet. 2001. "Early Childhood Education Programs." *Journal of Economic Perspectives* 15(2): 213–38.

Currie, Janet, and Duncan Thomas. 1995. "Does Head Start Make a Difference?" *American Economic Review* 85(3): 341–64.

Currie, Janet, and Duncan Thomas. 2000. "School Quality and the Longer-Term Effects of Head Start." *Journal of Human Resources* 35(4): 755–74.

Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3): 111–34.

Duncan, Greg, and Katherine Magnuson. 2011. "The Nature and Impact of Early Achievement

Skills, Attention Skills, and Behavior Problems." In *Whither Opportunity: Rising Inequality, Schools, and Children's Life Chances*, edited by Greg J. Duncan and Richard J. Murnane, 47–69. New York: Russell Sage.

Duncan, Greg, and Aaron Sojourner. Forthcoming. "Can Intensive Early Childhood Intervention Programs Eliminate Income-Based Cognitive and Achievement Gaps?" *Journal of Human Resources*.

Engel, Mimi, Amy Claessens, and Maida Finch. Forthcoming. "Teaching Students What They Already Know? The (Mis)alignment between Instructional Content in Mathematics and Student Knowledge in Kindergarten." *Education Evaluation and Policy Analysis*.

Forry, Nicole, and Elaine Sorenson. 2006. "The Child and Dependent Care Tax Credit: A Policy Analysis." *Marriage and Family Review* 39(1–2):159–76.

Garces, Eliana, Duncan Thomas, and Janet Currie. 2002. "Longer-Term Effects of Head Start." *American Economic Review* 92(4): 999–1012.

Gibbs, Chloe, Jens Ludwig, and Douglas L. Miller. 2011. "Does Head Start Do Any Lasting Good?" NBER Working Paper 17452.

Gormley, William T., Jr., Ted Gayer, Deborah Phillips, and Brittany Dawson. 2005. "The Effects of Universal Pre-K on Cognitive Development." *Developmental Psychology* 41(6): 872–84.

Gormley, William T., Deborah A. Phillips, Katie Newmark, Kate Welti, and Shirley Adelstein. 2011. "Social-Emotional Effects of Early Childhood Education Programs in Tulsa." *Child Development* 82(6): 2095–2109.

Gross, Ruth T., Donna Spiker, and Christine W. Haynes, eds. 1997. *Helping Low Birth Weight, Premature Babies: The Infant Health and Development Program*. Stanford, CA: Stanford University Press.

Ha, Yoonsook, Katherine Magnuson, and Marci Ybarra. 2012. "Patterns of Child Care Subsidy Receipt and the Stability of Child Care." *Children and Youth Services Review* 34(9): 1834–44.

Hamm, Katie. 2006. "More than Meets the Eye: Head Start Programs, Participants, Families, and Staff in 2005." Head Start Series, Brief No. 8, Center for Law and Social Policy, Washington, DC.

Haskins, Ron. 1985. "Public School Aggression among Children with Varying Day-Care Experience." *Child Development* 56(3): 689–703.

Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science*, June 30, 312(5782): 1900–1902.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010. "A New Cost–Benefit and Rate of Return Analysis for the Perry Preschool Program: A Summary." NBER Working Paper 16180.

Hill, Carolyn J., William T. Gormley, and Shirley Adelstein. 2012. "Do the Short-Term Effects of a Strong Preschool Program Persist?" Center for Research on Children in the United States Working Paper 18, Georgetown University.

Justice, Laura M., Andrew J. Mashburn, Bridget K. Hamre, and Robert C. Pianta. 2008. "Quality of Language and Literacy Instruction in Preschool Classrooms Serving At-Risk Pupils." *Early Childhood Research Quarterly* 23(1): 51–68.

Kelchen, Robert, Katherine Magnuson, Greg Duncan, Holly Schindler, Hilary Shager, and Hirokazu Yoshikawa. 2012. "Do the Effects of Early Childhood Programs on Academic and Adult Outcomes Vary by Gender? A Meta-Analysis." Unpublished paper.

Knudsen, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff. 2006. "Economic, Neurobiological and Behavioral Perspectives on Building America's Future Workforce." *Proceedings of the National Academy of Sciences of the United States of America* 103(27): 10155–62.

Leak, James, Greg Duncan, Weilin Li, Katherine Magnuson, Holly Schindler, and Hirokazu Yoshikawa. 2012. "Is Timing Everything? How Early Childhood Education Program Cognitive and Achievement Impacts Vary by Starting Age, Program Duration and Time since the End of the Program." Unpublished paper.

Lee, Valerie E., and Susanna Loeb. 1995. "Where Do Head Start Attendees End Up? One Reason Why Preschool Effects Fade Out." *Educational Evaluation and Policy Analysis* 17(1): 62–82.

Lipsey, Mark, Christina Weiland, Hirokazu Yoshikawa, Sandra Wilson, and Kerry Hofer. 2012. "The Prekindergarten Age-Cutoff Regression-Discontinuity Design: Methodological Issues and Implications for Application." Unpublished paper.

Love, John M. et al. 2003. "Child Care Quality Matters: How Conclusions May Vary with Context." *Child Development* 74(4): 1021–1033.

Ludwig, Jens, and Douglas L. Miller. 2007. "Does Head Start Improve Children's Life Chances: Evidence from a Regression Discontinuity Design." *Quarterly Journal of Economics* 122(1): 159–208.

Ludwig, Jens, and Deborah Phillips. 2007. "The Benefits and Costs of Head Start." *Social Policy Report* 21(3): 3–20.

Magnuson, Katherine A., Marcia K. Meyers, and Jane Waldfogel. 2007. "Public Funding and Enrollment in Formal Child Care in the 1990s." *Social Service Review* 81(1): 47–83.

Magnuson, Katherine, and Hilary Shager. 2010. "Early Education: Progress and Promise for Children from Low-income Families." *Children and Youth Services Review* 32(9): 1186–98.

Mashburn, Andrew, Robert Pianta, Bridget Hamre, Jason Downer, Oscar Barbarin, Donna Bryant, Margaret Burchinal, Diane Early, and Carolee Howes. 2008. "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills." *Child Development* 79(3): 732–49.

McGroder, Sharon M. 1990. "Head Start: What Do We Know About What Works?" Report prepared for the U. S. Office of Management and Budget. http://aspe.hhs.gov/daltcp/reports/headstar.pdf.

McKey, Ruth Hubbell, Larry Condelli, Harriet Ganson, Barbara J. Barrett, Catherine C. McConkey, and Margaret C. Plantz. 1985. *The Impact of Head Start on Children, Families and Communities: Final Report of the Head Start Evaluation, Synthesis and Utilization Project.* Washington, DC: CSR, Incorporated.

National Association of Child Care Resource and Referral Agencies. 2012. *Parents and the High Cost of Child Care: 2012 Report.* Arlington, VA: NACCRRA. http://www.naccrra.org/publications/naccrra-publications/2012/8/parents-and-the-high-cost-of-child-care-2012-report.

National Institute of Child Health and Human Development (NICHD) Early Childcare Research Network, and Greg J. Duncan. 2003. "Modeling the Impacts of Child Care Quality on Children's Preschool Cognitive Development." *Child Development* 74(5): 1454–75.

Nelson, Charles A., and Margaret A. Sheridan. 2011. "Lessons from Neuroscience Research for Understanding Causal Links between Family and Neighborhood Characteristics and Educational Outcomes." *Whither Opportunity: Rising Inequality, Schools, and Children's Life Chances,* edited by G. J. Duncan and R. J. Murnane, 27–46. New York: Russell Sage.

Olds, David L., Lois Sadler, and Harriet Kitzman. 2007. "Programs for Parents of Infants and Toddlers: Recent Evidence from Randomized Trials." *Journal of Child Psychology and Psychiatry* 48(3–4): 355–91.

Ramey, Craig T., and Frances A. Campbell. 1979. "Compensatory Education for Disadvantaged Children." *School Review* 87(2): 171–89.

Ramey, Craig T., and Francis A. Campbell. 1984. "Preventive Education for High-Risk Children: Cognitive Consequences of the Carolina Abecedarian Project." *American Journal of Mental Deficiency* 88(5): 515–23.

Ramey, Craig T., Frances A. Campbell, Margaret Burchinal, Martie L. Skinner, David M. Gardner, and Sharon L. Ramey. 2000. "Persistent Effects of Early Childhood Education on High-Risk Children and Their Mothers." *Applied Developmental Science* 4(1): 2–14.

Ramey, Craig T., and Sharon Landesman Ramey. 1998. "Early Intervention and Early Experience." *American Psychologist* 53(2): 109–120.

Raver, C. Cybele. 2004. "Placing Emotional Self-Regulation in Sociocultural and Socioeconomic Contexts." *Child Development* 75(2): 346–53.

Roberts, Brent, Kate Walton, and Wolfgang Viechtbauer. 2006. "Patterns of Mean-Level Change in Personality Traits across the Life Course: A Meta-Analysis of Longitudinal Studies." *Psychological Bulletin* 132(1): 1–25.

Sapolsky, Robert. 2004. "Mothering Style and Methylation." *Nature Neuroscience* 7(8): 791–92.

Schulman, Karen, and Helen Blank. 2012. "Downward Slide: State Child Care Assistance Policies 2012." National Women's Law Center. http://www.nwlc.org/sites/default/files/pdfs/NWLC2012_StateChildCareAssistanceReport.pdf.

Schweinhart, Lawrence J., Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Belfield, and Milagros Nores. 2005. *Lifetime Effects: The HighScope Perry Preschool Study through Age 40.* Monographs of the HighScope Educational Research Foundation, 14. Ypsilanti, MI: HighScope Press.

Tamis-LeMonda, Catherine S., Marc H. Bornstein, and Lisa Baumwell. 2001. "Maternal Responsiveness and Children's Achievement of Language Milestones." *Child Development* 72(3): 748–67.

US Department of Health and Human Services, Administration for Children and Families. 2005. *Head Start Impact Study: First Year Findings.* Washington, DC.

US Department of Health and Human Services, Administration for Children and Families. 2010. *Head Start Impact Study: Final Report.* Washington, DC.

Weiland, Christina, and Hirokazu Yoshikawa. Forthcoming. "Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills." *Child Development.*

Wong, Vivian, Thomas D. Cook, W. Steven Barnett, and Kuanghee Jung. 2008. "An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs." *Journal of Policy Analysis and Management* 27(1): 122–54.

Zhai, Fuhua , C. Cybele Raver, and Stephanie M. Jones. 2012. "Academic Performance of Subsequent Schools and Impacts of Early Interventions: Evidence from a Randomized Controlled Trial in Head Start Settings." *Children and Youth Services Review* 34(5): 946–54.

Zigler, Edward, and Penelope K. Trickett. 1978. "IQ, Social Competence, and Evaluation of Early Childhood Intervention Programs." *American Psychologist* 33(9): 789–98.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 133–152

What Can Be Done To Improve Struggling High Schools?

Julie Berry Cullen, Steven D. Levitt, Erin Robertson, and Sally Sadoff

The task faced by high schools is perhaps the most difficult in all of education. High schools not only are charged with preparing students for college, but also with serving the needs of the many students who will directly enter the workforce. High schools have to offer greater portfolios of options and sort students appropriately across those options. At the same time, compared with lower grade levels, high schools inherit student bodies with more discrepant abilities and competing outside options and influences (for example, Clotfelter, Ladd, and Vigdor 2009; Cascio and Staiger 2012). The balancing act is all the more challenging at disadvantaged high schools, where resources are of lower quality and more scarce.

The National Center for Education Statistics (Aud et al. 2012) provides a wealth of data showing that many US high schools are struggling. Across all public schools, only around 75 percent of students graduate on time, and approximately 8 percent of students drop out of high school altogether. In large urban school districts, like Chicago, New York City, and Los Angeles, cohort graduation rates hover around 60 to 65 percent. In the lowest income quintile, dropout rates are four times greater than in the highest income quintile. These statistics are especially shocking since

■ *Julie Berry Cullen is Associate Professor of Economics, University of California–San Diego, La Jolla, California. Steven D. Levitt is William B. Ogden Distinguished Service Professor of Economics and Director, Becker Center on Chicago Price Theory, both at the University of Chicago, Chicago, Illinois. Erin Robertson is a Research Professional, Becker Friedman Institute for Research in Economics, University of Chicago, Chicago, Illinois. Sally Sadoff is Assistant Professor of Management and Strategy, Rady School of Management, University of California–San Diego, La Jolla, California. Their email addresses are jbcullen@ucsd.edu, slevitt@uchicago.edu, elouise@uchicago.edu, and ssadoff@ucsd.edu.*

http://dx.doi.org/10.1257/jep.27.2.133. doi=10.1257/jep.27.2.133

dropping out of high school has increasingly become an economic death sentence.[1] A comparison of those who receive exactly twelve years of education versus those who stop just short of receiving a high school diploma yields a $300,000 difference in lifetime earnings, according to our calculations based on data from the 2010 American Community Survey.

In this paper, we suggest that underperforming high schools are failing in large part because traditional paradigms do not meet the needs of many of their students. The majority of high schools have sought to provide all students with academic skills from a primarily college-preparatory and nonexperiential perspective, with limited nonacademic supports. This mission has received renewed emphasis under the recent national school accountability movement. Yet, this emphasis is likely setting many high schools up to fight a losing battle, because a higher proportion of students from disadvantaged backgrounds lack the requisite skills to succeed by this definition. Test score gaps are around 0.7 standard deviations for minority students entering and exiting high school, and are equally large in subject areas like English and math (that are most often the focus of interventions) as in areas like history and science (Fryer 2011b). While increasing school inputs in failing schools or shifting disadvantaged students to high-performing schools have had limited effects on student outcomes, efforts to engage low-performing students through changes in the types of schools and classes available to these students appear to result in large gains in graduation rates and labor market outcomes.

In essence, our advice to high schools when it comes to underperforming students is to redefine the mission and eschew traditional success metrics like test scores, focusing instead on more pragmatic objectives like keeping kids out of trouble, giving them practical life skills, and helping with labor market integration. That conclusion will no doubt be unsatisfying to many readers. In an ideal world, high schools would perform miracles, bringing struggling students back from the brink and launching them towards four-year college degrees. Indeed, a few remarkable and innovative schools seem to be succeeding at that lofty objective. We discuss these programs, which offer a stark alternative to technical education, but with the important caveat that we are skeptical that these achievements can be generalized on a large scale. More likely, attempts to do so would be extremely costly and largely ineffective.

A fundamental question, even if one accepts the conclusions of the preceding paragraph, is how best to organize the delivery of these services. The success of vocational programs would seem to be especially dependent on a good match between students' interests and abilities and the types of career programs offered. One model is to create stand-alone career academies organized around a specific career theme. That approach has attractive features from an economies-of-scale perspective and will work well if the right kinds of students are willing to travel substantial distances to attend. A second vocational model is a "school within a school" model that injects

[1] Murnane (2013) provides a comprehensive review of the dropout problem, including evidence on how incentives to invest in education have evolved over time and on the effectiveness of interventions at different ages, including some of the high school interventions considered in this paper.

vocational options into schools that emphasize more traditional academic goals. The risk of this approach is that the range of vocational opportunities offered is likely to be limited at any one school, and good match quality may then be harder to achieve.

In fact, the quality of the match between students and high school programs, in our view, is the most critical issue facing high school reform today. Focusing policy on changes in resources may be effective in early grades, where there is potential for such investments to improve students' cognitive and noncognitive skill levels and trajectories. In contrast, the area of reform with the largest potential to improve high school outcomes like graduation is to provide struggling students with an increased variety of targeted educational models and schools (Jacob and Ludwig 2008; Murnane 2013). The most hopeful results have been seen in this area.

In this paper, we begin with a selective overview of the evidence regarding the effectiveness of standard school inputs in overcoming gaps in outcomes at the high school level. The key lesson that we highlight is that changes in school quality, retaining traditional models, do not appear to be adequate. We turn to the case for alternative models that emphasize the development of pragmatic job skills. We then consider models with loftier goals, such as KIPP and Harlem Children's Zone. We conclude with reflections on the most promising directions for research and reform.

Standard School Inputs—Vertical Differentiation

One way to view failing high schools is through the lens of a common high school education production function. Among the wide array of factors influencing student outcomes are capital and labor inputs, including administrators, teachers, and peers, along with the design and rigor of the curriculum. The level and quality of these inputs vertically differentiate schools, and low-performing high schools tend to rank poorly on these dimensions. The US Department of Education's *Toolbox Revisited* (Adelman 2006) documents that higher socioeconomic students have access to more-rigorous high school curricula; for instance, 72 percent of 1992 high school seniors in the highest quintile by socioeconomic status attended schools offering calculus compared to just 44 percent in the lowest quintile. Of students whose 5th-grade math scores placed them in the top half, 26 percent of African-Americans took Algebra I or another advanced math course in the 8th grade, while 60 percent of their white peers were enrolled in these courses (Ross et al. 2012). Low-income and minority students are also exposed to teachers with less experience and fewer qualifications than higher socioeconomic students (Lankford, Loeb, and Wyckoff 2002). In California, for example, students in the bottom income quartile have math and science teachers with an average of three fewer years of experience than do students in the top quartile (Socias, Chamber, Esra, and Shambaugh 2007). Even the conditions under which these students study vary greatly. Minority students are more likely to attend schools with trash on the floor, graffiti, and chipped paint (Planty and DeVoe 2005).

Evidence on Specific Inputs

Would increasing standard school inputs be sufficient to eliminate the wide gulfs in results between low-achieving and high-achieving high schools? The first step in responding to this question is to establish causal ties between these inputs and student outcomes, both overall and for disadvantaged students. Unfortunately, there is a dearth of evidence on the productivity of specific school inputs at the high school level. While the body of evidence on the role of inputs such as class size and teacher quality in earlier grades continues to expand, most studies relying on quasi-experimental methods exclude high school students. Often, the justification behind focusing analysis on lower grades is that the production process in high school is too complex, given that high school students take different sets of courses and are taught by teams of teachers. While this argument makes sense based on the desire of researchers to have clear identification strategies, it also leaves important gaps in the empirical evidence.

One result that does stand out from the existing literature is that increasing the overall level of resources is a blunt instrument for helping at-risk students. Perhaps the best evidence on this comes from the school finance equalization movement, which greatly mitigated disparities in resources across school districts. By exploiting variation in the timing of states' adoption of finance equalization policies, Card and Payne (2002) show that these policies erase only about 5 percent of the gap in SAT scores between high- and low-income students, with 95 percent remaining.

The limited evidence regarding capital inputs is equally discouraging. Though broader upgrades to high school facilities might matter, attempts to address the digital divide have not seemed to make much difference. Goolsbee and Guryan (2006) estimate the impact of a federal subsidy program on Internet investment across California schools from 1996 to 2000. Despite dramatic improvements in accessibility, particularly for the disadvantaged high schools that were assigned larger subsidy rates, they find little effect on student achievement. It may be that it is not the presence of computers, but rather the ability of computer-aided instruction to provide individualized lessons, that will offer the real payoff. For example, analyzing a randomized experiment implemented in three school districts in 2003 and 2004, Barrow, Markman, and Rouse (2009) find some evidence that computer-aided instruction in algebra and pre-algebra for 7th–9th graders improved student outcomes, especially in large classes.

Labor inputs, and in particular the quality of personnel, appear more important. More-effective principals have a positive impact on student test scores, but high-poverty schools have a large variance in principal quality (Brewer 1993; Branch, Hanushek, and Rivkin 2011). High value-added principals may be successful in part because they are particularly effective at selecting high-quality faculty and firing failing teachers. After a policy change in Chicago gave principals greater freedom to dismiss faculty, Jacob (2011) found that principals did use some measures of teacher productivity in making firing decisions. Productive teachers have been shown to increase graduation rates among students on the margin (Koedel 2008) and to be particularly valuable to students coming into high school with academic deficits.

For instance, Aaronson, Barrow, and Sander (2007) demonstrate that having a math teacher one standard deviation above average results in a 0.13 grade equivalent improvement in average test scores for low-ability 9th graders, explaining approximately one quarter of the improvement in test scores over the year.

Unfortunately, recruiting and retaining high-quality teachers at schools serving disadvantaged students may prove especially difficult (Lankford, Loeb, and Wyckoff 2002; Clotfelter, Ladd, Vigdor, and Wheeler 2007). An alternative to reallocating teachers is to offer incentives to existing teachers. While most evaluations of performance pay have not found positive effects in US high schools, recent studies reveal that small tweaks to the way incentives are structured can greatly impact their effectiveness. Fryer, Levitt, List, and Sadoff (2012) found that K–8 classrooms in which teachers received performance bonuses upfront, which they then had to return if student achievement did not improve enough, saw math gains equivalent to a one standard deviation increase in teacher quality. To get longer-term boosts in effort and grades, Levitt, List, and Sadoff (2012) found that deferred financial incentives offered either to students or to parents combined with personal "cheerleading" can work. Putting these bonus schemes into widespread operation, however, is no easy task.

Another important labor input that is not directly compensated and is perhaps even harder to manipulate is peer quality. Within high schools, the characteristics of classmates tend to be correlated with the level of coursework because of the common practice of tracking, making it difficult to tease out the role of peers. Though there has been a push to remove tracking from the typical high school model, research estimating the combined effects of peers and coursework suggest that concerns about increased inequality with tracking are unfounded. For example, Betts and Shkolnik (2000) and Figlio and Page (2002) find no relationship between tracking and outcome gaps in national samples of middle and high school students from the Longitudinal Survey of American Youth and the National Educational Longitudinal Study, respectively. In fact, Figlio and Page (2002) find that tracking may improve math scores for lower-performing students.

More direct evidence on the role of access to and participation in advanced coursework per se is more mixed. Several studies that exploit changes in graduation requirements show that increasing the number of required math courses decreases the wage gap between disadvantaged and middle-income students (Altonji 1995; Betts and Rose 2004; Goodman 2012). Yet the more prevalent exit exams and rigorous credit requirements embedded in states' school accountability systems as a result of the No Child Left Behind legislation have had the opposite effect on low-ability students than was intended, as these students are induced to drop out at higher rates (Dee and Jacob 2007; Papay, Murnane, and Willett 2010).

Evidence from Reassigning Students across Traditional Public Schools

Given the limited evidence on the efficacy (and heterogeneity in the efficacy) of specific inputs and the difficulty of aggregating results from studies that pull one lever at a time, the most convincing evidence on whether replicating high-performing high schools would be sufficient to eliminate gaps comes from reallocations of students across traditional schools. Open enrollment allows us to approximate this thought experiment. Under open enrollment, students are allowed to apply to any public school in the school district. Depending on how prevalent charter and magnet schools are, this strategy retains the traditional structure of public schools, but allows students who might otherwise attend low-performing neighborhood schools access to higher-performing alternatives.

An aspect of open enrollment that facilitates analysis is that schools that are oversubscribed typically admit a subset of interested students through a lottery. An associated limitation, though, is that the least advantaged students tend not to participate, so any findings may not extrapolate to them. For example, Cullen, Jacob, and Levitt (2005) document that three-quarters of rising freshmen from the top quartile of the ability distribution opt out of their neighborhood schools within the Chicago public schools, while only one-third of students from the bottom quartile do. Similar patterns are found in Charlotte-Mecklenburg, which also now has a well-established open enrollment program with high overall rates of participation (Deming, Hastings, Kane, and Staiger 2011).

Though students participate in the lotteries at high rates, the evidence for academic benefits from attending a high- rather than low-performing high school is not there. Among applicants to oversubscribed high schools, Cullen, Jacob, and Levitt (2006) find little to no effect of gaining access to a higher-achieving high school on academic outcomes, suggesting differences in outcomes across schools are driven by differences in student caseloads rather than inputs. Even when taking heterogeneity in student populations into account, there is little evidence that benefits accrue to different subsets of students, such as those students who face the greatest potential gains from attending a lottery school (for example, those students who attend schools with higher-quality peers than their next-best option). In fact, among those students, the likelihood of dropping out increases by nearly 11 percentage points in comparison to their peers who did not win the lottery.

In current work on the Charlotte-Mecklenburg schools, Deming, Hasting, Kane, and Staiger (2011) find similarly weak support for academic gains to attending a higher-performing high school on average, with null effects for test scores and college enrollment and an approximately 5 percent rise in high school graduation. The authors do emphasize that these average effects mask heterogeneous impacts. Strikingly, lottery winners who otherwise would have attended one of the four lowest-quality high schools experience no gain in 9th grade test scores but were 9 percentage points more likely to graduate high school and about 6.5 percentage points more likely to attend a four-year college.

In interpreting their findings, it is important to realize the treatment applied to this subgroup does not align with our thought experiment (where we simply increase standard school inputs). Only 15 percent of lottery applicants from the four lowest-quality high schools were effectively randomly assigned and so were included in the analysis, and more than two-thirds of the winners who took up the offer chose to attend one of the three district-wide magnets. These magnets have career and technical emphases, so these findings accord with Cullen, Jacob, and Levitt (2005), who find that positive returns were seen only for those students attending vocational schools. Students who attended vocational high schools were on the order of 15–20 percentage points more likely to graduate than their peers in other school models. Our view is that this vocational focus is key, and we return to this issue below.

Evidence from Restructuring Struggling High Schools

A possible explanation for why the typical "good" school does not improve outcomes for students opting out of "bad" schools is that such students might require specialized supports. Whole school reform models and small school initiatives seek to take the lessons learned about the relative importance of personnel, peer, and capital inputs in order to provide a more targeted education for low-achieving students. Both have delivered mixed results.

Whole school reform models are efforts that provide incentives for dramatic changes in personnel and policies and provide additional funding for wrap-around supports services for students. As part of the economic stimulus package enacted in 2009, the federal government greatly expanded the Title I School Improvement Grants subprogram. These new grants (of up to $2 million per year) were awarded to school districts according to the prevalence of low-performing schools and required adopting federally sanctioned school reform models. Comparing those barely eligible versus ineligible reveals that receipt does appear to have some positive effects on performance for the lowest-performing California high schools that replaced the school leader and most of the staff (Dee 2012). Improvements were uneven, though, across chosen models and targeted schools.

The small schools movement reorganizes large high schools into smaller autonomous schools in order to provide more-cohesive sets of teachers and peers and individualized attention. Several of these initiatives have been successful. For instance, a study of lottery participants applying to around 100 public small schools in New York City revealed that attendees experienced increases in the likelihood of graduating within four years of 8.6 percentage points (Bloom and Unterman 2012). In Chicago, Barrow, Claessens, and Schanzenbach (2013) also find that students in small schools are more likely to graduate, despite no signs of improvements on test scores. Results have been varied, though, and the movement that was once championed by the Bill Gates Foundation has since been largely dismissed. One of the issues is that small schools differ in their specific missions and the degree to which they adhere to the tenets. As in the case of open enrollment, it may be that it is the subset of small schools with career and technical missions that drive the results, which would imply that size is of second-order importance.

Alternative Models—Horizontal Differentiation

The bulk of the evidence discussed in the preceding section suggests that more inputs, structured in the usual fashion, or that access to high schools in other neighborhoods are unlikely to yield dramatically improved outcomes for struggling high school students. One limitation that students and families face is a dearth of different school models from which to choose. The majority of districts continue to provide access primarily to traditional, college-preparatory schools, with 80 percent of public high schools providing a traditional education (Snyder and Dillow 2012). Potential gains from match quality would be realized by providing a range of schooling options that better fit students' needs. Reviewing the literature across all grade levels, Jacob and Ludwig (2008) similarly conclude that expanding the school choices available to students and families can only help currently underserved students. In fact, in recent years, two nontraditional educational models in particular have emerged that have shown hopeful results.

The first of these can be viewed as capitulation or as realism: in either case, a recognition that students entering high school with low skills and little academic motivation are likely better served by a vocational model. The second approach is what one might call the Herculean effort strategy: radical programs that go beyond standard academic approaches, emphasizing noncognitive skills and social pressure to achieve, to sharply change students' motivations and goals. Beyond these two paths, the only other obvious answer is educational improvements earlier in life so that the gaps facing high schools are less daunting.

Vocational Focus

High schools and programs with a specialized focus can help the disadvantaged students who choose them by both playing to their interests and by overcoming informational and network deficits. Of the 70 percent of high school completers who enroll in two- or four-year colleges right after high school, only around 40 percent enroll in a four-year degree program, and only 60 percent of these students will graduate in six years or less. Among schools with open admissions policies—where the lowest ability students will likely enroll—graduation rates are half that, at 29 percent. Graduation rates at two-year institutions are equally low, with only 30 percent of enrolled students graduating within three years (Aud et al. 2012). For large numbers of high school students, preparing for a two- or four-year university does not match well with their trajectories.

A career-oriented track can potentially improve outcomes for low-achieving students on the margin of dropping out of school if it provides technical skills valued by the market and/or pushes them over the margin of enough perceived gains from schooling to avoid dropping out. While a college degree does offer high labor market returns, many growing and relatively profitable industries require only a high school degree. Table 1 shows average 2010 earnings of high school graduates (excluding those with any education beyond high school) and dropouts 35 to 44 years of age, by broad industry categories. Compared to people who have an 11th or 12th grade education but no diploma, high school graduates earn more across all industry categories. The returns to a high school degree are particularly high in areas that demand sector-specific skilled labor. While business and manufacturing and production continue to employ the majority of individuals with a

Table 1

Average 2010 Earnings of High School Graduates (Excluding Those with Education beyond High School) and Dropouts Ages 35 to 44 by Industry Category

(number of observations in parentheses)

	Graduates		*Dropouts*	
	Male	*Female*	*Male*	*Female*
Food & maintenance	$24,866 (441,746)	$16,271 (430,399)	$20,758 (102,280)	$14,712 (98,423)
Community & education	$29,406 (97,368)	$21,294 (570,649)	$22,745 (14,923)	$18,344 (78,000)
Manufacturing & production	$37,546 (2,406,371)	$23,481 (378,185)	$29,611 (419,210)	$20,177 (68,814)
Military & law enforcement	$45,616 (130,426)	$33,284 (39,284)	$37,659 (5,392)	$25,935 (3,473)
Business	$46,223 (901,278)	$29,907 (1,410,816)	$37,289 (100,523)	$23,935 (122,442)
Technology	$53,452 (122,519)	$35,268 (121,897)	$40,891 (9,476)	$31,826 (7,331)

Notes: Average earnings are estimated using the 2010 American Community Survey Public Use Microdata Sample (ACS PUMS). The first two columns show earnings for people who received a regular high school diploma (but no further schooling) and report annual earnings greater than $1,000. The third and fourth columns show earnings for people who received an 11th or 12th grade education but no diploma and report annual earnings greater than $1,000. The 2010 ACS Occupation Codes were used to categorize industries. "Business" includes management, business, science, and art, business operations specialists, financial specialists, legal, office and administrative support, and sales and related occupations; "Community & education" includes education, training, and library, and community and social services, healthcare support, and personal care and service; "Food & maintenance" includes food preparation and serving, and building and grounds cleaning and maintenance; "Manufacturing & production" includes construction and extraction, extraction workers, installation, maintenance, and repair, transportation and material moving, and farming, fishing, and forestry; "Military & law enforcement" includes military-specific occupations and protective service occupations; and "Technology" includes healthcare practitioners and technical, computer and mathematical, architecture and engineering, life, physical, and social science, and arts, design, entertainment, sports, and media.

high school diploma, technology appears to be a largely untapped potential field with relatively high income. Among 35 to 44 year-old males, graduates who work in the tech industry earn on average more than $50,000—or $7,000 more than high school graduates in business and over $10,000 more than high school dropouts in the same field. Those in food and maintenance and in community and education jobs, which tend to require unskilled labor, fare the worst, with women in food and maintenance making under $15,000 annually.

Evidence suggests that career-oriented programs improve both attainment and market-valued skills. In their analysis of career and technical magnet schools in New York City, Crain, Heebner, and Si (1992) find significant improvements in high school enrollment and graduation. This accords with the gains in graduation rates found by Deming, Hastings, Kane, and Staiger (2011) and Cullen, Jacob, and Levitt

(2005) under open enrollment, which in both cases are driven by such magnets. Beyond increasing graduation rates, there is some evidence that schools may also funnel students into vocational fields and classes, such as technology, that can particularly benefit them in the labor market. Bishop and Mane (2004) survey an array of evidence that suggests significant labor market returns to computer classes taken in school.

Most students have access to some vocational education. Around 80 percent of all high schools in 2008 offered career or technical courses, which include everything from business and computer classes to more-traditional high school classes like shop and home economics. But there are likely large differences between the majority of schools that provide some vocational courses within a college-prep centered curricula and schools like the magnet schools found to have sizable impacts that offer vocational tracks. In fact, while over 90 percent of students graduate with vocational credits, only a little over 20 percent complete an occupational concentration (Levesque, Laird, Hensley, Choy, Cataldi, and Hudson 2008). Enrollment in dedicated vocational schools has decreased from around 190,000 students in the 2000–2001 school year to approximately 125,000 in 2009–2010, despite an increase in the number of vocational schools from around 1,000 to over 1,300 (Snyder and Dillow 2012).

Given the evidence above, it seems likely that existing programs, on average, are undersubscribed. Why is there such low participation in vocational programs despite their potentially high returns? One reason may be lack of easy access to these schools. There is strong evidence that even under open enrollment, students lean heavily toward attending nearby schools (Cullen, Jacob, and Levitt 2005). Furthermore, information likely plays an important role in matching students to the correct educational model (Hastings and Weinstein 2008). If families live far from vocational options or do not know what those options are, students are unlikely to enroll in these programs. Furthermore, even in districts where vocational programs are in high demand, such as Chicago, many students who may benefit from a vocational education may not apply to these schools for these same reasons. In Chicago, for example, just 11 of 106 public high schools are vocational high schools, and while a centralized website provides a wealth of information on the types of career programs these schools offer, it does not have other basic metrics, such as graduation rates and school size, on which families may base schooling decisions.

One solution to both of these problems—lack of proximity and information—are smaller vocational schools, known as career academies, that operate as a subset of a larger, more traditional public school. While the broader small schools movement has lost steam, these small-school career academies—which can be found in over 6,000 high schools today (Snyder and Dillow 2012)—have been gaining momentum. By having focused, career-oriented tracks and partnering with local businesses and community colleges, career academies aim to graduate students with career and technical skills and an established business network, in addition to the skills necessary to enter two- and four-year colleges. In their evaluation of nine representative urban high schools, Kemple and Snipes (2000) find that career academy attendance significantly lowers dropout rates among high-risk students. Similarly, tracking students from one district, Maxwell and Rubin (2002) find that career academy students have higher graduation rates and a higher likelihood of starting a postsecondary education than students in traditional settings. And, studying JROTC (Junior Reserve Officer Training Corps) career academies across five major urban school districts, Elliot, Hanser, and Gilroy (2002) found improved graduation and attendance rates.

What is more, graduates of career academies see positive labor market outcomes, particularly among men. Over an eight-year period post-graduation, Kemple (2008) found in a randomized controlled study that male graduates of career academies saw a 17 percent increase in monthly income, earning a total of around $30,000 more than their non-career-academy peers. Career academies also seem to have had at least some success in funneling graduates into higher-income sectors; 7 percent of career academy graduates worked in tech fields compared to 4 percent in the control, for instance. Other school-to-work programs that integrate work-based learning have been found to increase the probability of employment after graduation, as Neumark and Rothstein (2006) show in an analysis of 1997 National Longitudinal Survey of Youth data in the aftermath of a temporary federal program that provided additional funding for such programs.

Given these results, expanding these programs to reach more at-risk students, particularly men, seems like a priority, and in fact, the Department of Education recently proposed funding for 3,000 additional career academies (US Department of Education 2012). As initiatives like these push vocational models, it is important that the right types of students are targeted. In 2000, for example, only 23 percent of high schools with more than 50 percent of students eligible for free or reduced price lunch had designated career academies. This is especially low given that among schools having less than 5 percent of students eligible for subsidized lunch, only 21 percent offered career academies (Silverberg, Warner, Fong, and Goodwin 2004).

Over the past decade, the primary federal policy aimed at vocational education—the Carl D. Perkins Vocational and Technical Education Act—has increasingly supported the further integration of academic and college-preparatory work with vocational education (Silverberg, Warner, Fong, and Goodwin 2004; Dann-Messier 2012). In other words, federal policies seem likely to encourage vocational programs to become less vocational. While the success of programs such as career academies may in part be due to the academic work required of students, one worry is that an increased focus on college prep activities will dilute the effectiveness of vocational tracks. Two models that have emerged from this increased college readiness focus have been largely successful, however. Tech-Prep programs place a strong emphasis on technology-related courses and partner with community colleges to help students earn college credits while in high school and guide them to two-year associate degree and certificate programs. The Department of Education estimates that nearly 50 percent of high schools and almost all community and technical colleges offer Tech-Prep. It has been found to increase high school graduation rates and enrollment at two-year colleges, although this may come at the expense of student enrollment in four-year degree programs (Cellini 2006). A much smaller program (in 2005, involving under 100 high schools nationwide), Talent Development, combines school-within-school career academies with a college-preparatory curriculum and offers a range of remedial opportunities. An interrupted time-series evaluation of early-adopting schools finds impacts on both test scores and graduation rates (Kemple, Herlihy, and Smith 2005).

Herculean Efforts

This focus on college preparatory work even with career and technical education programs has been heightened as the Obama administration pushes for "every American to commit to at least one year of higher education or postsecondary training" (Dann-Messier 2012). However, programs on the other extreme from vocational education that specialize in high-powered college preparation have not been as effective in meeting their goals. Exam schools, which admit students using entrance exams, have been linked possibly to improved short-term outcomes (Dobbie and Fryer 2011c; Abdulkadiroğlu, Angrist, and Pathak 2011), but not to improved long-term outcomes such as college attendance.

This is where the "Herculean model" of charter schools comes in to bridge the gap. Over the past 10 years, enrollment in secondary charter schools has grown four-fold, even though most empirical research has not found a positive impact of charter schools on student achievement. In a recent meta-analysis of 25 studies using experimental approaches, Betts and Tang (2011) find that there are no significant effects of charter high schools on average, although effects do tend to be larger in urban schools.

However, a few wildly successful charters—those Herculean schools that follow what is known as the "No Excuses" model—have emerged. This model puts a strong emphasis on a school culture that promotes academic rigor and high behavioral standards, uses data to select and retain high-quality teachers, and has a longer school day and year.[2] Looking across charter school models in New York City, Dobbie and Fryer (2011b) document that those that follow a No Excuses Model (or use similar practices) have the largest impact.[3] Angrist, Pathak, and Walters (2011) reach a similar conclusion in an examination of the effects of a large sample of charter schools in Massachusetts. Using randomized admissions lotteries, Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak (2011) find that average effects of attendance in these types of schools in Boston are large enough to close the math and reading achievement gaps between black and white high school students.

[2] For further discussion of the No Excuses model, see for example Carter (2000), Thernstrom and Thernstrom (2004), and Whitman (2008).

[3] They also find that standard input measures of class size, per pupil expenditure, and teacher qualifications are not correlated with school effectiveness. Similarly, Hoxby and Murarka (2009) find that the effectiveness of New York City charter schools is not associated with inputs like class size, but is strongly correlated with having a longer school year.

Dobbie and Fryer (2011a) find similar-sized effects in math (and smaller effects in reading) from attending Harlem Children's Zone (HCZ) No Excuses charter middle school. These students also have fewer absences, despite the fact that the lowest achieving students are required to be in school for roughly twice as many hours as the average New York City Public School student. The Knowledge is Power Program (KIPP) schooling model yields similar results, and low-ability students appear to have the largest gains (Angrist, Dynarski, Kane, Pathak, and Walters 2010; Betts and Tang 2011; Angrist, Dynarski, Kane, Pathak, and Walters 2012).

An open question is whether these small-scale educational successes can be scaled up and replicated. After all, these models do require an almost Herculean effort at all levels of the school—from administrators who foster a uniform school culture, to teachers who work longer hours, to students and parents who must be scholastically dedicated. There is some evidence that KIPP schools have been able to replicate their success. Gleason, Tuttle, Gill, Nichols-Barrer, and The (2012) use propensity score matching to measure the impact of admission to 22 KIPP schools across the country and find average test score effects equivalent to a year's worth of growth in math and three-quarters of a year in reading.

Fryer (2011a) examines an alternative to charter school expansion, testing whether key elements of the No Excuses models can be incorporated into traditional public schools. Nine underperforming middle and high schools in Houston first replaced all principals and half of the teaching staff, and then applied four tenets similar to those used in programs like KIPP and HCZ. These included increased instructional time, tutoring, data-driven instructional practices, and the fostering of a culture of high academic and behavioral expectations. Initial results after the first year saw achievement gains on par with those in No Excuses charter schools. This study provides the first set of results to indicate that the lessons learned from these charter school programs may generalize. This bodes well for reforms such as the federal government's recent Race to the Top program, which is designed to reduce the achievement gap by rewarding innovation and promoting educational tenets similar to those implemented in Houston.[4]

While these results are encouraging, they do not address key policy questions about implementing the No Excuses model on a large scale. KIPP, for example, is the nation's largest charter management organization, but serves only about 2 percent of charter school students and less than 0.1 percent of all public school students (Gleason et al. 2012).

The barriers to scale-up on the supply side are related to personnel and funding. As we discussed above, high-quality principals and teachers are a key input into successful schools. This is particularly true in the No Excuses model. Fryer (2011a) documents that over 200 principals had to be interviewed to find nine who

[4] As a caution, models that share these features but do not embed them in the regular school day do not seem to be as effective. Though there were positive impacts at some specific sites, the Quantum Opportunity Program, an after-school program that targeted at-risk youth had no long-term impacts (Maxfield, Schirm, and Rodriguez-Planas 2003).

demonstrated a commitment to the No Excuses model and a record of achievement. No Excuses schools also require teachers who will buy into a nontraditional educational agenda, accept less job security, and meet the heavy demands—including longer hours and greater emphasis on student performance—that these schools place on faculty. This is likely a primary reason why teachers in No Excuses schools tend to be much younger than in traditional public schools; in Boston, 6 percent of charter teachers are 49 or older compared to 40 percent district-wide (Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak 2011). This also implies that these schools must recruit from what is likely a limited pool: young teachers who are dedicated and talented despite their inexperience.

Funding also limits the ability of such schools to spread. Direct comparisons of costs in charter schools and traditional public schools are often difficult due to differences in funding structures. However, Dobbie and Fryer (2011a) estimate that HCZ spends $19,272 per pupil compared to $16,171 per pupil in the median school district in New York State. The authors argue that if the test score effects they measure translate into longer-term educational gains, HCZ easily passes a cost–benefit analysis. However, such arguments do not always produce adequate financial resources.

The No Excuses model also faces important barriers on the demand side. Like vocational schools, No Excuses schools seem to be undersubscribed given their effects on achievement. This is in part due to the same factors we discussed above—lack of information and defaulting into a neighborhood school. In Boston, for example, district-provided resources on schooling options do not typically include information about charter schools, and distance from a charter school is a strong predictor of attendance (Walters 2012).[5]

However, expanding information and access may have a limited impact—in part because many students may not be well matched with the No Excuses model. To make this concrete, imagine a public policy that makes a spot in a No Excuses school available to all current applicants. We know from the existing data that many of those who are admitted will not enroll—about 30 to 40 percent in HCZ (Dobbie and Fryer 2011a)—or will enroll and then leave. The relevant measure for this policy then is the impact of receiving admission to a No Excuses school regardless of actual attendance. In Boston charters, the effects of receiving admission are significant, but are about one-quarter the size of those for actual attendance that we discussed above (Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak 2011; for comparability, we use the authors' estimated effects per potential year spent in a charter school). While there are many reasons why students might gain admission and not enroll, or enroll and then leave, the gap between these effects gives some

[5] There is evidence that charter applicants do learn about school quality and are responsive to this information. Schools that are oversubscribed tend to have larger effects than schools that are undersubscribed (Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak 2011; Angrist, Pathak, and Walters 2011). Similarly, Hanushek, Kain, Rivken, and Branch (2007) argue that exit rates from charter schools are negatively associated with value added measures of school quality.

indication of the importance of exit rates, which may be due in part to students' mismatch with the demands and culture of No Excuses schools.[6]

Another policy we can imagine is one that expands access to those students who could benefit from the programs, but who do not currently apply. Walters (2012) simulates the effect of this policy in Boston, which as of 2011 planned to expand its number of charter schools by about 80 percent. He finds that those low-income, low-ability students who stand to gain the most from No Excuses schools are the least likely to opt-in. These students' high perceived costs of applying and strong preferences for traditional public schools dampen the effects of expansion.

This leads to the experiment, like that taking place in Houston, of integrating elements of successful charters into traditional public schools. This approach may increase the number of students experiencing the No Excuses model, although students for whom the model is a bad match may transfer schools or drop out. One early finding from the Houston experiment is instructive. Fryer (2011a) finds that the program leads to lower college attendance but that conditional on college attendance, students are more likely to enroll in a four-year institution. One interpretation of this result is that the program's explicit goal of 100 percent attendance at a four-year college or university pushes everyone down the same path when a two-year college or vocational training might be a better match for many students. However, it is not clear what the overall impact of these interventions will be, since more generally, we lack evidence on the long-term impact of No Excuses schools that would allow us to compare measures of educational attainment and labor market outcomes from these Herculean efforts to the impacts of vocational schools discussed above.[7]

Targeted Interventions

Why might the two approaches work? In part, their success may stem from addressing gaps in noncognitive skills and structuring learning to motivate and engage students. A growing literature links noncognitive ability—a broad set of skills captured by measures of behavior, personality, and work ethic—to educational achievement as well as longer-term outcomes including employment, wages, health, and crime (for example, Heckman, Stixrud, and Urzua 2006). A smaller set of studies documents racial and gender gaps in these skills. Black and Hispanic children demonstrate more antisocial behavior and receive lower ratings on measures of self-control and interpersonal skills (Carneiro, Heckman, and Masterov 2005; Goldammer 2012). Jacob (2002) finds that high school boys have lower noncognitive

[6] In response to evidence that some charters experience high exit rates, Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak (2011) and Angrist, Dynarski, Kane, Pathak, and Walters (2012) argue that exit rates among charter school students are similar to those for their peers in traditional public schools.

[7] Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak (2011) find that Boston No Excuses charter schools do not have a significant impact on high school graduation. More generally, there is a dearth of evidence on the longer-term impacts of charter schools of any type. In a lottery-based study, Strick (2009) finds that students in a single San Diego charter school are more likely to attend college. In an observational study, Booker, Sass, Gill, and Zimmer (2011) find that charter school students in Chicago and Florida high schools have higher graduation and college attendance rates.

ability than girls on measures of behavior and work ethic, and argues that this can partially explain the gender gap in college attendance. These findings point to potential mechanisms for several of the results discussed above—for example why career academies offering an alternative to college may be especially beneficial for boys, and why the No Excuses model's focus on behavior particularly benefits low-achieving and minority students.

Understanding the role of noncognitive skills in these larger interventions could help guide smaller-scale policies. There is limited but intriguing recent evidence on attempts to manipulate student noncognitive skills and effort that suggests, if targeted well, such efforts could be quite cheap. Hill, Roberts, Grogger, Guryan, and Sixkiller (2011) discuss several interventions that reduce delinquency by intervening on noncognitive rather than cognitive skills. In this spirit, the initial results from an intervention among disadvantaged high school boys in Chicago, which focused on developing skills related to emotional regulation and social-information processing, suggest large increases in schooling outcomes and large decreases in violent crime arrests (University of Chicago Crime Lab 2012). Equally striking is that short-term financial incentives, which address students' lack of ability to plan for the future, appear to lead to increased effort and improved test scores (Braun, Kirsch, and Yamamoto 2011; Levitt, List, Neckermann, and Sadoff 2012). Such targeted interventions can stand alone within a traditional school or can potentially complement the alternative models discussed in this section.

Implications

In spite of decades of well-intentioned efforts targeted at struggling high schools, outcomes today are little improved. A handful of innovative programs have achieved great success on a small scale, but more generally, the economic futures of the students at the bottom of the human capital distribution remain dismal. In our view, expanding access to educational options that focus on life skills and work experience, as opposed to a focus on traditional definitions of academic success, represents the most cost-effective, broadly implementable source of improvements for this group.

Increased school choice has been a centerpiece of educational policy reform over the last decade. In its current incarnation, it primarily provides underserved students access to higher-quality, traditional, college preparatory schools. School districts that allow some schools to deviate from this model in the form of career magnet schools and academies have seen the greatest impact on student achievement through improved match quality. Expanding vocational options should be a primary objective of school choice, not merely an afterthought. A possible area for innovation would be to focus curricula around those job market sectors—such as technology and business—that yield the highest returns to a high school diploma.

The other nontraditional schooling options that most help struggling students are Herculean models such as No Excuses. Unfortunately, perhaps the scarcest resource in education today is innovators like Geoffrey Canada, who started the Harlem Children's Zone, or Mike Feinberg and Dave Levin, who founded KIPP. Indeed, these innovations have occurred mostly at a small scale, and it is unclear whether schools that adopt these models without the guidance of the founders will prove as successful. One of the greatest gaps in our current understanding is the process by which small-scale interventions can be actualized on a grand scale. That area deserves to be the focus of intense research effort.

Ultimately, we believe that a program of education reform will work best when students and families are provided with a variety of schooling models from which to choose, and when information about these choices is disseminated effectively. Future reform efforts should aim to expand the variety of educational models available to underachieving high school students and reward broader measures of success than those embedded currently in school accountability systems. School districts could help foster innovations by using school choice legislation as a way to allow for sorting and better match quality among secondary school students and their schools. In this type of educational environment, empirical researchers have the opportunity to provide guidance to policymakers and innovators. More research is needed to sort successful school models and components from those that do little to improve graduation rates and subsequent labor market outcomes.

■ *We would like to thank David Autor, Julian Betts, John List, and Timothy Taylor for valuable feedback and suggestions. Steven Levitt is an unpaid board member of Blueprint, a small non-profit that does implementation of school turnarounds. Julie Berry Cullen receives consulting income from Educational Testing Service for evaluating the Council for Opportunity in Education's Go College initiative.*

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25(1): 95–135.

Abdulkadiroğlu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak. 2011. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." *Quarterly Journal of Economics* 126(2): 699–748.

Abdulkadiroğlu, Atila, Joshua D. Angrist, and Parag A. Pathak. 2011. "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools." NBER Working Paper 17264.

Adelman, Clifford. 2006. *The Toolbox Revisited: Paths to Degree Completion from High School through College.* US Department of Education.

Altonji, Joseph G. 1995. "The Effects of High School Curriculum on Education and Labor Market Outcomes." *Journal of Human Resources* 30(3): 409–38.

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2010. "Inputs and Impacts in Charter

Schools: KIPP Lynn." *American Economic Review* 100(2): 1–5.

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2012. "Who Benefits from KIPP?" *Journal of Policy Analysis and Management* 31(4): 837–60.

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters. 2011. Explaining Charter School Effectiveness. NBER Working Paper 17332.

Aud, Susan, William Hussar, Frank Johnson, Grace Kena, Erin Roth, Eileen Manning, Xiaolei Wang, and Jijun Zhang. 2012. *The Condition of Education 2012* (NCES 2012-045). National Center for Education Statistics, US Department of Education.

Barrow, Lisa, Amy Claessens, and Diane Whitmore Schanzenbach. 2013. "The Impact of Chicago's Small High School Initiative." NBER Working Paper 18889.

Barrow, Lisa, Lisa Markman, and Cecilia Rouse. 2009. "Technology's Edge: The Educational Benefits of Computer-Aided Instruction. *American Economic Journal: Economic Policy* 1(1): 52–74.

Betts, Julian R., and Heather Rose. 2004. "The Effect of High School Courses on Earnings." *Review of Economics and Statistics* 86(2): 497–513.

Betts, Julian R., and Jamie L. Shkolnik. 2000. "The Effects of Ability Grouping on Student Achievement and Resource Allocation in Secondary Schools." *Economics of Education Review* 19(1): 1–15.

Betts, Julian R., and Y. Emily Tang. 2011. *The Effect of Charter Schools on Student Achievement: A Meta-analysis of the Literature.* National Charter School Research Project.

Bishop, John H., and Ferran Mane. 2004. "The Impacts of Career-Technical Education on High School Labor Market Success." *Economics of Education Review* 23(4): 381–402.

Bloom, Howard S., and Rebecca Unterman. 2012. "Sustained Positive Effects on Graduation Rates Produced New York City's Small Public High Schools of Choice." MDRC Policy Brief. http://www.mdrc.org/sites/default/files/policy brief_34.pdf.

Booker, Kevin, Tim R. Sass, Brian Gill, and Ron Zimmer. 2011. "The Effects of Charter High Schools on Educational Attainment." *Journal of Labor Economics* 29(2): 377–415.

Branch, Gregory F., Eric A. Hanushek, and Steven G. Rivkin. 2011. "Estimating Principal Effectiveness." Working Paper 32, National Center for Analysis of Longitudinal Data in Education Reseach (CALDER), Urban Institute.

Braun, Henry, Irwin Kirsch, and Kentaro Yamamoto. 2011. "An Experimental Study of the Effects of Monetary Incentives on Performance on the 12th-Grade NAEP Reading Assessment." *Teachers College Record* 113(11): 2309–44.

Brewer, Dominic J. 1993. "Principals and Student Outcomes: Evidence from U.S. High Schools." *Economics of Education Review* 12(4): 281–92.

Card, David, and A. Abigail Payne. 2002. "School Finance Reform, the Distribution of School Spending, and the Distribution of Student Test Scores." *Journal of Public Economics* 83(1): 49–82.

Carneiro, Pedro, James J. Heckman, Dimitriy V. Masterov. 2005. "Understanding the Sources of Ethnic and Racial Wage Gaps and their Implications for Policy." In *Handbook of Employment Discrimination Research: Rights and Realities*, edited by Laura Beth Nielsen and Robert L. Nelson, 99–136. Springer: Amsterdam.

Carter, Samuel C. 2000. *No Excuses: Lessons from 21 High-Performing, High-Poverty Schools.* Heritage Foundation.

Cascio, Elizabeth U., and Douglas O. Staiger. 2012. "Knowledge, Tests, and Fadeout in Educational Interventions." NBER Working Paper 18038.

Cellini, Stephanie Riegg. 2006. "Smoothing the Transition to College? The Effect of Tech-Prep Programs on Educational Attainment." *Economics of Education Review* 25(4): 394–411.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2009. "The Academic Achievement Gap in Grades 3 to 8." *Review of Economics and Statistics* 91(2): 398–419

Clotfelter, Charles T., Helen F. Ladd, Jacob L. Vigdor, and Justin Wheeler. 2007. "High Poverty Schools and the Distribution of Teachers and Principals." *North Carolina Law Review* 85(June): 1345–80.

Crain, Robert L., Amy L. Heebner, Yiu-Pong Si. 1992. *The Effectiveness of New York City's Career Magnet Schools: An Evaluation of Ninth-Grade Performance using an Experimental Design.* Assisted by Will J. Jordan and David R. Keifer. Berkeley, CA: National Center for Research in Vocational Education.

Cullen, Julie Berry, Brian A. Jacob, and Steven D. Levitt. 2005. "The Impact of School Choice on Student Outcomes: An Analysis of the Chicago Public Schools." *Journal of Public Economics* 89(5–6): 729–60.

Cullen, Julie Berry, Brian A. Jacob, and Steven D. Levitt. 2006. "The Effect of School Choice on Participants: Evidence from Randomized Lotteries." *Econometrica* 74(5): 1191–1230.

Dann-Messier, Brenda. 2012. "Investing in America's Future: A Blueprint for Transforming Career and Technical Education." Remarks of Assistant Secretary Brenda Dann-Messier, Des

Moines Area Community College Town Hall. Washington, DC: US Department of Education.

Dee, Thomas S. 2012. "School Turnarounds: Evidence from the 2009 Stimulus." NBER Working Paper 17990.

Dee, Thomas S., and Brian A. Jacob. 2007. "Do High School Exit Exams Influence Educational Attainment or Labor Market Performance?" Chapter 6 in *Standards-Based Reform and the Poverty Gap: Lessons for No Child Left Behind*, edited by Adam Gamoran. Washington, DC: Brookings University Press.

Deming, David J., Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger. 2011. "School Choice, School Quality, and Postsecondary Attainment." NBER Working Paper 17438.

Dobbie, Will, and Roland G. Fryer, Jr. 2011a. "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone." *American Economic Journal: Applied Economics* 3(3): 158–87.

Dobbie, Will, and Roland G. Fryer, Jr. 2011b. "Getting beneath the Veil of Effective Schools: Evidence from New York City." NBER Working Paper 17632.

Dobbie, Will, and Roland G. Fryer, Jr. 2011c. "Exam High Schools and Academic Achievement: Evidence from New York City." NBER Working Paper 17286.

Elliott, Marc N., Lawrence M. Hanser, and Curtis L. Gilroy. 2002. "Career Academies: Additional Evidence of Positive Student Outcomes." *Journal of Education for Students Placed At Risk* 7(1): 71–90.

Figlio, David N., and Marianne E. Page. 2002. "School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?" *Journal of Urban Economics* 51(3): 497–514.

Fryer, Roland G., Jr. 2011a. "Injecting Successful Charter School Strategies into Traditional Public Schools: Early Results from an Experiment in Houston." NBER Working Paper 17494.

Fryer, Roland G., Jr. 2011b. "Racial Inequality in the 21st Century: The Declining Significance of Discrimination." *Handbook of Labor Economics*, vol. 4B, edited by O. Ashenfelter and D. Card, 855–971. Elsevier.

Fryer, Roland G., Jr., Steven D. Levitt, John List, and Sally Sadoff. 2012. "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment." NBER Working Paper 18237.

Gleason, Philip M., Christina Clark Tuttle, Brian Gill, Ira Nichols-Barrer, and Bing-ru The. 2012. "Do KIPP Schools Boost Student Achievement?" http://www.invalsi.it/invalsi/ri/improving_education/Papers/tuttle/155.pdf.

Goldammer, Christian. 2012. "Racial Gaps in Cognitive and Noncognitive Skills: The Asian Exception." http://lesacreduprintemps19.files.wordpress.com/2012/10/racial-gaps-in-cognitive-and-noncognitive-skills.pdf.

Goodman, Joshua S. 2012. "The Labor of Division: Returns to Compulsory Math Coursework." HKS Faculty Research Working Paper Series RWP12-032, John F. Kennedy School of Government, Harvard University.

Goolsbee, Austan, and Jonathan Guryan. 2006. "The Impact of Internet Subsidies in Schools." *Review of Economics and Statistics* 88(2): 336–47.

Hanushek, Eric A., John F. Kain, Steven G. Rivken, and Gregory F. Branch. 2007. "Charter School Quality and Parental Decision Making with School Choice." *Journal of Public Economics* 91(5–6): 823–48.

Hastings, Justine S., and Jeffrey M. Weinstein. 2008. "Information, School Choice, and Academic Achievement: Evidence from Two Experiments." *Quarterly Journal of Economics* 123(4): 1373–1414.

Heckman, James J., Jora Stixrud, and Sergio S. Urzua. 2006. "The Effect of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24(3): 411–82.

Hill, Patrick L., Brent W. Roberts, Jeffrey T. Grogger, Jonathan Guryan, and Karen Sixkiller. 2011. "Decreasing Delinquency, Criminal Behavior, and Recidivism by Intervening on Psychological Factors other than Cognitive Ability: A Review of the Interventional Literature." Chap. 8 in *Controllling Crime: Strategies and Tradeoffs*, edited by Philip J. Cook, Jens Ludwig, and Justin McCrary. National Bureau of Economic Research.

Hoxby, Caroline M., and Sonali Murarka. 2009. "Charter Schools in New York City: Who Enrolls and How They Affect their Students' Achievement." NBER Working Paper 14852.

Jacob, Brian A. 2002. "Where the Boys Aren't: Non-cognitive Skills, Returns to School and the Gender Gap in Higher Education. *Economics of Education Review* 21(6): 589–98.

Jacob, Brian A. 2011. "Do Principals Fire the Worst Teachers?" *Educational Evaluation and Policy Analysis* 33(4): 403–34.

Jacob, Brian A., and Jens Ludwig. 2008. "Improving Educational Outcomes for Poor Children." NBER Working Paper 14550.

Kemple, James J. 2008. *Career Academies: Long-term Impacts on Labor Market Outcomes, Educational Attainment, and Transitions to Adulthood.* New York: MDRC.

Kemple, James J., Corinne M. Herlihy, and Thomas J. Smith. 2005. *Making Progress toward Graduation: Evidence from the Talent Development High School Model.* New York: MDRC.

Kemple, James J., and Jason C. Snipes. 2000. *Career Academies: Impacts on Students' Engagement and Performance in High School.* New York: MDRC, Manpower Demonstration Research Corporation.

Koedel, Cory. 2008. "Teacher Quality and Dropout Outcomes in a Large, Urban School District." *Journal of Urban Economics* 64(3): 560–72.

Lankford, Hamilton, Susanna Loeb, and James Wyckoff. 2002. "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis." *Educational Evaluation and Policy Analysis* 24(1): 37–62.

Levesque, Karen, Jennifer Laird, Elisabeth Hensley, Susan P. Choy, Emily Forrest Cataldi, and Lisa Hudson. 2008. *Career and Technical Education in the United States: 1990 to 2005: Statistical Analysis Report.* (NCES 2008-035). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, US Department of Education.

Levitt, Steven D., John A. List, and Sally Sadoff. 2012. "The Effect of Performance-based Incentives on Educational Achievement: Evidence from a Randomized Experiment." Unpublished paper.

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff. 2012. "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance." NBER Working Paper 18165.

Maxfield, Myles, Allen Schirm, and Nuria Rodriguez-Planas. 2003. "The Quantum Opportunity Program Demonstration: Implementation and Short-term Impacts." Submitted to the US Department of Labor by Mathematica Policy Research. http://wdr.doleta.gov/research/FullText_Documents/2003-05.pdf.

Maxwell, Nan L., and Victor Rubin. 2002. "High School Career Academies and Post-Secondary Outcomes." *Economics of Education Review* 21(2): 137–52.

Murnane, Richard J. 2013. "U.S. High School Graduate Rates: Patterns and Explanations." NBER Working Paper 18701.

Neumark, David, and Donna Rothstein. 2006. "School-to-Career Programs and Transitions to Employment and Higher Education." *Economics of Education Review* 25(4): 374–93.

Papay, John P., Richard J. Murnane, and John B. Willett. 2010. "The Consequences of High School Exit Examinations for Low-Performing Urban Students: Evidence from Massachusetts." *Educational Evaluation and Policy Analysis* 32(1): 5–23.

Planty, Mike, and Jill F. DeVoe. 2005. *An Examination of the Conditions of School Facilities Attended by 10th-grade Students in 2002.* (NCES 2006–302). US Department of Education, National Center for Education Statistics. Washington, DC: US Government Printing Office.

Ross, Terris, Grace Kena, Amy Rathbun, Angelina KewalRamani, Jijun Zhang, Paul Kristapovich, and Eileen Manning. 2012. *Higher Education Gaps in Access and Persistence Study* (NCES 2012-046). US Department of Education, National Center for Education Statistics. Washington, DC: Government Printing Office.

Silverberg, Marsha, Elizabeth Warner, Michael Fong, and David Goodwin. *National Assessment of Vocational Education: Final Report to Congress.* Prepared by U.S. Department of Education Office of the Under Secretary Policy and Program Studies Service. Washington, DC: US Department of Education, Office of the Under Secretary Policy and Program Studies Service.

Snyder, Thomas D., and Sally A. Dillow. 2012. *Digest of Education Statistics 2011.* (NCES 2012-001). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, US Department of Education.

Socias, Miguel, Jay Chamber, Phil Esra, and Larisa Shambaugh. 2007. "The Distribution of Teaching and Learning Resources in California's Middle and High Schools." Issues & Answers REL 2007 – No. 023, Regional Educational Laboratory, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, U.S. Department of Education.

Strick, Betsy. 2009. "College Enrollment and Persistence of Preuss Alumni: Preuss and Comparison Students in the Classes of 2005 and 2006." http://create.ucsd.edu/_files/Strick11-13-2009.pdf.

Thernstrom, Abigail, and Stephan Thernstrom. 2004. *No Excuses: Closing the Racial Gap in Learning.* Simon & Schuster.

US Department of Education, Office of Vocational and Adult Education. 2012. "Investing in America's Future: A Blueprint for Transforming Career and Technical Education.

US Department of Education. 2012. Expanding successful career and technical education through career academies.

Walters, Christopher. 2012. "A Structural Model of Charter School Choice and Academic Achievement." http://economics.mit.edu/files/8429.

Whitman, David. 2008. *Sweating the Small Stuff: Inner-City Schools and the New Paternalism.* Thomas B. Fordham Institute.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 153–172

Beyond BA Blinders: Lessons from Occupational Colleges and Certificate Programs for Nontraditional Students[†]

James E. Rosenbaum and Janet Rosenbaum

Postsecondary education mostly focuses on the four-year BA degree. One often hears claims about how such degrees have "million dollar payoffs," an estimate which is based on average earnings for those who complete a four-year BA degree compared to those with only a high school education (for example, see how Couch 2012 presents the results of Carnevale, Rose, and Cheah 2011). While these "million dollar" claims are intended to be encouraging, they lead to predictable disappointments for many students who don't complete a BA or don't receive the higher pay.

Students' educational goals have dramatically increased in recent decades (Schneider and Stevenson 1999). Among high school graduates, 89 percent plan to get BA degrees (authors' analysis of data from the National Education Longitudinal Survey 2004), and over 80 percent actually enter college in the eight years after graduation (Adelman 2004). In keeping with these rising aspirations, enrollment has almost doubled over the last 40 years in public four-year colleges to almost 8 million students, with another 5 million in private four-year colleges (National Center for Education Statistics (NCES) 2012). However, many students pursue postsecondary education outside the context of four-year colleges. Enrollment has more than tripled over the last 40 years in public two-year colleges to over 7 million students,

■ *James E. Rosenbaum is Professor of Sociology, Education, and Social Policy and Faculty Fellow, Institute for Policy Research, Northwestern University, Evanston, Illinois. Janet Rosenbaum is Assistant Professor in the Department of Epidemiology and Biostatistics, School of Public Health, State University of New York Downstate Medical Center, Brooklyn, New York. Their email addresses are j-rosenbaum@northwestern.edu and janetrosenbaum @downstate.edu.*

[†] To access the Appendix, visit
http://dx.doi.org/10.1257/jep.27.2.153.

doi=10.1257/jep.27.2.153

and increased at a similar rate in private two-year colleges. Two-year colleges' total enrollment was a small fraction of four-year college enrollment in 1970 (37 percent), but it had grown to 58 percent in 2010 (NCES 2012, table 199). Comparing just public colleges, total enrollment at public two-year colleges was 51.9 percent of public four-year college enrollment in 1970, but it was 91.1 percent of public four-year enrollment in 2010. Community colleges have reduced the formal barriers of time, distance, and cost that often hindered students in the past from pursuing a postsecondary degree by offering convenient locations, flexible schedules, and low tuitions. Even low achievement in high school is not a barrier because many community colleges have open admissions policies.

Community colleges are often promoted as the first step toward the ultimate goal of a four-year degree. However, community colleges have extremely poor degree completion rates, with only 37 percent of students finishing an AA (associate's) or BA (bachelor's) degree in eight years. In contrast, there are indications that their private two-year college counterparts enroll similar students, but have 56 percent degree completion rates (Stephan, Rosenbaum, and Person 2009). While recent evidence estimates less of a difference for AA degrees, it does suggest substantially higher certificate completion in the private institutions (Deming, Goldin, and Katz 2012). In any case, private colleges challenge many of our preconceptions about college. They are less wedded to college traditions, which raises interesting questions. Do private colleges offering certificates or AA degrees use different procedures? Should community colleges consider some of these procedures to reduce student difficulties and improve their completion rates?

In this paper, we examine a specific kind of college known as *occupational colleges*—private accredited colleges that offer career preparation in occupational fields like health care, business, information technology, and others. Occupational colleges are accredited, and they offer certificates, associate's degrees, and sometimes bachelor's degrees. Occupational colleges are private but not selective, and they enroll many low-achieving and low-income students, who are typically funded by federal and state financial aid. In the next section of this paper, we offer a brief overview of occupational colleges, and we compare their degree completion rates with those of community colleges.

We then describe findings from a detailed study comparing occupational and community colleges (Rosenbaum, Deil-Amen, and Person 2006; Rosenbaum, Rosenbaum, Stephan, Foran, and Schuetz forthcoming). Our concern is not with promoting occupational colleges themselves, but with identifying their distinctive procedures and how they work. Our observations suggest that the occupational college sector is heterogeneous: that is, it includes some colleges with exemplary outcomes, and others that apparently are frauds. Rather than assess the average outcomes of this heterogeneous sector, we identify some better colleges in this sector, which design procedures to reduce difficulties students have with traditional college procedures. We find that these innovative procedures at the better occupational colleges are based on economic principles: enhanced incentives, structured choice, and investments in signals. In particular, we find that occupational colleges *enhance incentives* by creating quick payoffs, postponing obstacles, avoiding failure, and identifying unseen nonpecuniary rewards. They *structure choice* by creating pathways, timeslots, frequent monitoring, and mandatory advising. We find that besides investing in human capital, they also *invest in signals* by paying job placement staff to develop relationships that employers will trust to assure them that recommended graduates will not pose serious risks.

Of course, college incentives mostly come from labor market outcomes, and we find monetary and nonpecuniary gains resulting from certificate and AA programs of which many students are unaware. After reviewing evidence on monetary gains, we consider nonpecuniary gains, analyzing data from the National Longitudinal Study of Adolescent Health. We find that young adult workers ages 25–32 identify many nonpecuniary rewards of jobs such as job status, autonomy, relevance to later career, and others; that these rewards are more strongly associated with job satisfaction than are earnings; and that certificates and AA degrees have significant payoffs on these nonpecuniary rewards (compared to high school graduates). While community colleges assume that students are aware of all incentives for various credentials, students who seek the million dollar gains from a four-year BA may have difficulty seeing the nonpecuniary returns from a one- or two-year credential. In contrast, occupational colleges bolster students' incentives by identifying job rewards students don't see at the time but typically appreciate after they complete their credential.

For many community college students, earning the more likely, quick sub-BA credential—perhaps followed by a four-year degree in the future—will be preferable to the relatively unlikely pathway from a community college program directly to a four-year BA. In sum, this paper suggests that nontraditional colleges and nontraditional credentials (certificates and AA degrees) deserve much closer attention from researchers, policymakers, and students.

As faculty at institutions of higher education, this comparison may help us examine college procedures that we take for granted and help us perceive unnecessary obstacles they impose. Our aim is to understand nontraditional procedures, and how they implement economic insights and nontraditional goals to assist their students' success. Although "college" is usually assumed to be a single entity, we discover that some colleges use radically different procedures that enhance incentives, identify nonpecuniary incentives, structure choice, and invest in trusted signals.

What are Occupational Colleges?

Occupational colleges are private institutions, both for-profit and nonprofit, that confer accredited degrees and certificates in occupational fields. This category includes large corporations like DeVry, Phoenix, and others as well as small colleges that offer preparation in many mid-skilled jobs in a variety of fields: for example, health sciences, consumer services, business, protective services, computer information sciences, engineering technicians, marketing, and legal services. Although annual tuitions at occupational colleges are much higher than tax-subsidized community colleges, federal and state grants cover some of the added costs, and occupational colleges help students obtain these grants (which require complex paperwork). Critics contend that for-profit colleges will be deceptive, and Tierney (2012, p. 155, 159) agrees that some are, but he also argues that "the need for a better educated workforce will continue to grow . . . The reality is that at a time when states need to increase college participation [to meet labor market needs], the public sectors are cutting back. Without the active involvement of the private non- and for-profit sectors, there is simply no way to reach the educational levels that will enable the United States to be a leader in college access, completion, and attainment." At the same time that public community colleges are facing budget cutbacks, for-profit and nonprofit private colleges are growing.

Analyzing the National Education Longitudinal Survey of 1988 (NELS), Stephan, Rosenbaum, and Person (2009) found that occupational colleges enroll similar students as community colleges but their degree completion rates are 20 percentage points higher. These results are robust to different methods of estimation. For example, limiting the sample to comparable students (using propensity-weighted regression), the estimated difference in overall attainment rates is 20 percentage points or more. (This data source is not useful for comparing earnings, because too few people reported wages from occupational colleges.)

However, our purpose in this essay is not to characterize or defend the average performance of this sector. There have been a number of highly publicized cases of institutions that provided little for their students and, in some cases, even engaged in outright fraud. Even if average outcomes of for-profit colleges are no better than community colleges—and actually, the evidence seems to show that their graduation rates are better—occupational colleges are a heterogenous group and averages should not prevent us from looking at the better institutions. In the next section, we focus on describing procedures used by some of the better private occupational colleges, what economic principles they may use, and whether they suggest procedures for reforming traditional community colleges.

Lessons from Occupational Colleges: Six Nontraditional Procedures

Many who are involved in higher education, whether as participants or as researchers, assume implicitly or explicitly that traditional four-year BA degrees are the only meaningful goal of college, and that traditional college procedures are the only conceivable procedures. For example, community colleges require over 60 percent of students to take remedial courses in the hope of transforming them into traditional students who can meet traditional academic demands, regardless of how long this takes. This BA-centric approach has been tried many times, and it has failed repeatedly (Bailey, Jeong, and Cho 2010). These failures lead some observers to believe that disadvantaged students inherently lack the academic ability to succeed in college (Murray 2008).

However, blaming this failure solely on students' abilities to master a specific academic curriculum overlooks the many nonacademic demands that colleges make on students. Traditional college procedures require students to have culturally specific information and financial support to succeed in college: college and labor market knowledge, schedule flexibility, and resources to persist without payoffs for four years—and often much longer. In interviews, community college students report a wide variety of mistakes in course choices, time allocation, and degree plans. Such mistakes threaten college persistence by creating serious problems for students: courses without credits, credits without credentials, and credentials without job payoffs (Rosenbaum, Rosenbaum, Stephan, Foran, and Schuetz forthcoming). These problems arise from students' difficulties with traditional college procedures, but, as we will argue, the better occupational colleges use different procedures that reduce these difficulties.

Are there lessons from occupational colleges that community colleges could use to improve their completion rates? In Rosenbaum, Deil-Amen, and Person (2006), we analyzed the institutional procedures in seven community colleges and seven occupational colleges in the Chicago metropolitan area. Because our aim was to discover alternative procedures, rather than to evaluate average outcomes, we purposely chose better occupational colleges that we expected would be more effective at graduating students. Our group of seven private colleges includes three nonprofits and four for-profits, three of which are part of for-profit national chains across the United States.

Based on our observations, interviews with college staff and students, and surveys of 4,000 students, we discovered specific problems that students experience in community colleges that are less common in occupational colleges. Our study led to six lessons that broaden our conceptions of college procedures, which we identify below. Moreover, more recent research discovers a broad array of job rewards, many of which young adults find more satisfying than earnings. We find that occupational colleges apply economic principles to their procedures so students are more likely to see incentives, make good choices, and get trusted labor market signals.

Lesson 1: Quicker Successes with Sub-BA Credentials

Most students enter community college to pursue a "four-year BA degree." However, these plans lead to all-too-predictable disappointments. Among community college students pursuing a BA, only 4 percent complete a BA in four years (Stephan 2010, based on data from the National Education Longitudinal Survey). Another 8 percent complete a BA degree in five years, and yet another 16 percent do so in 6–8 years. In sum, only 28 percent get a BA in eight years. Over the next 20 years, another 10 percent of students may complete a BA degree (based on the authors' unpublished analyses from the National Longitudinal Survey of Youth 1979), but such long periods before completion will inevitably leave fewer years for payoffs in earnings or other nonpecuniary forms.

Thus, for the average community college student, the "four-year BA" is nearly a myth and the BA is still rare in eight years. The BA is likely to become an increasingly weaker incentive as students discover these outcomes.

In contrast, occupational colleges enhance incentives by offering quicker credentials and making them an automatic part of the curriculum path. All seven of the occupational colleges in our study offered certificates in one year and associate's degrees in two years, and these credentials were conferred automatically along the way for those who were continuing on to a BA. While these credentials are also offered in community colleges, they are deemphasized, and many students don't realize they can get a quick certificate with job payoffs on the way to a BA degree. In occupational colleges, students may still pursue a BA, but along the way they earn shorter credentials that have higher completion rates and may lead to high-demand, mid-skilled jobs by age 25, which is a good way to start a career and likely boosts students' confidence to go further.

Lesson 2: Postpone Remediation, Avoid an Obstacle

Community colleges have opened their doors to new kinds of students, including many with previously low educational achievement (Long and Riley 2007). Nonetheless, these students are encouraged to plan BA goals. However, BA goals require college-level academic skills, which means that low-achieving students must take remedial courses. While these students take classes in college buildings and pay college tuition, they are not in "college classes."

Over 60 percent of students take remedial courses, which are high school–level courses that give no college credit (Adelman 2004), and many students take several remedial courses (Rosenbaum, Deil-Amen, and Person 2006).

While community colleges frontload remedial courses to prepare students for later BA courses, that policy only works if students complete the remedial courses—which most students do not. On average, about 46 percent of students complete the reading sequence, and 33 percent complete the math sequence. For students referred to the lowest-level remedial courses, completion rates are abysmal: Bailey, Jeong, and Cho (2010) report just 29 percent complete the remedial reading sequence and 17 percent complete the remedial math sequence. Very rarely does anyone warn students in low-level remedial courses that the math sequence only has a 17 percent success rate. Indeed, remedial classes are often concealed behind the euphemism of "developmental education," which many students don't understand. The incentives to continue in college may be undermined as students discover that they are getting no credits for some courses and that remedial courses usually don't work.

This is all the more unfortunate since, in fact, "college-level academic skills" are not necessary to benefit from postsecondary education. In our interviews, faculty at both community and occupational colleges report that ninth-grade math and reading skills are sufficient for completing certificates in many occupations, including high-demand fields in computer networking, medical technicians, allied health, and accounting (Rosenbaum, Cepa, and Rosenbaum 2013). For underprepared students, these alternatives to the BA do not require immediate remediation, while the traditional BA route, with its frontloaded remedial courses, is a major obstacle.

Occupational colleges enhance student incentives by avoiding or postponing obstacles that cause delays or failures, like remedial noncredit courses. Students can earn certificates and AA degrees without remedial requirements. Any academic remedial lessons that are needed are integrated into occupational courses, and remediation is gradually provided as needed, without becoming a separate noncredit obstacle.

Lesson 3: Degree Ladders, Not a "Fail-First" Sequence

Most community colleges not only postpone success by emphasizing remedial courses, they also frontload failure. In the national Beginning Postsecondary Survey, researchers found what we call "a fail-first sequence:" that is, 42 percent of community college students drop out in the first year, 50 percent return to community college, and 53 percent then *drop out again* (Horn 1999). Only 14 percent of early dropouts acquire any credential.

"Fail first" is a costly method: it rarely leads to a credential, even among students who have the qualifications for a certificate or AA degree. Advisors often don't tell students about these quicker interim credentials they could get while pursuing a BA. Certificates and AA degrees take less time, require fewer remedial courses, and have labor market value, and many credits will count for BAs (Rosenbaum and Cepa, 2013, discussed below). The fail-first sequence prevents students from seeing how they can succeed in college and undermines their incentives to continue trying.

In contrast, occupational colleges enhance incentives by creating "incremental success," essentially reversing the fail-first sequence. In this approach, early courses are relatively easy, engaging, career relevant, and teach skills of general value, "the new basic skills," which are computer skills, soft skills, and task management (Murnane and Levy 1996). Within twelve months, students can earn certificates in high-demand fields like health, computers, business, and other fields that often lead to better jobs (Carnevale, Rose, and Cheah 2011), and this may even permit students to begin gaining relevant job experience during college. For students who have never done well in school, these sub-BA credentials avoid early failure, give students a quick payoff and confidence that they can succeed in college, and offer the first step on a degree ladder to AA and BA degrees. Community colleges could greatly increase their rates for students completing some kind of credential if they placed more emphasis on degree ladders.

Traditionally, certificates were the end of an individual's postsecondary education, but not anymore (Gill and Leigh 2004). Over 30 percent of people with certificates also get AA or BA degrees (Carnevale, Rose, and Hanson 2012). Some certificate courses count for the BA; in our community college sample: 47 percent of certificate courses count toward a BA, although the portion varies across colleges and majors (Rosenbaum and Cepa 2013). Getting a certificate on the way to a BA may add a little more time to the BA timetable, but this extra time is insurance that gives short-term job payoffs, access to a career, career-related job experience, and some credits toward AAs and BAs (Schuetz, Rosenbaum, Foran, and Cepa 2012). In particular, students at risk of dropping out might first aim for certificates, which can be completed in one year versus the six or more years it often takes to finish a BA (Bound, Hershbein, and Long 2009; Stephan 2010).

Lesson 4: Structured Curriculum Pathways and Timeslots

College traditions encourage an exploratory process where students choose their own courses. This approach can work fairly well for many traditional four-year college students whose parents will support them through four years or more and can offer advice and guidance. But for disadvantaged students without such financial support and guidance, this exploratory approach offers a bewildering plethora of choices, which many students report to be highly confusing, and it can increase the risks of not completing any credential at all. Among high school graduates in 1992, 8 percent got associate's degrees by the year 2000, and another 10 percent had enough credits for an AA, but no degree (60+ credits, Adelman 2004). Students have difficulty knowing which courses count so they take the wrong ones and fail to get degrees. The labor market mostly rewards credentials, not isolated credits, so 10 percent of this nation's high school graduates get *credits without credentials*, which will yield little or no payoff (Grubb 1996, 2002).

By tradition, college courses fill a patchwork of timeslots in a week. For students who live off campus, such a schedule may require commuting for each class and adjusting to schedule changes every semester. Many students report that each semester's unpredictable timeslots for required courses can conflict with work or childcare schedules, making the choice process more difficult and sometimes preventing required courses, which lengthens degree timetables.

To reduce these difficulties, occupational colleges impose structure on students' choices. They offer structured curricula in pre-set time slots. Like a package-deal vacation, students at such colleges choose their career goals, and then the college packages all the details at the outset, so students know at the beginning which classes to take and what timeslots to keep free. Echoing the findings of Schwartz (2004) about the disadvantages of too many choices, one student told us, "I am balancing childcare, work hours, shopping, cooking, cleaning, and college. I don't want more choices." While policymakers often assume that limiting choice is undesirable, for community college students worried about costly mistakes, structure can help them make dependable progress towards their goals.

Lesson 5: Mandatory Advising and Monitoring Student Progress

Community colleges require students to make complex decisions about large numbers of courses and programs. Many of these students can't get advice from their parents—because their parents didn't attend college. As a result, students make many mistakes. They choose courses that are too easy, and so don't make sufficient progress; they choose courses that are too difficult, and face a greater chance of failure; or they choose courses that don't give the right kind of credit for their program, degree, transfer, or employment. Students miss deadlines, underestimate degree timetables, and even discover that early credits can expire if they progress too slowly. The rules for degree completion are often complex and confusing (Rosenbaum and Cepa 2013).

In contrast, occupational colleges structure the choice process by monitoring students' progress and offering frequent mandatory advising (Rosenbaum, Deli-Amen, and Person 2006). A monitoring system makes sure that students take the right courses and keeps track of absences, grades, and teacher concerns. Advisors quickly contact students before problems become serious. Unlike the student-initiated advising that is common in community colleges, occupational colleges make advising meetings mandatory several times each term and target these meetings at preventing common student mistakes. Unlike individual advising in community colleges, advising occurs in small groups of students in the same program (that is, in cohorts), and one student will often ask questions that had not yet occurred to others.

Lesson 6: School-Directed Job Placement

Traditionally, college degrees were expected to guarantee good jobs. That may have been true several decades ago, but no longer. Nonetheless, many colleges still operate as if it were true. At the community colleges we studied, career services offices offered optional workshops in interviewing and resume preparation to a few students, but these workshops were not aggressively marketed, perhaps because the offices had too few staff to handle a sizable influx of students. (One advisor at a community college discouraged the student newspaper from mentioning his work because the career office could not handle more students.) At many community colleges, the career services offices don't have time to interact with employers. They post ads on bulletin boards or websites for job openings, and many of these jobs are unrelated to college programs.

In contrast, occupational colleges do job placement, and they *invest in signals* employers will value. Job placement services are required, structured, and comprehensive. While community colleges offer a few students optional workshops to make prettier resumes, in occupational colleges, job placement staff help students translate course titles into work-relevant skills that employers will recognize, they provide extensive job search assistance to every graduate, they closely oversee and advise the job search process, they advise students on self-presentation, and they identify skill-relevant job openings. They are savvy. They understand the spatial-mismatch hypothesis (Kain 1968) and urge students to consider residential moves to improve employment prospects.

While community colleges focus mostly on improving human capital value, occupational colleges also invest in improving the value of their *signals* about graduates. Job placement staff invest time in developing long-term relationships with employers. Employers trust the recommendations and ratings of the placement staff because they know that staff won't jeopardize their future credibility. Employers can be confident that they know what risks they are taking in hiring applicants and what assets they are getting (Rosenbaum, Deil-Amen, Person 2006).

Monetary Payoffs from AA Degrees and Certificates

Of course, the main college incentives are the payoffs associated with college credentials, and college procedures can improve students' awareness of incentives they will value after they graduate. We focus on payoffs after college, and how colleges can enhance incentives by improving awareness of future payoffs that students don't anticipate. While most students know the monetary payoffs to BA degrees, the returns to sub-BA credentials are poorly understood. Indeed, much research ignores these credentials and focuses on "years of education," which may not match well with credentials. When BA degrees take eight years, we don't expect the recipients to have a greater payoff than a four-year BA; in reality, they may have less. While students who get one year of college but no credential may have little earnings payoff (Grubb 2002), one-year certificates in some fields have substantial payoffs (Jacobson and Mokher 2009).

Some public service ads have proclaimed that BA degrees have a $1 million payoff in lifetime earnings, although some recent estimates are much lower (Lederman 2008). While BA degrees lead to higher median earnings than certificates and AAs, earnings are widely dispersed and overlap; for example, 24 percent of certificate graduates have higher earnings than the median earnings of BA graduates (Carnevale, Rose, and Hanson 2012; see also Baum, Ma, and Payea 2010; Jacobson and Mokher 2009).

In regression analyses of adults of all ages, Carnevale (2012) find that BAs, AAs, and certificates all have significantly higher earnings than high school graduates (72, 47, and 19 percent, respectively, after controls for gender, ethnicity, experience, and experience squared), and the certificate payoff is larger for males than for females (22 versus 15 percent, see their table A1). Focusing on young adults, Carnevale finds that BAs, AAs, and certificates all have significantly higher earnings than high school graduates (67, 39, and 33 percent respectively, after controls for the same demographics and also academic skills; see their table A3). The higher payoff for certificates in the younger sample may indicate that certificates have greater payoffs either for young adults or in the current decade (compared with earlier decades, when fewer jobs required certificates).

Of course, even when payoffs for different degree levels are adjusted for observable factors like experience or demographic factors, those who complete a degree may well be more likely to have valuable unobservable factors like persistence or social skills. As a result, the estimated gains to finishing a degree are a mixture of the learning from the degree itself and these unobservable factors and thus are biased upward. This paper is not the place to try to sort out these well-known endogeneity problems. But we would note that the students who start at occupational colleges and community colleges are observationally similar, and the socioeconomic status, high school grade point average, and achievement test distributions of the two groups of students strongly overlap (Stephan, Rosenbaum, Person 2009). Since students at both types of college are continuing on toward two-year degrees, they are likely to be more similar in some unobservable characteristics as well. Moreover, when students' progress is closely monitored and they meet in frequent mandatory counseling sessions, or when graduating students get personal job-search support from job placement staff who are well connected to employers, it is hard to ignore the possibility that these high-contact institutional procedures at occupational colleges contribute to strong incentives, and thus to positive completion rates and employment outcomes.

Nonpecuniary Gains from AA Degrees and Certificates

Occupational colleges also enhance incentives by identifying nonpecuniary payoffs that students usually don't anticipate, but they will value in the future after they graduate. While many students and policymakers focus on earnings payoffs from postsecondary education, our interviews indicate that community college students often ignore nonpecuniary job rewards like autonomy, career relevance, job status, healthy work conditions, and others. Although prior research has found that older adults value these job rewards, young college students may rightly dismiss the ratings of older adults, whose careers were shaped in very different labor market conditions. However, young students might be impressed if they were to learn that young adults value nonpecuniary rewards. Thus, it seems important to consider the extent to which young working adults value nonpecuniary job rewards, and whether some job rewards are strongly related to certificates and associate's degree credentials. If young workers value these nonpecuniary rewards, then colleges can help students recognize their value and increase students' incentives to pursue such credentials.

In occupational colleges, job placement staff report that they urge students to consider nonpecuniary outcomes when they are choosing jobs. As economic theory predicts, they report that many high-paid jobs are not necessarily "good jobs." High earnings may compensate for jobs having the serious disadvantages that we call the five Ds: dangerous, demanding (strenuous), disruptive (unpredictable work shifts), dead end, and deceptive (for example, jobs that promise large sales commissions that are unlikely to materialize). These staff warn graduates to think twice before taking such jobs. Instead, they urge graduates to seek nonpecuniary rewards—jobs that offer career preparation, on-the-job training, specialized skills, and advancement (Redline and Rosenbaum 2010). They contend that although young students are rarely aware of these nonpecuniary rewards, they will value them later when they are in the work world. This message often resonates with young adults, if not always with college students. A study of 69 young adults (27 year-olds) found that many were "disappointed in their career development because they had not achieved a career-type job" (Mortimer, Zimmer-Gembeck, Holmes, and Shanahan 2002, p. 447).

To investigate more systematically whether young adults actually find such rewards satisfying and gain such rewards from various postsecondary educational credentials, Janet Rosenbaum (2013) analyzed data from the National Longitudinal Study of Adolescent Health (Add Health).[1] This is "a longitudinal study of a nationally representative sample of adolescents in grades 7–12 in the United States during the 1994–95 school year. The Add Health cohort has been followed into young adulthood with four in-home interviews, the most recent in 2008, when the sample was aged 25–32" (for more on Add Health, see http://www.cpc.unc.edu/projects/addhealth).

The Add Health study asks young adult workers about many job attributes: perceived status, repetitive tasks, career-related, career preparation, and autonomy. These questions address the career concepts that counselors and students express in

[1] Add Health is a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by grant P01-HD31921 from the National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (http://www.cpc.unc.edu/addhealth).

Table 1
Correlation between Job Satisfaction and Job Rewards in each Education Level and for All Students

	Highest degree					
Job rewards	*High school*	*Certificate*	*AA*	*BA*	*Post-BA*	*All*
Personal earnings	.11	.17	.07	.10	.02	.10
Perceived status	.21	.20	.22	.22	.11	.21
Job autonomy	.29	.37	.32	.33	.33	.32
Job not repetitive	.16	.14	.14	.19	.11	.17
Job related to career goal	.31	.32	.36	.35	.28	.33
Job part of career	.35	.36	.35	.38	.37	.37
Achieved desired educational level	.12	.11	.11	.12	.01	.12
N	(4,470)	(938)	(1,058)	(2,838)	(1,155)	(10,459)

open-ended interviews. These findings can also be interpreted as showing whether young adults who get sub-BA credentials get jobs with career-related attributes. For instance, certificates and applied associate's degrees can lead to mid-skill jobs in various fields like medical technology, computer networking, or paralegal jobs, which in turn may offer skill increases and career advancements.

Although these mid-skill jobs are not as highly paid as professional jobs, they do share some of the attributes of these jobs, as defined by sociologist Richard Ingersol (2004): "rigorous training and licensing requirements, clear standards for practice, substantial workplace responsibility, positive working conditions, an active professional organization or association." They may also have career ladders that build on their specialization, particularly when combined with higher credentials. For instance, computer networking may give valuable experience for becoming a network administrator, although additional credentials may also be required. Some counselors tell students to attend college to get a career, and students themselves often say they are seeking a "career, not just a job" and "I want a job I can stay in for my whole life" (Rosenbaum and Cepa 2013).

As a starting point, we examine whether young adults find these nonpecuniary job attributes rewarding. We look at how young working adults, ages 25–32 in this data, rate their jobs on earnings and other job attributes and whether jobs with such attributes are satisfying. We find that many job attributes are strongly correlated with job satisfaction, as shown in the last column of Table 1, and these correlation coefficients also persist at comparable levels in each education category, indicating that the relationships are not mediated by education (even though education is associated with many of these job attributes). We infer that young adults consider these attributes as rewards, since they are strongly associated with overall job satisfaction, even more strongly than earnings.

Then we use each of the job-reward variables as a series of dependent variables, and run regressions in which the postsecondary credentials shown in the columns are explanatory variables, along with controls for many individual attributes, listed in the note beneath Table 2. Table 2 shows that compared with high school graduates, those with certificates and AA degrees see an increase in many job rewards, including perceived status, job satisfaction, autonomy, career relevance, career preparation, and whether their job is part of a career. AAs (but not certificates) are statistically significantly associated with increases in earnings, fringe benefits, and desk jobs, and decreases in various job demands like whether the job has characteristics like physically hard, irregular hours, night shift, or repetitive. Indeed, inspecting this long array of job rewards, we see that AA degrees have payoffs on virtually all job rewards for which BAs have payoffs.

In sum, besides certificates' earnings payoffs, certificate graduates get some of the same nonpecuniary job rewards as BAs, and sometimes at the same magnitude as BAs (for career related and career preparation). While the earnings advantage for those with an AA degree is much smaller than BAs, AAs get nearly all of the nonpecuniary job rewards as BAs, sometimes at a similar magnitude (autonomy, health benefits, career preparation), but often less (earnings, strenuous, night shift).

There is some evidence on other nonpecuniary benefits of AA degrees and credentials. Janet Rosenbaum (2012) identified health payoffs to greater education, although these credentials may have these health payoffs through access to better job conditions. For example, Presser (2005) has identified negative psychosocial and health consequences from nontraditional work schedules. Sleep problems have been cited as a cause of disparities in morbidity and mortality throughout the life course (Grandner, Hale, Moore, and Patel 2010). A substantial number of jobs with nontraditional schedules—either night shifts or irregular shifts—is a relatively new development in the US labor market. AA graduates are less likely to work night shifts or irregular shifts than high school graduates, but certificate graduates show no such advantage. Preliminary analyses find that after exact matching on job qualities, including shift work, sub-BA and BA graduates do not differ in sleep problems. Sub-BA credentials may improve health status compared to high school graduates, but parity with BAs may require better job conditions, such as avoiding night shifts, (Janet Rosenbaum 2012; Grandner 2010).

Are certificates preparing youth for narrow vocations with a short shelf life, diverting them from more substantial skill sets that would help them throughout their careers? There are many indications that the payoffs from certificates justify their investments. First, certificates are not an educational dead end, given that over 30 percent of people with certificates also get college degrees (Carnevale, Rose, and Hanson 2012). Second, for youth struggling to escape dead-end jobs (Doeringer and Piore 1971), these results suggest that certificates lead to career-related jobs, career preparation, and perhaps career advancement. Third, even if youth don't get further education, one-year certificates cost less (in time and money) and pose lower risks of interruption than BA degrees that often take 6–8 years (Bound, Hersbein, and Long 2009; Stephan 2010). Certificates are more likely among low-income youth (Carnevale, Rose, and Hanson 2012), but this segmentation is a problem for US society more generally, not the fault of colleges. Of course, certificates are not

Table 2

Multivariate Regressions Show Job Rewards for Different Educational Levels, Relative to High School Graduates

	Certificate	*AA*	*BA*	*>BA*
Poisson regression				
Job relates to career				
Unrelated	0.59****	0.76****	0.51****	0.25****
	(0.52, 0.67)	(0.68, 0.84)	(0.47, 0.56)	(0.21, 0.31)
Preparation	1.35****	1.18**	1.22****	1.08
	(1.21, 1.50)	(1.06, 1.32)	(1.12, 1.33)	(0.96, 1.22)
Part of career	1.36****	1.35****	1.60****	2.09****
	(1.25, 1.49)	(1.24, 1.47)	(1.50, 1.71)	(1.94, 2.24)
Benefits offered				
Health benefits	1.01	1.13****	1.18****	1.27****
	(0.97, 1.06)	(1.09, 1.17)	(1.15, 1.22)	(1.22, 1.31)
Retirement benefits	1.03	1.15****	1.25****	1.34****
	(0.97, 1.09)	(1.09, 1.20)	(1.21, 1.30)	(1.29, 1.40)
Vacation benefits	1.01	1.11****	1.17****	1.22****
	(0.97, 1.06)	(1.07, 1.15)	(1.13, 1.20)	(1.18, 1.27)
Job conditions				
Day shift	0.99	1.08***	1.23****	1.23****
	(0.94, 1.04)	(1.03, 1.13)	(1.19, 1.27)	(1.18, 1.29)
Irregular hours	1.06	0.83**	0.77****	0.78***
	(0.93, 1.22)	(0.72, 0.96)	(0.69, 0.86)	(0.68, 0.90)
Work hard physically	0.90	0.58****	0.28****	0.12****
	(0.75, 1.09)	(0.46, 0.72)	(0.22, 0.35)	(0.07, 0.21)
Work desk job	0.96	1.20****	1.71****	1.44****
	(0.85, 1.07)	(1.09, 1.32)	(1.60,1.83)	(1.32, 1.58)
Supervise managers	0.86	1.18+	1.28**	1.05
	(0.68, 1.09)	(0.97, 1.44)	(1.09, 1.50)	(0.84, 1.32)
Supervise others	0.97	1.00	0.96	1.08
	(0.86, 1.10)	(0.89, 1.12)	(0.88, 1.05)	(0.97, 1.21)
Ordinary least squares regression				
Personal earnings	2.25	4.35**	12.9****	19.8****
	(–0.79, 5.29)	(1.43, 7.27)	(10.6, 15.1)	(16.7, 22.8)
Perceived status (0–10)	0.13*	0.27****	0.86****	1.48****
	(0.02, 0.24)	(0.16, 0.37)	(0.78, 0.94)	(1.37, 1.59)
Job satisfaction	0.03**	0.02**	0.01*	0.05****
	(0.01, 0.04)	(0.01, 0.04)	(0.00. 0.03)	(0.03, 0.07)
Job autonomy	0.05****	0.03**	0.04****	0.07****
	(0.03, 0.07)	(0.01, 0.05)	(0.02, 0.06)	(0.05, 0.09)
Job repetitive	–0.01	–0.04****	–0.13****	–0.19****
	(–0.03, 0.01)	(–0.06, –0.02)	(–0.15,–0.12)	(–0.22, –0.17)
Number times fired	–0.01	–0.09*	–0.19****	–0.32****
	(–0.11, 0.09)	(–0.19, 0.00)	(–0.27, –0.12)	(–0.42, –0.22)

Notes: Multivariate regression results ($n = 10{,}582$). Columns correspond to educational levels, and rows correspond to employment outcomes. The entries correspond to the multivariate regression coefficient predicting the outcome from the educational level. Control variables: demographics (race/ethnicity (black, Latino, Asian), gender); educational factors (grade average, test score, grades not reported by respondent); acculturation (nativity, parent nativity, speak English versus another language at home); and parent's socioeconomic status (parent's self-reported educational level, household income, and whether they have enough money to pay bills.)

+, *, **, ***, and **** correspond to .10, .05, .01, .001, and .0001 levels of significance.

for everyone. However, a fairly broad range of high school students and their advisors should consider occupational colleges and certificate programs as potentially useful pathways to higher education. In contrast to the abysmal BA completion rates in community colleges, certificates provide quick high-odds payoffs on the way to a more distant and less-certain BA degree.

Conclusion

Our aim is to broaden the conception of possible college procedures and the range of job rewards. Most of us wear BA blinders. For anyone who attended colleges and universities, traditional practices that target BA goals are taken for granted, and it is hard to imagine other procedures that might pose fewer obstacles to disadvantaged students. Moreover, we usually focus on earnings outcomes, ignoring other job rewards that might offer important incentives, particularly to youths starting their careers.

In studying occupational colleges, we discovered six nontraditional procedures that differ from traditional college practices and make novel use of economic mechanisms: incentives, choice, and signals. Consistent with neoclassical economic theory that stresses incentives, occupational colleges *enhance incentives* by providing quick milestones and valued payoffs, and they identify unseen nonpecuniary job rewards that students will appreciate later. Consistent with some recent research (Schwartz 2004; Thaler and Sunstein 2008), occupational colleges *structure choice* by offering a few options, package deals, and substantial supervision (through frequent mandatory advising and monitoring). Consistent with signaling theory (Stigler 1961; Spence 1973), occupational colleges *invest in trusted signals*, not just in human capital. Just as prep schools invest in trusted signals by encouraging counselors to build long-term relationships with selective colleges (Persell and Cookson 1985), occupational colleges invest in the credibility of their signals by encouraging job placement staff to build trusted relationships and reputations with employers. These investments make the college's signals about their graduates trustworthy and reduce employers' uncertainties about graduates' hard-to-measure attributes. Even if occupational colleges' graduates had only "adequate" human capital, employers can depend on trusted placement staff not to recommend students who will seriously disappoint.

Particularly for middle- and lower-achieving high school students, traditional community college procedures offer a bewildering array of confusing choices and uncertain incentives. They offer many programs, multiple credentials for each, and uncertain job payoffs, so students have difficulty figuring out what outcomes are desirable for them. These problems only become worse for students who have achievement limitations, weekly schedule constraints, limited budgets, or uncertain timelines. Such complexity and confusion often lead to poor choices (Schwartz 2004). Many community college students complain that they can't anticipate the implications of their decisions.

Table 3
Traditional Procedures at Community Colleges and Nontraditional Procedures at Occupational Colleges

Community colleges: Traditional procedures	*Occupational colleges: Nontraditional procedures*
Deferred 6+ year payoffs	Quick payoffs
Early obstacles (remedial)	Postponed obstacles
Direct BA goal	Incremental success degree ladders
Complex choices & schedules	Package deal pathways & preset timeslots
Unassisted course choices	Mandatory advising & monitored progress
Self-directed job search	College-guided job choice & job search

In contrast, occupational colleges reduce complexity and the associated risks by alternative procedures, shown in Table 3. They offer a relatively few package-deal options, where each option offers high odds of desirable outcomes. Students make a single choice—their program goal—and this choice then illuminates a specific pathway toward completion, reducing mistakes and increasing the odds of success. Moreover, when 80 percent of entering students have BA plans (Deming, Goldin, and Katz 2012, Table 1), community colleges encourage them to focus on BA credentials, even if they don't understand the risks and costs. In contrast, even though, in Deming's sample, 64 percent of for-profit college students expect to earn BA or higher, these for-profit colleges also encourage students to aim for interim credential goals, which often have monetary and nonpecuniary job payoffs along the way to the BA. While community colleges emphasize remedial courses which delay payoffs and emphasize academic skills that are weaknesses for many students, occupational colleges also emphasize job skills and soft skills, which employers often value more than academics (Zemsky 1994). While community colleges focus on improving human capital, occupational colleges may *increase the payoffs* to human capital by investing in signals that employers trust. In effect, occupational colleges design processes that simplify choices, frontload successes, maximize payoffs, and reduce mistakes. Community colleges could adapt many of these procedures, and they might gain similar benefits.

Some of these procedures might be implemented in community colleges, and there are already a few examples. Some Tennessee community colleges have structured choice with package-deal programs that specify all courses and minimize noncredit remedial coursework (Carnevale, Rose, and Hanson 2012, p. 17). Some community colleges have improved supervision with computerized monitoring of student progress, and they notify students about teacher concerns or wrong class choices (for example, http://www.cuny.edu/asap). Some community colleges have structured incentives, and they identify certificates with dependable job payoffs (for example, Ivy Tech in Indiana). Some community college program staff invest time in employer contacts which enable them to provide evaluations of their graduates that employers trust (Rosenbaum, Deil-Amen, and Person 2006). However, many of these reforms are piecemeal, initiated by a few staff in a few programs. Most community colleges seem wedded to traditional conceptions of unstructured choices, vague unstructured incentives, and signaling graduates' competencies by credentials alone, which employers often mistrust.

These findings have implications for research, for college policies, and for public policy. From a research perspective, our findings suggest that community college students rarely perceive credential options, their odds of success, or nonpecuniary job rewards, so these students often make uninformed choices. They are deprived of seeing the full range of potential rewards—the nonpecuniary incentives they will later value, for which they would be willing to exert effort. Research can identify these alternatives and future payoffs, so that students can see a wider range of incentives.

In particular, research needs to consider how students are affected when they face predictable disappointments pursuing BAs, and whether they might be more motivated if they began their postsecondary experience with an *incremental success model*: where their first goals are quick credentials and nonpecuniary payoffs, particularly ones that lead to careers. More generally, while BA degrees and earnings may provide dependable incentives for high-achieving students, they may be long-shots for other students. By conceiving of credentials and outcomes more broadly, research may discover how incentives and outcomes can be improved for all students in community colleges, and perhaps in other schools and colleges.

Our results suggest that institutions can shape their procedures to make incentives stronger: clearer, quicker, more certain, and more dependable. Indeed, some of these lessons may also hold true in high schools targeting low-achieving students (Rosenbaum and Becker 2011). While community colleges cannot offer that single model exclusively, they could make it an option, and quick credentials with high odds of success would create stronger incentives for low-achieving students (Jones 2011; Rosenbaum, Deil-Amen, and Person 2006). The all-or-nothing BA degree can remain an option, but students with low-achievement need to know about interim credentials with better odds and quicker timetables on the way to pursuing a BA degree.

When education reformers proclaim that "Time is the enemy" (Complete College America 2011), they are noting that many students have a limited window of time to complete a postsecondary credential of some sort, before their time and money run out. Long-duration credentials, noncredit courses, early failures, mistaken courses, and the wrong job search strategies are costly, especially to students who don't have much time, but colleges can avoid these costs with nontraditional procedures. Middle-class policymakers (and researchers) need to understand this lesson. Choices may also change if students consider a broader range of job rewards. Instead of reifying college procedures, credentials, and outcomes as if they were single entities, research can identify alternatives and study how different kinds of students fare with them.

For public policy, these issues have implications for addressing labor market needs. Even in the weak economy of the last few years, some industries report shortages for filling mid-skilled jobs (Deloitte and Manufacturing Institute 2011; ManpowerGroup 2011; Haymes 2012). Mid-skill jobs in technician and health occupations continue to have strong demand (Holzer, Lane, Rosenblum, and Andersson

2011, p. 149), and technician occupations grew even during the 2007–2009 recession (Acemoglu and Autor 2010). Many youth don't have such skills, which they could get from community college sub-BA programs. Unfortunately, most students don't know about these credentials or their payoffs, and they often don't consider them an option. Alternative procedures that make these credentials and their nonpecuniary payoffs more salient may improve students' incentives and outcomes, and indeed may broaden their conceptions of success. Like the mythical Lake Wobegon, "where all the children are above average," perhaps most community college students can get above-average jobs—at least on some dimension.

Research must be more focused to examine procedures, but this approach may be more useful in understanding reforms. At a time when community colleges are trying new reforms, research can go beyond average outcomes to look at variation and underlying processes. Such research can use economic conceptions to clarify how reforms operate and how reforms can be better designed to conform to economic principles.

Over the past 40 years, college aspirations and enrollment have dramatically increased, and social science research may have contributed to encouraging new kinds of students to attend college (including low-achieving students). However, researchers have not considered whether these students would have greater success with different college procedures, different interim credentials, different sequences, and different job goals. Researchers can make an important contribution by testing these preconceptions and helping policymakers see beyond their BA blinders.

■ *The authors thank Kennan Cepa, Greg Duncan, David Figlio, Norton Grubb, Jon Guryan, Diane Schanzenbach, and the editors of this journal for helpful suggestions. The authors also appreciate the support of the Spencer Foundation, the Gates Foundation, the Institute for Social Policy Research at Northwestern University, the School of Public Health, State University of New York, the American Educational Research Association, and the National Institute for Child Health and Human Development grant R24-HD041041 to the Maryland Population Research Center.*

References

Acemoglu, Daron, and David H. Autor. 2010. "Skills, Tasks and Technologies: Implications for Employment and Earnings." Chap. 12 in *Handbook of Labor Economics*, Volume 4B, edited by Orley Ashenfelter and David Card. Amsterdam: Elsevier.

Adelman, Clifford. 2004. *Principal Indicators of Student Academic Histories in Postsecondary Education, 1972–2000.* Washington DC: US Department of Education, Institute of Education Sciences.

Bailey, Thomas, Dong Wook Jeong, Sung-Woo Cho. 2010. "Referral, Enrollment, and Completion in Developmental Education Sequences in

Community Colleges." *Economics of Education Review* 29(2): 255–70.

Baum, Sandy, Jennifer Ma, and Kathleen Payea. 2010. *Education Pays 2010: The Benefits of Higher Education for Individuals and Society.* Princeton, NJ: College Board.

Bound, John, Brad Hershbein, and Bridget Terry Long. 2009. "Playing the Admissions Game: Student Reactions to Increasing College Competition." *Journal of Economic Perspectives* 23(4): 119–46.

Carnevale, Anthony P., Stephen J. Rose, and Ban Cheah. 2011. "The College Payoff: Education, Occupations, Lifetime Earnings." Georgetown University Center on Education and the Workforce. http://www9.georgetown.edu/grad/gppi/hpi/cew/pdfs/collegepayoff-complete.pdf.

Carnevale, Anthony P., Stephen J. Rose, Andrew. R. Hanson. 2012. *Certificates: Gateway to Gainful Employment and College Degrees.* Georgetown University Center on Education and the Workforce. http://www9.georgetown.edu/grad/gppi/hpi/cew/pdfs/Certificates.FullReport.061812.pdf.

Complete College America. 2011. *Time is the Enemy.* http://www.completecollege.org/docs/Time_Is_the_Enemy.pdf.

Couch, Christina. 2012. "The Million-Dollar Edge of a College Degree." Bankrate.com, August 31. http://www.bankrate.com/finance/jobs-careers/million-dollar-edge-college-degree.aspx.

Deloitte, and The Manufacturing Institute. 2011. "Boiling Point? The Skills Gap in US Manufacturing." October, 17. http://www.deloitte.com/us/mfgskillsgap.

DeLuca, Stefanie, Anna Rhodes, and Robert Bozick. 2012. "Mind the Gap (Year): College Delay, Time Use and Postsecondary Pathways." Unpublished working Paper, Johns Hopkins University, Department of Sociology.

Deming, David J., Claudia Goldin, and Lawrence F. Katz. 2012. "The For-Profit Postsecondary School Sector: Nimble Critters or Agile Predators?" *Journal of Economic Perspectives* 26(1): 139–64.

Doeringer, Peter, and Michael J. Piore. 1971. *Internal Labor Markets and Manpower Analysis.* Lexington, MA: Lexington Books.

Educational Longitudinal Study of 2002. N.d. National Center for Education Statistics. US Department of Education. http://nces.ed.gov/surveys/els2002/.

Gill, Andrew M., and Duane E. Leigh. 2004. *Evaluating Academic Programs in California's Community Colleges.* San Francisco: Public Policy Institute of California.

Grandner, Michael A., Lauren Hale, Melisa Moore, and Nirav P. Patel. 2010 "Mortality Associated with Short Sleep Duration: The Evidence, the Possible Mechanisms, and the Future." *Sleep Medicine Reviews* 14(3): 191–203.

Grubb, W. Norton. 1996. *Working in the Middle: Strengthening Education and Training for the Mid-Skilled Labor Force.* San Francisco: Jossey-Bass.

Grubb, W. Norton. 2002. "Learning and Earning in the Middle, Part I: National Studies of Pre-Baccalaureate Education." *Economics of Education Review* 21(4): 299–321.

Haymes, Lindsay. 2012. "System Overhall: City Colleges Get a Revision." *Chicago Policy Review,* May 9. http://chicagopolicyreview.org/2012/05/09/system-overhaul-city-colleges-get-a-revision/.

Holzer, Harry J., Julia I. Lane, David B. Rosenblum, and Frederik Andersson. 2011. "Where Are All the Good Jobs Going? What National and Local Job Quality and Dynamics Mean for U.S. Workers." New York: Russell Sage Foundation.

Horn, Laura. 1999. *Stopouts or Stayouts? Undergraduates Who Leave College in Their First Year.* Project Officer: Dennis Carroll. National Center for Education Statistics, Statistical Analysis Report. U.S. Department of Education, NCES 1999-087.

Ingersol, Richard. 2004. "The Status of Teaching as a Profession." In *Schools and Society,* edited by Jeanne Ballantine and Joan Spade, 102–16. Belmont, CA: Wadsworth.

Jacobson, Louis, and Christine Mokher. 2009. *Pathways to Boosting the Earnings of Low-Income Students by Increasing Their Educational Attainment.* Prepared for the Bill & Melinda Gates Foundation by Hudson Institute and CNA. Washington, DC: Hudson Institute Center for Employment Policy. http://www.hudson.org/files/publications/Pathways%20to%20Boosting.pdf.

Jones, Stan. 2011. "Freedom to Fail? The Board's Role in Reducing College Dropout Rates" *Trusteeship,* January/February, 2–5.

Kain, John F. 1968. "Housing Segregation, Negro Employment, and Metropolitan Decentralization." *Quarterly Journal of Economics* 82(2): 175–97.

Lederman, Doug. 2008. "College Isn't Worth a Million Dollars." *Inside Higher Ed,* April 7. http://www.insidehighered.com/news/2008/04/07/miller.

Long, Bridget Terry, and Erin K. Riley. 2007. "Sending Signals to Students: The Role of Early Placement Testing in Improving Academic Preparation." In *Minding the Gap: Why Integrating High School with College Makes Sense and How to Do It,* edited by Nancy Hoffman, Joel Vargas, Andrea Venezia, and Marc Miller, 105–112. Harvard Education Press.

ManpowerGroup. 2012. *2012 Talent Shortage Survey.* Milwaukee. http://www.manpowergroup.us/campaigns/talent-shortage-2012/.

Mortimer, Jeylan T., Melanie J. Zimmer-Gembeck, Mikki Holmes, and Michael J. Shanahan. 2002. "The Process of Occupational Decision-Making." *Journal of Vocational Behavior* 61(3): 439–65.

Murnane, Richard J., and Frank Levy. 1996. *Teaching the New Basic Skills: Principles for Educating Children to Thrive in a Changing Economy.* New York: The Free Press.

Murray, Charles. 2008. *Real Education: Four Simple Truths for Bringing America's Schools Back to Reality.* New York City: Crown.

National Center for Education Statistics (NCES). 2012. *Digest of Educational Statistics.* Washington, DC: National Center for Educational Statistics.

Persell, Caroline H., and Peter W. Cookson. 1985. "Chartering and Bartering: Elite Education and Social Reproduction." *Social Problems* 33(2): 114–29.

Presser, Harriet B. 2005. *Working In a 24/7 Economy: Challenges for American Families.* Russell Sage Foundation.

Redline, Julie E., and James E. Rosenbaum. 2010. "School Job Placement: Can it Avoid Reproducing Social Inequalities?" *Teachers College Record* 112(3): 843–75.

Rosenbaum, James, and Kennan Cepa. 2013. "Community College Students' Plans and Disappointments." Draft paper, Institute for Policy Research, Northwestern University.

Rosenbaum, James, Kennan Cepa, and Janet Rosenbaum. 2013. "Beyond the One-Size-Fits-all College Degree." *Contexts* 12(1): 48–52.

Rosenbaum, James E., Regina Deil-Amen, and Ann E. Person. 2006. *After Admission: From College Access to College Success.* New York: Russell Sage Foundation Press.

Rosenbaum, James E., and Kelly Iwanaga Becker. 2011. "The Early College Challenge: Navigating Disadvantaged Students' Transition to College." *American Educator* 35(3): 14–20.

Rosenbaum, James E., Janet Rosenbaum, Jennifer Stephan, Amy E. Foran, Pam Schuetz. Forthcoming. "Beyond BA Blinders: Cultural Impediments to College Success. My 4-Year Degree was the Longest 8 Years of My Life." In *Bringing Culture Back In: Rethinking the African-American Youth Crisis,* edited by Orlando Patterson. Cambridge, MA: Harvard University Press.

Rosenbaum, Janet. 2012 "Degrees of Health Disparities: Health Status Disparities between Young Adults with High School Diplomas, Sub-baccalaureate Degrees, and Baccalaureate Degrees." *Health Services and Outcomes Research Methodology* 12(2–3): 156–68.

Rosenbaum, Janet. 2013. "Money Isn't Everything: Do Sub-BA Credentials Lead to Non-Monetary Job Rewards?" Draft paper. State University of New York, Brooklyn, NY.

Schneider, Barbara, and David Stevenson. 1999. *The Ambitious Generation: Motivated but Directionless.* Yale University Press.

Schuetz, Pam, James Rosenbaum, Amy Foran, and Kennan Cepa. 2012. "Degree Ladders." Unpublished working paper, Northwestern University.

Schwartz, Barry. 2004. *The Paradox of Choice.* New York: Harper.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics* 87(3): 355–74.

Stephan, Jennifer. 2010. "Is an Associate's Degree a Dead End?" Unpublished analysis, National Education Longitudinal Survey, Northwestern University, Institute for Policy Research.

Stephan, Jennifer L., James E. Rosenbaum, and Ann E. Person. 2009. "Stratification in College Entry and Completion." *Social Science Research* 38(3): 572–93.

Stigler, George J. 1961. "The Economics of Information." *Journal of Political Economy* 69(3): 213–25.

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* Yale University Press.

Tierney, William G. 2012. "The Conundrum of Profit-Making Institutions in Higher Education." In *Preparing Today's Students for Tomorrow's Jobs in Metropolitan America,* edited by Laura Perna, 149–76. University of Pennsylvania Press.

Zemsky, Robert. 1994. "What Employers Want: Employer Perspectives on Youth, the Youth Labor Market, and Prospects for a National System of Youth Apprenticeships." National Center on the Educational Quality of the Workforce, University of Pennsylvania, Philadelphia.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 173–192

Economics versus Politics: Pitfalls of Policy Advice

Daron Acemoglu and James A. Robinson

The fundamental approach to policy prescription in economics derives from the recognition that the presence of market failures—like externalities, public goods, monopoly, and imperfect competition—creates room for well-designed public interventions to improve social welfare. This tradition, already clear in Pigou (1912), was elaborated by Samuelson (1947), and still provides the basis of most policy advice provided by economists. For example, the first development economists in the 1950s used market-failure–inspired ideas as the intellectual basis for the need for government intervention to promote development in poor countries (Killick 1978). Though belief in the ability of the government or the effectiveness of aid has waxed and waned, current approaches to development problems have much in common with this early tradition, even if they have become more sophisticated—in recognizing second-best issues, for instance, by incorporating informational frictions explicitly in policy design (for example, Townsend 2011); in highlighting the specificity of the appropriate policy depending on context (for example, Rodrik 2007); and in emphasizing the role of rigorous empirical methods in determining which sorts of interventions can be effective (for example, Banerjee and Duflo 2011). But in all of these approaches, politics is largely absent from the scene.

This neglect of politics is often justified—implicitly or explicitly—in one of three ways. The first is to maintain that politicians are basically interested, or induced to be interested, in promoting social welfare, for example, because socially

■ *Daron Acemoglu is the Elizabeth and James Killian Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. James A. Robinson is David Florence Professor of Government, Harvard University, Cambridge, Massachusetts. Their email addresses are daron@mit.edu and jrobinson@gov.harvard.edu.*

http://dx.doi.org/10.1257/jep.27.2.173. doi=10.1257/jep.27.2.173

efficient policy is what helps politicians to stay in power or get re-elected, as in models like Whitman (1989, 1995) and Mulligan and Tsui (2006, 2008).

The second is to view politics as a random factor, just creating potentially severe but unsystematic grit on the wheels of economic policymaking (for example, Sachs, 2005, or as in the Banerjee, 2012, argument that the Liberian dictator Samuel Doe's economic policies were disastrous because he did not understand "what was involved in being president").

The third justification recognizes that political economy matters, but maintains that "good economics is good politics," meaning that good economic policies necessarily relax political constraints (for examples, Boycko, Shleifer, and Vishny 1995; Banerjee and Duflo 2011, in particular, p. 261; Sachs et al. 2004). The implication is the same as the first two views: one could unwaveringly support good economic policies, assured that they will not only solve market failures but also unleash beneficial political forces—whatever those may be.

In this essay, we argue not only that economic advice will ignore politics at its peril but also that there are systematic forces that sometimes turn good economics into bad politics, with the latter unfortunately often trumping the economic good. Of course, we are not claiming that economic advice should shy away from identifying market failures and creative solutions to them, nor are we suggesting a blanket bias away from good economic policy. Rather, our argument is that economic analysis needs to identify, theoretically and empirically, conditions under which politics and economics run into conflict, and then evaluate policy proposals taking into account this conflict and the potential backlashes it creates.

Our basic argument is straightforward: the extant political equilibrium may not be independent of the market failure; indeed it may critically rest upon it. Faced with a trade union exercising monopoly power and raising the wages of its members, many economists would advocate removing or limiting the union's ability to exercise this monopoly power, and this is certainly the right policy in some circumstances. But unions do not just influence the way the labor market functions; they also have important implications for the political system. Historically, unions have played a key role in the creation of democracy in many parts of the world, particularly in western Europe; they have founded, funded, and supported political parties, such as the Labour Party in Britain or the Social Democratic parties of Scandinavia, which have had large effects on public policy and on the extent of taxation and income redistribution, often balancing the political power of established business interests and political elites. Because the higher wages that unions generate for their members are one of the main reasons why people join unions, reducing their market power is likely to foster de-unionization. But this may, by further strengthening groups and interests that were already dominant in society, also change the political equilibrium in a direction involving greater efficiency losses. This case illustrates a more general conclusion, which is the heart of our argument: even when it is possible, removing a market failure need not improve the allocation of resources because of its effect on future political equilibria. To understand whether it is likely to do so, one must look at the political consequences of a policy—it is not sufficient to just focus on the economic costs and benefits.

To develop this argument more fully, we offer a simple theoretical framework clarifying the links between economic policy and the political equilibrium. We emphasize why, in the presence of political economy considerations, economic cost–benefit analysis is not sufficient, and also how, in contrast to standard second-best reasoning, our argument provides some pointers for what types of market failures, if removed, are most likely to have deleterious impacts on the political equilibrium. We highlight economic policies that strengthen the already dominant groups in society—conversely, weakening their political counterweights—as those that need to be studied more holistically, combining politics with economics, to avoid major unintended political consequences.

We then discuss three broad mechanisms generating circumstances under which good economic policy may make bad politics. First, economic rents in the present can affect political equilibria; policies that seek to address market failures can reduce the economic rents for certain groups and thus may have unintended political consequences, particularly when the rents that are destroyed are those of groups that are already weak, further tilting the balance of power in society. Second, even in the absence of changing rents, the distribution of income can affect the political equilibrium, which implies that the distributional effects of the policies that enhance economic efficiency cannot be ignored for an additional, political reason. Once again, policies that lead to a further increase in inequality would be the ones most likely to have counterproductive political implications. Third, political incentive compatibility constraints, which determine the interests a politician has to satisfy to remain in power, may be violated as a result of removing market failures, creating a political backlash. In each case, we provide a few examples to illustrate the mechanisms in action.

At this point, our mechanisms are mainly illustrative. Our purpose is to show that the issues highlighted by our framework are present in a number of important historical and current episodes, and that there are some important commonalities consistent with a basic political economy approach—in particular, linking the counterproductive political implications to economic policies that improve the standing of already dominant groups and interests in society. A more systematic empirical and theoretical analysis of these issues is necessary to uncover the major regularities and lessons, to enrich our views of how economics and politics interact, and to delineate the circumstances, if any, where economists can go on abstracting from politics.

A Theoretical Framework

To assist in clarifying these ideas and to organize the discussion of mechanisms in the next section, consider a two-period model. Suppose an economic policy has to be chosen in both periods and there are no economic linkages between these two periods. In addition suppose that in the first period politicians have some freedom of choice over policy—in some sense, there is a "window of policy opportunity" so that policy is not completely determined by vested interests or some political calculations. This policy choice might also be influenced by advice from economists, for example, aimed at correcting a market failure. In the second period, policies will be determined in a political equilibrium.

Let us first focus on the world of economics without politics, with no political (or economic) linkage between the two periods. In such a world, the first-period policy choice can be made without any concern for the political equilibrium in the second period.[1] However, the reality is that policy choices in the first period often strengthen some groups and weaken others, and thus will likely affect the political equilibrium in the second period. In turn, the political equilibrium will determine the choices made in the second period. Therefore, the objective of the welfare maximizing policymaker, and the advice given by economists, should not just be to solve market failures today, but should take into account the later political ramifications of this first period choice.[2]

The argument so far is similar to a political version of the famous second-best caveat to economic policy analysis (Lancaster and Lipsey 1956). But often, more can be said. Much political economy analysis highlights the role of the balance of political power in society, emphasizing in particular that (1) economic and political power are linked; and (2) the political dominance of a narrow interest group or segment of society will have deleterious effects (for example, Acemoglu and Robinson 2012). In this light, policies that economically strengthen already dominant groups, or that weaken those groups that are acting as a counterbalance to the dominant groups, are especially likely to tilt the balance of political power further and have unintended, counterproductive implications.[3] In addition, economic reforms that leave the fundamental political and institutional sources of

[1] Mathematically, in the world of economics without politics, policies in the two periods, x_1 and x_2, are chosen independently to maximize welfare, $\sum_{t=1}^{2} W_t(x_t)$ (where discounting is suppressed without any loss of generality). Here W_t captures social welfare in period t. In this case, the social welfare maximizing policy/advice in the first period would be x_1^{SW} such that $W_1'(x_1^{SW}) = 0$.

[2] Mathematically, we can think of second-period policy being determined as $x_2 = \xi(p_2)$, where p_2 is an index of the distribution of political power in the second period. This distribution of political power is itself determined in part by today's policies, which can be summarized by a function π, so that $p_2 = \pi(x_1)$. In contrast to the situation in footnote 1, social welfare maximization in this world, where economic policies and politics in the future are endogenous, will require (assuming differentiability):

$$W_1'(x_1) \;+\; W_2'(\xi(\pi(x_1)))\,\frac{d\xi(\pi(x_1))}{dp_2}\,\frac{d\pi(x_1)}{dx_1} \;=\; 0.$$

Therefore, unless $d\xi/dp_2 = 0$ (so that future policies are independent of future politics) or $d\pi/dx_1 = 0$ (so that future politics is independent of today's policies), the second term in this equation will be nonzero, implying that the objective of the welfare-maximizing policymaker, and the advice given by economists, should not just be to solve market failures today but should factor in politics.

[3] Following up on footnote 2, one first needs to order policies, for example, such that higher x favors the already politically powerful groups. With this ordering, denote the status quo policies which will apply without any intervention by x_1^0 and x_2^0. Suppose that $x_2^0 > x_2^{SW}$, so that the status quo in the future is already biased in favor of the politically powerful, and that p_2 increases (shifts in favor of the dominant groups) when x_1 increases. Then any policy reform that involves $x_1 > x_1^0$ (so that it favors the politically powerful relative to the status quo today) will tend to increase p_2 and shift the political equilibrium in the second period further to the benefit of the politically powerful. This tends to lead to yet higher values of x_2 (thus increasing the gap between actual and socially optimal policies in the second period). Our framework suggests that the political consequences of these types of policies should be carefully studied.

inefficiencies unchanged and instead deal with some of their symptoms in a superficial way also risk a political backlash by violating "political incentive compatibility constraints"— effectively destroying existing political equilibria or coalitions. We show later how this has been an endemic problem with policy reform in Africa, where rather than being targeted at the fundamental political economy problems that create poor policy, reforms often focus on an outcome of these problems, such poor monetary or fiscal policy.

Of course, the devil is in the details. How might current economic policy choices affect future political equilibria? How do political equilibria affect the level of welfare that will be achieved in the future? Clearly, these effects may differ across settings, like democracies versus nondemocracies, but we will argue that in many instances they seem to be present and of first-order importance.

The Organizational Importance of Economic Rents

Economic rents create incentives to organize—in particular, to extract and/or take advantage of those rents or to protect them. The existence of organizations has potentially powerful political consequences. Thus, eradicating market failures and removing the resulting rents will often change investments in organizations by certain individuals and groups, and via this channel influence the political equilibrium. This intuition suggests that economic policy making should take into account—or at least study—the impact of policy on the political organization of various groups.

Rents, Unionization, and Democracy

In most situations, unions clearly create economic distortions by pushing the wages of their members up relative to nonunionized employees. Unions may also create other distortions, like discouraging employers from adopting certain technologies and efficiency-enhancing practices. As a result, reducing the power of unions to push up wages is often mainstream economic advice. The counterarguments rooted in economic theory typically refer either to the role of unions in securing a more equal distribution of income, especially by improving the pay of lower-wage workers, or to arguments that firms have some monopsony power in setting wages and unions can counterbalance that power.

In the context of our framework, the key point is that any policy choice that reduces the ability of unions to push for high wages— even if it does not directly involve making it harder to organize unions—will indirectly reduce union activity. After all, many workers may no longer find joining unions worthwhile when the premium they receive is limited. In the context of our framework, today's policies affect tomorrow's organizational investments and thus the distribution of political power—in this case, the power of unions. Moreover, in many settings, despite the power of unions in the status quo, the balance of power is already tilted in favor of large employers so that weakening unions might create a more tilted balance of political power in society, with the potential dynamic costs that this will engender.

This outcome results because, as we have already noted, unions do not just fight for higher wages or attempt to influence the internal organization of firms; they have also been very active politically in ways that seem likely to affect the political equilibrium. One of the most important consequences of the political power of unions is the role they have played in creating and supporting democratic institutions around the world, particularly starting from a situation in which political power was very unequally distributed in a nondemocratic context. A recent literature on the factors influencing the creation of democracy has moved away from earlier work, such as that of Moore (1966), which emphasized the role of the middle class or the bourgeoisie, and has instead pointed out that it is often the working classes or poor segments of society that have played a defining role in the emergence and flourishing of democracy (Rueschemeyer, Stephens, and Stephens 1992; Collier and Mahoney 1997; Acemoglu and Robinson 2000, 2006). This literature argues that the extent to which the working class is organized or able to engage in collective action is critical for its ability to push for institutional change. Since unions are in the business of organizing working people, it makes sense that the presence of unions should facilitate collective action that pushes for regime change. A great deal of case study and econometric research supports this emphasis on social conflict (for example, Aidt and Jensen 2012).

Some examples where unions have played a pivotal role in democratization range from the "first wave of democratization" in Europe prior to World War I (Eley 2002) through the battle of Solidarity against the communist regime in Poland, to the fight against the apartheid regime in South Africa by the Congress of South African Trade Unions (COSATU). One of the clearest recent cases is the formation of the Workers Party (PT) in Brazil in 1979. This party emerged in the context of a strike at the Scânia truck factory in São Bernardo. The leader of the São Bernardo metalworkers was a 33 year-old activist called Luis Inácio Lula da Silva, also known as Lula, who helped to organize what was the first in a series of strikes which swept across Brazil, challenging the military dictatorship. On the face of it, these strikes were about wages and working conditions, but as Lula later recalled: "I think we can't separate economic and political factors . . . The . . . struggle was over wages, but in struggling for wages, the working class won a political victory" (quoted in Keck 1992, p. 65). Formed the year after the strike, the PT was in the vanguard of the successful movement to force the military from power in Brazil.

In summary, policies reducing the effectiveness of unions in negotiating over wages and working conditions for their members will reduce their political power. Though Lula and COSATU managed to organize in hostile political environments, the evidence suggests that unionization rates are sensitive to government policies that facilitate the creation of monopoly power and rents (Rothstein 1992, Western 1999, Schmitt and Mitukiewicz 2012). If the political power of unions is important in supporting a range of other economic and political outcomes, then correcting the labor market failures associated with union-induced high wages may backfire.

This perspective can also be applied to US experience with unions. The share of US workers belonging to a union peaked back in the early 1950s. There was an element of policy choice here: after encouraging the growth of unions with the passage of the National Labor Relations Act (the Wagner Act) in 1935 and by various actions during World War II, the bargaining environment became less favorable for unions with the passage of the Labor Management Relations Act of 1947 (the Taft–Hartley Act). Moreover, starting in the 1970s, policies that encouraged free trade increased the level of competition in the US economy, undercutting the ability of a number of private sector unions to raise wages. Of further significance was the anti-union stance of the Reagan administration (Farber and Western 2002). The decline in union membership may have had various political economy consequences, for example, as an important contributing factor to the rise in income inequality (Western and Rosenfeld 2011). More speculatively, it may have also contributed to the explosion in compensation of chief executive officers (DiNardo, Hallock, and Pischke 1997, 2000) and to the rapid deregulation of the financial sector.

Consequences of the Organization of Resource Wealth

A popular argument claims that the specific form of a country's natural resources have a first-order impact on economics and politics. Botswana's deep-mined diamonds, according to this argument, have different consequences for political stability than Sierra Leone's alluvial diamonds (Ross 2006). However, the consequences of minerals can depend not so much on their intrinsic characteristics or on whether their extraction is organized efficiently from an economic point of view, but on the political consequences of how their exploitation is structured. We illustrate this with the comparison of the exploitation of the alluvial gold deposits of Australia and the diamond deposits of Sierra Leone. In both cases, for basic economic reasons, free entry into mining was inefficient. Natural resource extraction provides a clear case of the "congestion" problem: the more others extract, the less will be left for each participant. One way to address this market failure is to assign exclusive property rights to the natural resources to a large producer, who would plan long-term and eliminate the dissipation of rents that is likely to arise through excessive entry. Another way to organize mining is to allow large numbers of individuals or small firms to search for the resources, possibly subject to registration or other fees. The Australian case suggests that when a large number of independent, small-scale miners are doing the extraction and realize their rents may be dissipated, this contributes to their political organization as a group, ultimately creating a more balanced political landscape and contributing to the development of democratic politics. On the other side, Sierra Leone's experience showcases how the prevalence of large and very profitable mining interests, as well as the political infighting to control and benefit from these mining interests, often has negative and nondemocratic implications for the distribution of political power.

In Australia, gold was discovered in New South Wales and then in the newly-formed state of Victoria in 1851. The immediate reaction of Australian political elites was to try to ban gold mining in the fear that the labor force on farms and ranches would vanish. If gold mining was to be allowed, it would be only after proper surveys had been made and the land leased out to large enterprises (Hirst 2008, p. 375). But the gold was on Crown land, and as such outside the direct control of the elites in the Legislative Assembly of New South Wales, who could neither ban mining nor allocate mining rights in large lots. The concerns of the British colonial state about the growing power of local Australian economic and political elites led to the decision to allow anyone to take out a mining license as long as they paid 30 shillings per month. Though this fee was high, it did not stop a massive gold rush. Soon 50 percent of the men in Victoria were working in the gold fields. Of Melbourne's 40 police constables, 38 resigned to go and dig gold. Ships were unable to sail from Melbourne's harbor because their crews deserted (Blainey 2006, p. 40).

As Australia's mining camps spread, resentment grew about the license, which had to be paid whether or not the miner found gold, and further restrictions were placed on the size of claims miners could stake. Punishments were also increased for those found without a license. The miners began to organize to protect their interests and increase their rents by reducing mining licenses. In 1854, they founded the Ballarat Reform League in the town of that name in the middle of the goldfields. In November 1854, the diggers in Ballarat delivered a set of "Resolutions" to the governor, which were heavily inspired by the agenda of the Chartists, a movement of working-class people seeking greater political participation in Britain. (Indeed, the secretary of the Ballarat Reform League, John Humffray, had been a Chartist in Wales before emigrating to Australia.) The demands included manhood suffrage, no property requirements to become a member of the Victorian Legislative Council, and payment of the members of that council (Hirst 2002, p. 48). They also demanded an end to licenses and the disbanding of the commissioners who collected the license fees on the goldfields. A group of miners, diggers as they were known, led by Peter Lalor, decided to refuse to pay for their licenses, took up arms, and built a stockade at Eureka. On December 3, 1854, armed police stormed the stockade, and 30 diggers and five policemen died.

In the outrage that followed, a Royal Commission recommended reform of the license system, also fixing a fee of one pound for a yearly license. Under the 1853 Victoria Constitution, someone in possession of an annual mining license was deemed to have sufficient wealth to be eligible to vote. Thus, at one fell swoop, any digger willing to pay one pound—and many were—was enfranchised. At the same time, to placate the diggers further, the Legislative Council was expanded to allow for representation from the goldfields. In 1855, Humffray and Lalor were elected to the Victoria Legislative Council along with six other diggers. In March 1856, the Legislative Council introduced the world's first effective secret ballot, henceforth known as the Australian Ballot. All eight diggers voted in favor, and the measure passed by 33 votes to 25 (Hirst 2006).

The contrast between the Australian experience, where the organization of gold deposits created a large pro-democratic force, and that in Sierra Leone is striking. Prospecting for diamonds began in Sierra Leone in the 1920s, with the first discoveries being made in 1931. Small-scale mining started in 1933 in the east of the country, and in 1935 the colonial government gave the Sierra Leone Selection Trust (SLST) practically exclusive prospecting and mining rights for the entire country. To protect these rights from illegal mining, SLST had its own security force. In his study, van der Lann (1965, p. 79) poses the question: "What is better for the

Sierra Leone economy: to have the diamond deposits slowly and steadily exploited by a mining company, or rapidly worked by diggers?" He argues that the SLST was (economically) better both "because the recovery rate of the diggers . . . falls far short of the rate of close to 100% achieved by SLST" (p. 80), and also because the SLST monopoly generated more revenues for the government.

Yet the really important feature of the organization of the diamond mining in Sierra Leone was not the economic costs and benefits of the SLST, but its political consequences. While the SLST struggled to control illegal mining by diggers, the organization of the mining did not create the type of democratic impulse as it had in Australia. One consequence of this was that the independence movement of the 1950s was spearheaded by paramount chiefs and other elites favored by British colonialism. In 1952, when these elite Sierra Leoneans began to control the Legislative Council, they chose not to open up diamond mining to Sierra Leoneans, but rather to extract greater taxes from the SLST. In exchange, they helped to enforce the monopoly rights by aggressively punishing illegal mining. The Minister of Mines in charge of this was the future kleptocratic prime minister and president of Sierra Leone, Siaka Stevens. In 1956, the number of illegal diggers had become so large—possibly 75,000 (van der Laan 1965, p. 65)—that security forces were overwhelmed. SLST's monopoly was now restricted to two areas, but these were still the prime deposits of Kono and Tongo Fields. Elsewhere, mining licenses were issued but not to "strangers"—meaning anyone who was not an indigenous resident of the chieftaincy where the mining was to take place.

Naturally, as we discuss in Acemoglu and Robinson (2012), there were other historical and institutional factors stacking the cards against the development of the type of inclusive economic and political institutions and policies that would have stimulated economic growth in Sierra Leone. But the arrangements that had been made for accessing natural resources were a key contributor to the fact that during the critical period of the founding of the first political parties in Sierra Leone, they were formed by elites, particularly by the Paramount Chiefs and those connected to them, without the input of the broad mass of Sierra Leoneans (Cartwright 1970, provides an overview). The scene was set for the creation of one-party and authoritarian rule after independence in 1961.

Political Consequences of Inequality

Removing a market failure will also generally alter the distribution of income in society. For example, when unions are less able to exercise monopoly power, not only will their organization dwindle, but (at least in the absence of robust competition) profits will rise. Income will typically be redistributed from workers to the managers and owners of firms. However, this shift will also influence the political equilibrium.

An example of how an altered distribution of income can have a first-order impact on future politics is provided by the effect of Atlantic trade opportunities on the English political system in the seventeenth century. Because Atlantic trading activities were not the monopoly of the Crown in England at this time, this trade was dominated by independent merchants, adventurers, and privateers. Profits from Atlantic trade enriched many of these men, who opposed the Stuart monarchs' absolutism and sought to limit the Crown's prerogatives. As they became richer, they also became more powerful and bolder, and were even able to field armies to defeat the monarchy in the English Civil War of the 1640s and then during the Glorious Revolution of 1688. The papers Acemoglu, Johnson, and Robinson (2005) and Jha (2010) provide historical and empirical evidence linking the rise of inclusive institutions in England and the Dutch Republic to the rise of merchants and industrialists benefiting from Atlantic trade. Tellingly, this trajectory is very different than the one observed in Portugal or Spain, where Atlantic trading activities were monopolized and allocated by the Crown. In these countries, the riches of trade flowed into the coffers of the already dominant monarchs, strengthening the monarchy, weakening the parliaments of these nations, and contributing to the tilted balance of political power, which persisted and underpinned the lack of economic and political development in these parts of Western Europe.

Money and Politics in the United States

The experience of financial deregulation over the past 30 years in the United States, as analyzed by Johnson and Kwak (2010), provides an illustration of how economic policy designed with a disregard for political implications can be injurious to social welfare. The system of financial and banking regulation that emerged from the Great Depression had many features that were irrational from a purely economic viewpoint. These included the prohibition of interstate banking and the separation of commercial from investment banking. Jayaratne and Strathan (1996), among others, found that the removal of some of these banking restrictions spurred rapid economic growth. Such reforms are akin to those directly addressing market failures in the sense that they were removing distortions partly introduced by previous policies. But in common with the other economic policies with potentially counterproductive political consequences, these reforms also tended to strengthen an already powerful constituency, the financial sector.

Financial deregulation started small, for example, ending fixed commissions on stock trading in 1975. Then Regulation Q, which limited interest rates on savings accounts, was abolished in 1980. As Johnson and Kwak (2010) argue, while the banking and financial services industry was not powerful enough at this time to get all the deregulation it wanted, it was strong enough to block new regulation. This was relevant because considerable financial innovation was starting to take place: as one example, Salomon Brothers originated interest rate swaps in 1981. As these new financial instruments developed, and as regulations that limited what financial services banks could perform were incrementally relaxed during this time by regulators and courts, the financial sector became bigger and more profitable. Between 1980 and 2005, financial sector profits grew 800 percent in real terms, while nonfinancial profits rose by 250 percent (Johnson and Kwak 2010, chap. 3). Between 1998 and 2007, financial sector profits were on average about 30 percent of total profits in the private sector. During this period, the financial sector expanded from 3.5 percent to almost 6 percent of GDP.

As the banks got bigger and more profitable, they also became more assertive and influential. They started to lobby more and contribute more to political campaigns. While in 1990 the financial sector donated $61 million dollars to political campaigns, by 2006 this was $260 million (the industry which was the next largest donor, health care, gave only $100 million in 2006). Of course, rising wealth and campaign contributions were not the only source of rising political power for the financial industry. There was a revolving door between Wall Street and executive appointments in Washington as well. As Johnson and Kwaak (2010) point out, there was also an intellectual revolution in academic finance involving the pricing of derivative financial instruments and a body of studies arguing for deregulation, all of which was interpreted as bolstering the financial sector's position.

So financial deregulation continued. In 1994, the Riegle-Neal Interstate Banking and Branching Efficiency Act relaxed constraints on interstate banking and led to a series of mergers which constructed large nationwide banks. JPMorgan Chase and Citicorp were formed, and the Bank of America transformed. In 1999, the Gramm–Leach–Bliley Act effectively codified the demolition of most of the barriers between commercial and investment banking, barriers that had already been falling incrementally for several decades as a result of regulatory and court decisions. But perhaps more important than these changes was the avoidance of regulations: for example, regulations that might have altered how accountants and regulators treated the collateralized debt obligations based on mortgage-backed securities and the credit default swaps sold by insurance companies like the American Insurance Group (AIG). The political power of the financial industry also accentuated the moral hazard problem in finance (that large financial institutions can take risks expecting to be bailed out by the government when things get bad). Ultimately, these regulatory changes and the regulatory void, in conjunction with the moral hazard problem, created an environment that encouraged excessive risk-taking and contributed to the 2007–2008 financial crisis.

In terms of our framework, this account illustrates how potentially efficiency-enhancing deregulation may have increased the size and political power of the financial industry, which then altered the structure of future regulations and allocations in favor of the financial industry, with potentially adverse consequences for the rest of society. Put differently, any analysis of these economic policies that focused only on their economic costs and benefits but did not take into account the political consequences of the changes they unleashed would dramatically understate the likelihood of the costs that actually occurred.

Russian Privatization

Most economists favor privatization of industry, and few argue that government ownership of industry is efficient from a cost–benefit perspective. Like deregulation, privatization is also proposed as a way of improving economic efficiency by reversing existing (government-imposed) distortions. Yet the privatization of firms in Russia during the 1990s is another example of a policy with a major effect on income distribution, creating a group of very wealthy individuals and putting in motion significant political changes—not only in terms of the direct negative consequences of the policy, but also in terms of the potential weakening of the reform process and the backlash that these policies created, paving the way for the rise of Vladimir Putin's authoritarian regime.

In summer 1991, Boris Yeltsin won the election for the newly created Russian presidency. His platform, on the basis of which he beat four Communists and a hardcore nationalist, included a radical program of market-oriented reform. To implement it, he picked Yegor Gaidar, who in turn asked Anatoly Chubais to be in charge of privatization. Of all the policies that Yeltsin wanted to implement, the privatization of the country's thousands of state-owned firms was perhaps the most critical; but he had no specific plan about how to accomplish it. Gaidar and Chubais came up with a strategy to put the main assets of the Soviet Union into private hands.

Starting in the spring of 1992, small firms like stores and restaurants began to be sold off. People could take ownership of their own apartment for free or almost for free. In late 1992, Chubais turned to the big firms. Yeltsin's team tried to get the public involved in this initial distribution of assets. Large- and medium-sized enterprises were required to sell 29 percent of their shares in voucher auctions, and in October 1992, each Russian adult was issued vouchers with a nominal value of 10,000 rubles; one's vouchers could be acquired at a local bank for a fee of just 25 rubles. By January 1993, 97 percent of Russians had claimed their vouchers. These vouchers could be sold or used to bid for the shares of specific companies when they privatized. The first voucher auctions were held in December 1992, and in total, about 14,000 enterprises held such auctions. However, most assets of these firms went to their workers and managers. The law allowed for workers and managers to buy 51 percent of the voting shares of a firm at a discount and using the firms' own funds. In effect, the majority of privatizing firms' assets were handed to insiders at huge discounts.

The most controversial stage of the privatization—and in hindsight the most clearly deleterious—was the loans-for-shares deal in 1995. State shares in twelve highly profitable enterprises concentrated in the energy sector were used as collateral for bank loans to the government. If the loans were not paid off, and the government never had any intention of paying them off, the banks would have the right to sell the shares. Between November 1996 and February 1997, sales happened for the shares of several large firms including Yukos, Sidanko, and Surgutneftegaz, and in each case, the shares were bought by the banks themselves in auctions where outside bids were ignored or disqualified. Freeland (2000) and Hoffman (2002) provide overviews of these events and a description of the resulting rise of the oligarchs. Not only did this type of privatization massively enrich and empower the oligarchs, but it also failed to create a large number of small shareholders. In 1994, workers owned 50 percent of the average Russian enterprise; by 1999, this figure had dropped to 36 percent. By 2005, 71 percent of medium and large industry and communications enterprises had a single shareholder who owned half the stock (Treisman 2011, pp. 223–24).

The driving force behind privatization was textbook economics, to move Russia from central planning and state ownership to a much more efficient market economy. This was certainly the view of many economists at the time, and the main debate was about how fast to privatize (Aghion and Blanchard 1994), not whether to maintain state ownership or not. (Arguments that privatization might create a private monopoly, with even worse economic consequences than public ownership, as suggested by Borenstein (2002) in the context of California, were not commonly raised at that time. Since then some authors, including Black, Kraakman, and Tarassova (2000), Stiglitz (2002), and Goldman (2003), have argued against privatization on purely economic grounds.) To the extent that economists worried about the political economy of the process, they did not consider that privatization might have adverse political consequences. Rather, they focused on how to structure the transition so that the political coalition in favor of privatization would stay on track (Dewatripont and Roland 1992), or on the political constraints shaping what form of privatization would be able to occur (Shleifer and Treisman 2000). In fact, a common view was that the particular details of Russia's privatization were not of first-order importance, essentially because of the "good economics is good politics" argument. Boycko, Shleifer, and Vishny (1995, pp. 10–11), for example, asserted: "[A]t least in Russia, political influence over economic life was the fundamental cause of economic inefficiency, and the principal objective of economic reform was, therefore, to depoliticize economic life . . . Privatization fosters depoliticization because it robs politicians of control over firms."

There is indeed evidence that Russia's privatization was initially good for the economy, and even the oligarchs appear to have at first invested heavily in their new firms (Treisman 2011; Åslund 2007, chap. 6). For example, Shleifer and Treisman (2004, p. 29) ask: "Have the oligarchs stripped assets from the companies they acquired in privatization, rather than investing in them? The audited financial statements of these companies suggest that their assets have grown dramatically, especially since 1998 . . . And the major oligarchs have been investing hundreds of millions of dollars annually in their companies . . ."

But our emphasis here is on the political consequences of the privatization, which turned out to be highly damaging. The privatization enriched and also temporarily politically empowered a group of unscrupulous oligarchs; in fact, so much so that inequality in Russia rose significantly following privatization (Alexeev 1999). Even more importantly, the economic and political inequality it created induced a backlash against the process of economic and political reform in Russia, ultimately re-creating authoritarianism and firmly entrenching a form of state-led crony capitalism (see Guriev and Sonin 2008, for a theoretical analysis). There are several layers to understanding how this political equilibrium evolved. First, privatization failed to create the type of broad distribution of assets which would have provided the economic underpinning for the nascent democracy and socially desirable economic policies. Second, the distribution of gains was not just narrow, it was illegitimate—because the large increase in inequality favored the politically enterprising and the connected. Third, the concentrated nature of the assets which emerged from this process and the huge rents that were up for grabs made it very easy for the KGB, re-energized under the leadership of Putin, to wrestle back control of the economy. Finally, the way in which the privatization took place may have undermined the incentives of the oligarchs to push for better institutions (Sonin 2003), and may also have fueled popular support for Putin's authoritarian political strategy.

Our bottom line on the experience of Russian privatization is that a purely economic approach to moving from collectively to privately owned assets turned out to be woefully inadequate—as was a political economy approach based on the assertion that "good economics is good politics." The evidence instead suggests that privatization, particularly its form, had a defining impact on Russian politics and contributed to the rise of an authoritarian and repressive regime ruling over a much more unequal society.

Violating Political Incentive Constraints

Politicians typically face "political incentive compatibility constraints," which determine the expected utility that a political leader in power must obtain himself or give to organized interests if he or she wishes to stay in power. Removing market failures, without recognizing and addressing the fundamental political and institutional sources of distortions, may violate these constraints. Put differently, a set of policies which may seem deeply misguided by the standards of basic textbook economics may nonetheless be serving the political economy purpose of holding together a governing coalition. By implication, removing such market failures can weaken existing coalitions or disrupt equilibria. The result may be the rise of new coalitions or new types of equilibria, which might reinstate the market failures or create new ones, because they are useful in binding together the governing coalition or creating rents for the rulers. In our study of central bank independence under weak institutions, Acemoglu, Johnson, Querubín, and Robinson (2008), we called this type of re-creation of distortions "the seesaw effect." But more ominously, the results of violating political incentive compatibility constraints might also be a period of civil unrest, with high costs of its own, or even civil war. Thus, addressing market failures in this setting without appropriate consideration of political consequences may be ineffective or even counterproductive in broad social welfare terms.

Policy Reform, Instability, and Violence

The experience of Ghana's Prime Minister in 1971, Kofi Busia, illustrates that policy advice should take into account that politicians face political constraints and that in this case as well, good economics is not necessarily good politics.

Busia had come to power in 1969 after the military junta that had ejected the increasingly autocratic government of Kwame Nkrumah in 1966 had finally given up its power. Busia immediately faced a deep economic crisis, underpinned by unsustainable expansionary economic policies, distortionary price controls implemented through marketing boards, and an overvalued exchange rate. But these policies were not adopted because Ghanaian leaders, Busia included, believed that they were good economics. Nor were they embraced as a way to develop the country. They were chosen to satisfy political constraints. The expansionary economic policies and overvalued exchange rates enabled Busia and his predecessors to transfer resources to urban groups. Price controls also had a strong political logic, first recognized by Bates (1981) in his classic book: market distortions and price controls create valuable rents which can then be allocated to generate political support. Ghanaian pricing policies squeezed agriculture, delivering cheap food to the politically more powerful urban constituencies, and generated revenues which financed government spending—and lined politicians' pockets.

Taken together, these policies did mean that balance of payments crises and foreign exchange shortages, as well as economic recession, became unavoidable. To outside institutions such as the World Bank and the International Monetary Fund (IMF), the problem and its solution were clear: distortionary policies had to be removed. Faced with economic crisis and international pressure, Busia caved in and signed an agreement with the IMF on December 27, 1971, which included a massive 44 percent devaluation of the currency. Whatever the textbook economic logic behind the reforms, the political outcome was dire. The devaluation was followed by rioting, continuous demonstrations and discontent. Two weeks after the announcement of the devaluation, Busia was overthrown in a military coup, which immediately reversed the devaluation. State controls over prices, wages, marketing boards, and exchange rates were the heart of a Ghanaian politician's patronage network, and any politician who lost the support of this network was susceptible both at the polls and against the military.

The combination of policy reform followed by violence in Ghana is not an isolated instance. As Herbst (1990) and Reno (1995, 1998) have pointed out, there is a general pattern in countries across West Africa of policy reform being followed by violence: indeed, one reason that policy reform is so seldom implemented in Africa (van de Walle 2001) is because politicians know that it is expected to lead to the breakdown of political order.

Reno's (1995, 1998) analysis of Sierra Leone is telling. After the rise to power of former Minister of Mines Siaka Stevens and his All People's Congress Party in 1968, a political compact emerged in Sierra Leone based on the creation and distribution of rents. Patrimonialism and redistribution of these rents was burnished to a fine art by Stevens, who manipulated traditional political institutions such as the chieftaincy and bought support on a massive scale using rents, patronage, and jobs. Stevens ruled until 1985 when he gave way to his hand-picked successor Joseph Momoh, who ruled the country until he was overthrown by a military coup in 1992. Barely any public goods were provided in the country in the 40 years prior to the end of the civil war and re-democratization in 2002. Roads fell to pieces and schools disintegrated. National television broadcasts stopped in 1987 when the transmitter was sold by the Minister of Information, and in 1989 a radio tower which relayed radio signals outside Freetown fell down, ending transmissions outside the capital (Reno 2003, p. 48). The Sierra Leone Produce Marketing Board had a monopsony over all export crops and paid farmers very low prices—as low as 40 percent of the world level (Davies 2007). The exchange rate was massively overvalued, creating a black market. GDP per capita fell almost monotonically from the early 1970s onwards and reached about 40 percent of the level recorded at independence by the end of the civil war in the early 2000s (Davies 2007).

But ironically, Reno (1995, 1998) argues that attempts by the international community to improve the economic policies of Sierra Leone had the unintended consequence of intensifying the existing violence and arguably even pushing the country into a bloody civil war. Sierra Leone first called in the IMF in 1979, and after that entered into a long series of negotiations. As its economy declined in the 1980s, the government's need for international resources escalated, but the problem from the point of view of President Momoh, according to Reno (1995), was that "fiscal responsibility and budget cutting in this context only hastened the urgency of finding alternative means of ensuring associates' loyalty" (p. 156). More importantly, "Momoh was losing resources to enforce political control. As revenue shortfalls and IMF austerity measures shut down parts of the state bureaucracy that had survived . . . Momoh's allies sought other means of supporting themselves as they lost access to benefits . . . The president could no longer control disobedience . . . The reform of 'bad policies' neither restored the president's political control, nor tapped 'entrepreneurial energies' which were now directed at evading the president's authority" (p. 161).

In short, well-intentioned economic policies imposed on the regime by economists trying to redress market failures and policy distortions robbed President Momoh of the instruments he had used to buy political support. As a result, he switched to a different political strategy, substituting direct force and coercion for buying people off. In January 1990, Momoh launched "Operation Clean Slate" which was in effect an attempt to use the army to take over the diamond mining areas. Without the usual instruments of patronage, such as public sector employment and contracting, Momoh turned to coercion to try to grab what rents remained in the country. The resentment this caused in the east helped to fuel a bloody ten-year civil war (Richards 1996). Though Momoh's regime was clearly extractive, kleptocratic, and repressive, the subsequent civil war was certainly not the intended objective or a desirable outcome, and Reno's analysis highlights how unintended political consequences are commonplace when reform is imposed from the outside without understanding the political equilibrium and the political incentive compatibility constraints on the ground.

Rents and the Natural State

The book by North, Wallis, and Weingast (2009) also indirectly underscores that the "good economics is good politics" dictum is fallacious by providing several counterexamples in the context of what they call the "natural state." In their conceptual framework for explaining economic development, they argue that there is a basic dichotomy between two types of social orders: on one hand, "open access" characterized by economic development, democracy, rich and vibrant civil society with lots of organizations, and widespread impersonal social relationships, including the rule of law, and secure property rights; and on the other hand, "limited access" characterized by poor economic growth, a small number of organizations and social relations along personal lines with privileges, unequal enforcement of laws and insecure property rights. All social orders, they argue, are constructed to control the threat and use of violence, but they do so in different ways with different consequences for economic incentives and development. In particular, a "natural state" is a limited access order where the key to controlling violence is the creation of rents. Echoing Bates's analysis we discussed above, they write (p. 17):

> [S]ystematic rent creation through limited access in a natural state is not simply a method of lining the pockets of the dominant coalition; it is the essential means of controlling violence. Rent-creation, limits on competition, and access to organizations are central to the nature of the state, its institutions, and the society's performance. Limiting the ability to form contractual organizations only to members of the coalition ties the interests of powerful elites directly to the survival of the coalition, thus ensuring their continued cooperation.

In the world of the natural state or limited access order, which they claim is a general model for the political economy of poor countries, good economics is almost never good politics. As North, Wallis, Webb, and Weingast (2013, p. 18) put it: "Because elites know that violence will reduce their own rents, they have incentives not to fight. Furthermore, each elite understands that other elites face similar incentives. In this way, the political system of a natural state manipulates the economic system to produce rents that then secure political order." They summarize their argument by stating (p. 7): "[T]he appropriate counterfactual from eliminating rents is not a competitive market economy . . . but a society in disorder and violence."

Concluding Remarks

There is a broad—even if not always explicitly articulated—consensus amongst economists that, if possible, public policy should always seek ways of reducing or removing market failures and policy distortions. In this essay, we have argued that this conclusion is often incorrect because it ignores politics. In fact, the extant political equilibrium may crucially depend on the presence of the market failure. Economic reforms implemented without an understanding of their political consequences, rather than promoting economic efficiency, can significantly reduce it.

Our argument is related to but different from the classical second-best caveat of Lancaster and Lipsey (1956) for two reasons. First, it is not the interaction of several market failures but the implications of current policy reforms on future political equilibria that are at the heart of our argument. Second, though much work still remains to be done in clarifying the linkages between economic policies and future political equilibria, our approach does not simply point out that any economic reform might adversely affect future political equilibria. Rather, building on basic political economy insights, it highlights that one should be particularly careful about the political impacts of economic reforms that change the distribution of income or rents in society in a direction benefiting already powerful groups. In such cases, well-intentioned economic policies might tilt the balance of political power even further in favor of dominant groups, creating significant adverse consequences for future political equilibria.

We are of course not the first ones to point out that the political economy of economic policy matters, nor that a standard cost–benefit framework for the analysis of policy is inadequate because it leaves out politics. Since the 1980s, a vibrant literature in political economy has sought to develop positive models of how policy actually gets chosen, which involves modeling politics and the decision-making process (for overviews, see Drazen 2000; Persson and Tabellini 2000; Besley 2007; Acemoglu and Robinson 2006). That being said, existing political economy analyses either do not focus on this question or else emphasize that, if politically possible, market failures should be removed. Dixit (1997) and Drazen (2002) have argued that policy (or institutional) advice must be given in a way that takes seriously the constraint that policy is chosen as part of a political equilibrium—implying that policy advice should be tempered by what is incentive-compatible for politicians. Nevertheless, to the best of our knowledge, our main argument in this paper has not been made before. Our argument is that economic policy should not just focus on removing market failures and correcting distortions but, particularly when it will affect the distribution of income and rents in society in a direction that further strengthens already dominant groups, its implications for future political equilibria should be factored in. It thus calls for a different framework, explicitly based in political economy, for the analysis of economic policy. Much of the conceptual, theoretical, and empirical foundations of such a framework remain areas for future work.

■ *We thank David Autor, Chang-Tai Hsieh, Simon Johnson, John List, Suresh Naidu, and Timothy Taylor for comments.*

References

Acemoglu, Daron, Simon Johnson, Pablo Querubín, and James A. Robinson. 2008. "When Does Policy Reform Work? The Case of Central Bank Independence." *Brookings Papers on Economic Activity*, Spring, 39(1): 351–417.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2005. "The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth." *American Economic Review* 95(3): 546–79.

Acemoglu, Daron, and James A. Robinson. 2000. "Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective. *Quarterly Journal of Economics* 115(4): 1167–99.

Acemoglu, Daron, and James A. Robinson. 2006. *Economic Origins of Dictatorship and Democracy*. Cambridge University Press.

Acemoglu, Daron, and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York, NY: Crown.

Aghion, Philippe, and Olivier J. Blanchard. 1994. "On the Speed of Transition in Central Europe." In *NBER Macroeconomics Annual*, vol. 9, edited by Stanley Fischer and Julio J. Rotemberg, 283–330. MIT Press.

Aidt, Toke S., and Peter S. Jensen. 2012. "Workers of the World Unite! Franchise Extensions and the Threat of Revolution in Europe, 1820–1938." March, 14. http://www.econ.cam.ac.uk/faculty/aidt/papers/web/workers/workers.pdf.

Alexeev, Michael. 1999. "Privatization and the Distribution of Wealth in Russia." *Economics of Transition* 7(2): 449–65.

Åslund, Anders. 2007. *How Capitalism was Built: The Transformation of Central and Eastern Europe, Russia and Central Asia.* Cambridge University Press.

Banerjee, Abhijit V. 2012. "Poor Economics—Effective Poverty Reduction Policies." Kapuscinsky Development Lecture. Available in various formats at: http://kapuscinskilectures.eu/lectures/poor-economics/.

Banerjee, Abhijit V., and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* New York, NY: Public Affairs Press.

Bates, Robert H. 1981. *Markets and States in Tropical Africa.* Berkeley, CA: University of California Press.

Besley, Timothy. 2007. *Principled Agents? The Political Economy of Good Government.* New York, NY: Oxford University Press.

Black, Bernard, Reinier Kraakman, and Anna Tarassova. 2000. *Russian Privatization and Corporate Governance: What Went Wrong?* Stanford Law Review 52(6): 1731–1808.

Blainey, Geoffrey. 2006. *A History of Victoria.* Cambridge University Press.

Borenstein, Severin. 2002. "The Trouble with Electricity Markets: Understanding California's Restructuring Disaster." *Journal of Economic Perspectives* 16(1): 191–211.

Boycko, Maxim, Andrei Shleifer, and Robert W. Vishny. 1995. *Privatizing Russia.* MIT Press.

Cartwright, John R. 1970. *Politics in Sierra Leone 1947–67.* University of Toronto Press.

Collier, Ruth Berins, and James Mahoney. 1997. "Adding Collective Actors to Collective Outcomes: Labor and Recent Democratization in South America and Southern Europe." *Comparative Politics* 29(3): 285–303.

Davies, Victor A. B. 2007. "Sierra Leone's Economic Growth Performance, 1961–2000." Chap 19 in *The Political Economy of Growth in Africa, 1960-2000*, vol. 2, edited by Benno J. Ndulu et al. Cambridge University Press.

Dewatripont, Mathias, and Gérard Roland. 1992. "Economic Reform and Dynamic Political Constraints." *Review of Economic Studies* 59(4): 703–30.

DiNardo, John, Kevin Hallock, Jörn-Steffen Pischke. 1997. "Unions and Managerial Pay." NBER Working Paper 6318.

DiNardo, John, Kevin Hallock, Jörn-Steffen Pischke. 2000. "Unions and the Labor Market for Managers." Institute for the Study of Labor (IZA) Discussion Paper 150.

Dixit, Avinash. 1997. "Economists as Advisers to Politicians and to Society." *Economics and Politics* 9(3): 225–30.

Drazen, Allan M. 2000. *Political Economy in Macroeconomics.* Princeton University Press.

Drazen, Allan M. 2002. "Conditionality and Ownership in IMF Lending: A Political Economy Approach." *IMF Staff Papers* 49(Special Issue): 36–67.

Eley, Geoff. 2002. *Forging Democracy: The History of the Left in Europe, 1850–2000.* New York, NY: Oxford University Press.

Farber, Henry S., and Bruce Western. 2002. "Ronald Reagan and the Politics of Declining Union Organization." *British Journal of Industrial Relations* 40(3): 385–402.

Freeland, Chrystia. 2000. *Sale of the Century: Russia's Wild Rise from Communism to Capitalism.* New York, NY: Crown Business.

Goldman, Marshall I. 2003. *The Privatization of Russia: Russian Reform Goes Awry.* New York, NY: Routledge.

Guriev, Sergei M., and Konstantin Sonin. 2008. "Dictators and Oligarchs: A Dynamic Theory of Contested Property Rights." *Journal of Public Economics* 93(1–2): 1–13.

Herbst, Jeffrey I. 1990. "The Structural Adjustment of Politics in Africa." *World Development* 18(7): 949–58.

Hirst, John B. 2002. *Australia's Democracy: A Short History.* Sydney, Australia: Allen and Unwin.

Hirst, John B. 2006. *Making Votes Secret: Victoria's Introduction of a New Method of Voting that has Spread around the World.* Melbourne, Australia: Victorian Electoral Commission.

Hirst, John B. 2008 *Freedom on the Fatal Shore: Australia's First Colony.* Melbourne, Australia: Black Inc.

Hoffman, David E. 2002. *The Oligarchs: Wealth and Power in the New Russia.* New York, NY: Public Affairs.

Jayaratne, Jith, and Philip E. Strathan. 1996. "The Finance–Growth Nexus: Evidence from Bank Branch Deregulation." *Quarterly Journal of Economics* 111(3): 639–70.

Jha, Suamitra. 2010. "Financial Innovation and Political Development: Evidence from Revolutionary England." Stanford University Graduate School of Business Research Paper No. 2005. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=934943.

Johnson, Simon, and James Kwak. 2010. *13 Bankers: The Wall Street Takeover and the Next Financial Meltdown.* New York, NY: Pantheon.

Keck, Margaret E. 1992. *The Workers' Party and Democratization in Brazil.* Yale University Press.

Killick, Tony. 1978. *Development Economics in Action.* London: Heinemann Educational Books.

Lancaster, Kelvin, and Richard G. Lipsey. 1956. "The General Theory of Second Best." *Review of Economic Studies* 24(1): 11–32.

Moore, Barrington, Jr. 1966. "The Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World." Boston, MA: Beacon Press.

Mulligan, Casey B., and Kevin K. Tsui. 2006. "Political Competitiveness." NBER Working Paper 12653.

Mulligan, Casey B., and Kevin K. Tsui. 2008. "Political Entry, Public Policies, and the Economy." NBER Working Paper 13830.

North, Douglass C., John Joseph Wallis, Steven B. Webb, and Barry R. Weingast. 2013. "Limited Access Orders: An Introduction to the Conceptual Framework." Chap. 1 in *In the Shadow of Violence: Politics, Economics, and the Problem of Development in Limited Access Societies,* edited by D. C. North, J. J. Wallis, S. B. Webb, and B. R. Weingast. Cambridge University Press.

North, Douglass C., John Joseph Wallis, and Barry R. Weingast. 2009. *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History.* Cambridge University Press.

Persson, Torsten, and Guido Tabellini. 2000. *Political Economics: Explaining Economic Policy.* Cambridge: MIT Press.

Pigou, Arthur C. 1912. *Wealth and Welfare.* London: Macmillan.

Reno, William. 1995. *Corruption and State Politics in Sierra Leone.* Cambridge University Press.

Reno, William. 1998. *Warlords Politics and African States.* Boulder, CO: Lynne Rienner.

Reno, William. 2003. "Political Networks in a Failing State: The Roots and Future of Violent Conflict in Sierra Leone." *Internationale Politik und Gesellschaft* no. 2, pp. 44–66.

Richards, Paul. 1996. *Fighting for the Rainforest: War, Youth and Resources in Sierra Leone.* Oxford, UK: James Currey.

Rodrik, Dani. 2007. *One Economics, Many Recipes: Globalization, Institutions, and Economic Growth.* Princeton University Press.

Ross, Michael L. 2006. "A Closer Look at Oil, Diamonds, and Civil War." *Annual Review of Political Science* vol. 9, pp. 265–300.

Rothstein, Bo. 1992. "Labor-Market Institutions and Working-Class Strength." Chap. 2 in *Structuring Politics: Historical Institutionalism in Comparative Analysis,* edited by Sven Steinmo, Kathleen Thelen, and Frank Longstreth. Cambridge University Press.

Rueschemeyer, Dietrich, Evelyn Huber Stephens, and John D. Stephens. 1992. *Capitalist Development and Democracy.* Cambridge University Press.

Sachs, Jeffrey D. 2005. *End of Poverty: Economic Possibilities for Our Time.* New York, NY: Penguin Press.

Sachs, Jeffrey D., John W. McArthur, Guido Schmidt-Traub, Margaret Kruk, Chandrika Bahadur, Michael Faye, and Gordon McCord. 2004. "Ending Africa's Poverty Trap." *Brookings Papers on Economic Activity* 35(1): 117–240.

Samuelson, Paul A. 1947. *Foundations of Economic Analysis.* Cambridge: Harvard University Press.

Schmitt, John, and Alexandra Mitukiewicz. 2012. "Politics Matter: Changes in Unionisation Rates in Rich Countries, 1960–2010." *Industrial Relations Journal* 43(3): 260–80.

Shleifer, Andrei, and Daniel Treisman. 2000. *Without a Map: Political Tactics and Economic Reform in Russia.* Cambridge: MIT Press.

Shleifer, Andrei, and Daniel Treisman. 2004. "A Normal Country." *Foreign Affairs* 83(2): 20–38.

Sonin, Konstantin. 2003. "Why the Rich May Favor Poor Protection of Property Rights." *Journal of Comparative Economics* 31(4): 715–31.

Stiglitz, Joseph E. 2002. *Globalization and Its Discontents.* New York: Norton & Co.

Townsend, Robert M. 2011. *Financial Systems in Developing Economies: Growth, Inequality and Policy Evaluation in Thailand.* Oxford University Press.

Treisman, Daniel. 2011. *The Return: Russia's Journey from Gorbachev to Medvedev.* New York, NY: The Free Press.

van de Walle, Nicolas. 2001. *African Economies and the Politics of Permanent Crisis, 1979–1999.* Cambridge University Press.

van der Laan, H. L. 1965. "The Sierra Leone Diamonds: An Economic Study Covering the Years 1952–61." Oxford University Press.

Western, Bruce. 1999. *Between Class and Market: Postwar Unionization in the Capitalist Democracies.* Princeton University Press.

Western, Bruce, and Jake Rosenfeld. 2011." Unions, Norms, and the Rise in U.S. Wage Inequality." *American Sociological Review* 76(4): 513–37.

Wittman, Donald. 1989. "Why Democracies are Efficient." *Journal of Political Economy* 97(6): 1395–1424.

Wittman, Donald. 1995. *The Myth of Democratic Failure: Why Political Institutions Are Efficient.* University of Chicago Press.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 193–218

Latin America's Social Policy Challenge: Education, Social Insurance, Redistribution

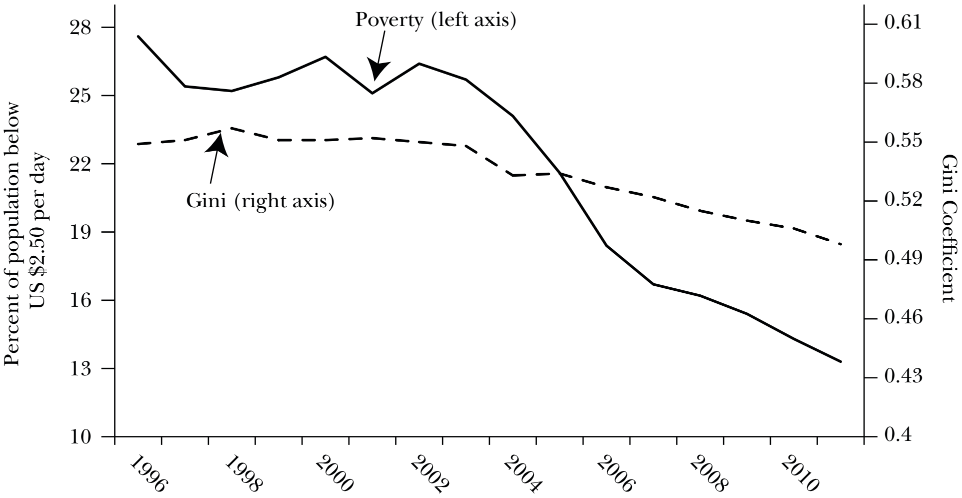
Santiago Levy and Norbert Schady

Long regarded as a region beset by macroeconomic instability, high inflation, and excessive poverty and inequality, Latin America has undergone a major transformation over the last 20 years. After the "lost decade" of the 1980s, many countries underwent successful macroeconomic stabilization programs, accompanied in some cases by large trade reforms and fundamental institutional innovations like granting autonomy to their central banks. During the 1990s and extending into the 2000s, the region's GDP grew at an average annual rate of 3.3 percent (on a population-weighted basis), more than double the 1.5 percent rate observed in the 1980s. Inflation, long the region's scourge, has by-and-large been contained in most countries—in 2011, inflation was 10 percent in Argentina (although there is some controversy about this figure), 7 percent in Brazil, and below 4 percent in Chile, Colombia, Mexico, and Peru. This marks a sharp improvement from the 1990s, which included hyperinflationary episodes in Argentina, Brazil, and Peru. With much better monetary policy, substantially lower fiscal deficits, and improved debt management, most countries in the region were resilient to the 2008–2009 world financial crisis. For Latin America as a whole, unemployment rates rose by less than 1 percentage point, and poverty rates continued to decline, albeit at a lower rate than in previous years. Indeed, for the first time in living memory, many governments were able to conduct effective countercyclical macroeconomic

■ *Santiago Levy is Vice President for Sectors and Knowledge, and Norbert Schady is Principal Economic Advisor for the Social Sector, both at the Inter-American Development Bank, which is headquartered in Washington, DC. Their email addresses are slevy@iadb.org and norberts@iadb.org.*

http://dx.doi.org/10.1257/jep.27.2.193. doi=10.1257/jep.27.2.193

Figure 1
Poverty and Inequality in Latin America



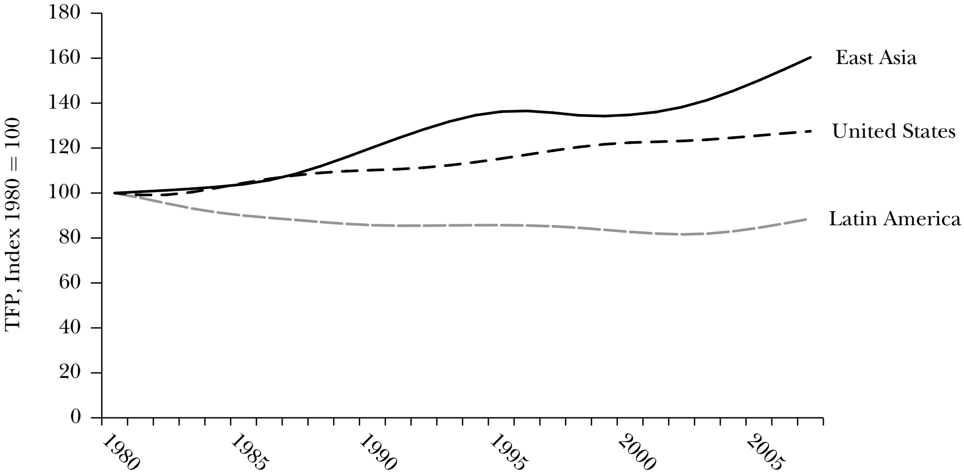
Source: Authors' calculations based on data from the SEDLAC database (Socio-Economic Database for Latin America and the Caribbean) maintained by CEDLAS (Center for Distributional, Labor and Social Studies at the Universidad Nacional de la Plata) and the World Bank, as well as data from the Economic Commission for Latin America and the Caribbean, the Inter-American Development Bank and the World Bank.

policies (Inter-American Development Bank 2012).[1] In 2011, the stock of international reserves for the region stood at $752 billion, including $351 billion in Brazil, $149 billion in Mexico, and more than $40 billion each in Argentina, Chile, and Peru. By comparison, total international reserves for the region were $151 billion a decade ago.

The region's achievements go beyond improved macroeconomic management. The last decade has witnessed substantial and sustained reductions in poverty and inequality, as shown in Figure 1. Poverty fell in virtually every country, and for the region as a whole, the fraction of people living on less than $2.50 per capita per day fell from 26.8 to 13.3 percent, implying that 55 million fewer people lived under the poverty line in Latin America in 2011 than in 2000. The declines in inequality are impressive, too. In 2000, the Gini coefficient was above 0.5 in Argentina, Brazil, Chile, Colombia, Mexico, and Peru, and above 0.45 in Venezuela. By 2011, it had fallen by 6 percentage points or more in Argentina, Brazil, Peru, and Venezuela; by more than 3 percentage points in Chile and Mexico; and by 2 points in Colombia.

[1] In previous crises, many countries in Latin America had weak fiscal positions, high inflation, or scarce access to international financial markets, constraining their ability to use fiscal and monetary policy to boost growth in the face of negative output shocks.

Figure 2
Total Factor Productivity (TFP) in Latin America and East Asia



Source: Authors' calculations based on data from Daude and Fernández-Arias (2010).
Notes: Latin America: Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Honduras, Mexico, Nicaragua, Panama, Peru, Paraguay, El Salvador, Uruguay, and Venezuela. East Asia: Hong Kong, Korea, Malaysia, Singapore, and Thailand.

There have also been steady improvements in various nonmonetary measures of well-being. Between 1990 and 2010, infant mortality in Latin America fell from about 120 to 60 deaths per 1,000 live births, maternal mortality fell from 50 to 25 per 100,000 live births, and chronic malnutrition (or stunting) among children age 5 and younger fell from 25 percent to 12 percent of the population. As we discuss below, there have been continued increases in both school enrollment rates and in the mean years of schooling attained. Finally, and of great importance, the health and education outcomes of girls in the region are now on par with, or surpass, those of boys.

Interregional comparisons help to put these achievements in perspective, however. Growth rates in Latin America continue to be much lower than in East Asia—in the 2000s, GDP grew by 10.4 percent per year in China, 7.1 percent in Vietnam, and 4.1 percent in Korea. At current average growth rates, it will take Brazil 21 years to reach the current GDP per capita of Korea, and Mexico will need 25 years. Moreover, one can also argue that at least part of the good growth performance observed in the last decade is the result of an old-fashioned commodity boom associated with growth in Asia and, at least through 2008, unusually favorable circumstances in international capital markets. Figure 2 compares total factor productivity growth in East Asia, Latin America, and the United States from 1980 to 2007: in East Asia, productivity growth has been faster than in the United States, while in Latin America it was negative up to 2000 (IDB 2010). Since then, productivity growth has been similar in Latin America and the United States, but the region has continued to lose ground relative to East Asia.

In parallel, savings and investments rates in the region continue to be very low relative to those observed in Asia. Savings rates in Latin America range from 18 percent of GDP (in Brazil) to 24 percent (in Mexico and Chile), which contrasts with 31 percent in India, 34 percent in Malaysia, and 51 percent in China. And, despite recent gains, Latin America continues to be the most unequal region of the world and, because of this, is characterized by poverty rates that are much higher than one would expect given the region's income level.[2]

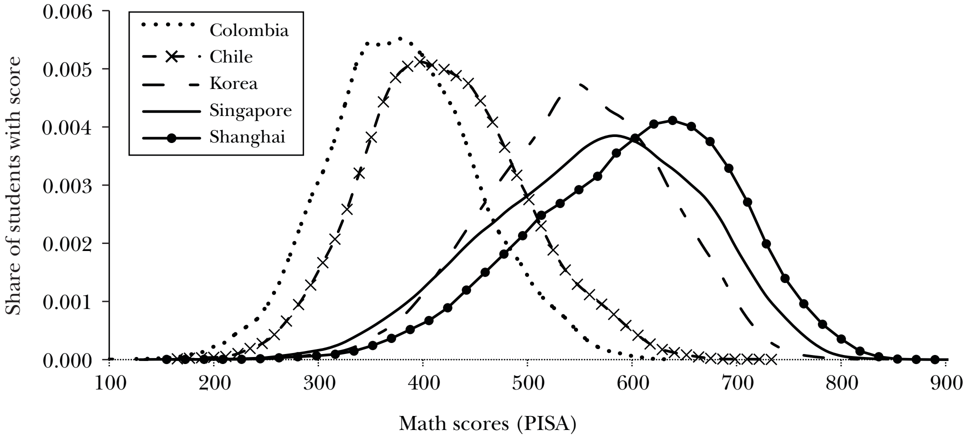
For growth to be both faster and more equitable, Latin America needs to focus sharply on accelerating productivity growth; on raising its savings rate to sustain a larger investment effort (particularly in infrastructure); on increasing the human capital of its workforce; and on new policies to reduce inequality.

In this paper, we focus on a subset of these issues. In particular, we argue that social policy, including human capital and education, social insurance, and redistribution, need special attention if Latin America's achievements over the last two decades are to be sustained and amplified. Starting in the mid 1990s, many governments in the region introduced a variety of programs, including noncontributory pensions and health insurance, and cash transfers targeted to the poor. Social spending in Latin America increased sharply. A growing political and technocratic consensus developed around the need for policies to ensure that the poor have a minimum income floor, are protected from various risks, and have the human capital that will allow them (or their children) to escape poverty. These policies have been widely praised and, like others, we believe they have resulted in substantial improvements in the lives of the poor in the region. However, a more nuanced view shows some worrisome trends. Moving forward, we believe it is necessary to pay much closer attention to the quality of services, particularly in education; to the incentives generated by the interplay of some programs, particularly in the labor market; to a more balanced intertemporal distribution of benefits, particularly between young and old; and to sustainable sources of finance, particularly to the link between contributions and benefits.

An important caveat is in order before we begin: Latin America is a heterogeneous region. Countries differ in size, income levels, institutions, and endowments. What follows are broad strokes that may not always apply to all countries, and any implications for policy would have to be scrutinized and adapted to specific circumstances.

[2] Simple simulations illustrate this point. Uruguay is the least unequal country in Latin America, with a Gini coefficient of 0.40. We simulate poverty levels in Mexico, Brazil, and Chile if their mean income levels were unchanged but the distribution of income was that of Uruguay. These simulations suggest that, in Mexico, which has a Gini coefficient of 0.48, the number of poor people would fall by half; in Brazil, which has a Gini of 0.53, by more than two-thirds; and in Chile, which has a Gini coefficient of 0.52, by more than four-fifths. Of course, there is no realistic set of policies that could redistribute income while keeping mean income at the same level. But the simulations illustrate that the region's unequal distribution of income mechanically results in high levels of poverty.

Figure 3
Math Scores among 15-Year Olds, Latin America and East Asia (2009)



Source: Authors' calculations based on data from OECD Program for International Student Assessment (PISA), 2009.

Schooling

Latin America has an impressive record of expanding the coverage of basic education. Net enrollment rates, given by the fraction of children who are enrolled at the appropriate level for their age, now exceed 90 percent in primary school and are between 60 and 80 percent in secondary school in most countries. Broadly speaking, countries in the region have enrollment rates that are similar to those of other countries with similar income levels (Inter-American Development Bank 2011). On average, individuals born in 1945 in Latin America completed six grades of schooling, while those born in 1985 have completed ten grades.

Unfortunately, increases in schooling levels have not been accompanied by increases in quality. The performance of Latin American students on standardized tests is dismal. Argentina, Brazil, Chile, Colombia, Mexico, Panama, Peru, and Uruguay participated in the International Programme for International Student Assessment (PISA), which tested competencies in language, mathematics, and science for 15 year-olds. Their scores are relatively similar to each other, with Chile and Uruguay performing somewhat better and Panama and Peru somewhat worse. However, students in Panama score below those in Indonesia, even though GDP per capita in Indonesia is about one-third that in Panama, and those in Argentina have scores that are approximately 100 points (one standard deviation) below those in Poland. Only 5 percent of Chilean students score at or above the median score of students in Singapore, and only 1.5 percent score at or above the median of students in Shanghai, as shown in Figure 3; the same is true of Uruguay. Even children in the best schools in the region appear to perform poorly. We calculated the PISA math scores limiting the sample to the 10 percent best-performing schools that participated in this test in Chile and Uruguay, and found that only 10 percent of children in these high-performing schools in Chile, and 13 percent in Uruguay, have scores as high as the average child in Shanghai. Moreover, such comparisons likely understate the differences across countries, because the PISA only tests 15-year-old children currently enrolled in 7th grade or above, and grade repetition and dropout rates are higher in Latin America than in East Asia.

There is a broad consensus among policymakers and researchers that the very low performance of Latin American students on standardized tests has negative implications for productivity. A simple accounting exercise by Hanushek and Woessman (2012) suggests that at least half of the Latin American low-growth "puzzle" can be attributed to low levels of cognitive skills among students, as measured by test scores. Further, because poor children in Latin America generally attend lower-quality schools than their better-off counterparts, it also has negative implications for equity.

Why do Latin American students perform so poorly on international tests? Two reasons are particularly important: factors that affect children before they enter school, and the poor quality of teachers. In the region, many children arrive at the beginning of formal schooling with serious deficits in health and development. Rates of chronic malnutrition (low height-for-age, or stunting) are very high in some countries, especially among the poor. In Guatemala, more than half the children under the age of five are more than two standard deviations behind in height, relative to a reference population of well-nourished children. In Bolivia, Ecuador, and Peru, the number is between 20 and 30 percent, but among the poorest households, especially those in rural areas, the fraction is more than double. Poor nutritional status in early childhood has serious implications for cognitive functioning, and the damage may be largely irreversible.

Other indicators also suggest that poor Latin American children begin schooling already behind. Schady et al. (2012) show that, by the time they enter school, the poorest children in rural Chile are about two-thirds of a standard deviation behind where they should be on their performance on a test (the Spanish-speaking version of the Peabody Picture Vocabulary Test) that has been shown to be highly predictive of school failure; in Colombia and Ecuador these delays are about one-and-a-half standard deviations; and in Nicaragua and Peru, the poorest children in rural areas are more than two standard deviations behind, which implies delays of about two years in their cognitive development.

Although the evidence from Latin America is sparse, there seems to be considerable scope for interventions targeted at young children—especially those in poor households, who exhibit the biggest delays (see Schady 2012 and Vegas and Santibañez 2010 for reviews). In Argentina, children in cohorts and regions that were exposed to a preschool program have test scores in third grade that are 0.23 standard deviations higher than those who were not exposed, have fewer behavior problems and are more likely to pay attention in class and participate (Berlinski, Galiani, and Gertler 2009). In Uruguay, plausibly exogenous variation in access to preschool is associated with 0.8 more years of completed schooling by the time children are 15 years of age (Berlinski, Galiani, and Manacorda 2008). In Colombia, a pilot home-visiting and parenting program improved cognitive development among young children by about 0.3 standard deviations (Attanasio, Fitzsimons, Granthan-McGregor, Meghir, and Rubio-Codina 2012).[3] In Guatemala, a program that distributed a high-protein energy drink known as *Atole* to poor children in early childhood improved chronic malnutrition, schooling, test scores, and men's wages almost 20 years later (Behrman, Calderon, Preston, Hoddinott, Martorell, and Stein 2009; Hoddinott, Maluccio, Behrman, Flores, and Martorell 2008; Maluccio, Hoddinott, Behrman, Martorell, Quisumbing, and Stein 2009).

Teachers in many Latin American countries have deficiencies in content knowledge and basic teaching practices. Peru applied tests of content knowledge to all teachers in 2007. Almost 50 percent of math teachers could not perform basic arithmetic operations and about one-third lacked basic reading comprehension skills. Using data for Peru, Metzler and Woessman (2012) estimate that higher levels of teacher content knowledge in language and (especially) math translate into better learning outcomes for children. In Chile, in any given year, roughly one-third of teachers are deemed to have unsatisfactory performance on the performance evaluation system known as *Docente Más*. In Ecuador, the Classroom Assessment Scoring System (CLASS), a measure of the quality of teaching practices which focuses on socioemotional support, classroom management, and instructional support provided by teachers (Mashburn, Downer, Hamre, Justice, and Pianta 2010; Pianta 2011; Pianta and Hamre 2009) was applied to a sample of teachers between first and third grade. Roughly 90 percent received the lowest possible score of 1, on a scale of 1 to 7, in terms of the instructional support they provide (Araujo, Cruz-Aguayo, and Schady 2012). In Chile, teacher scores are somewhat better than in Ecuador, but only marginally so (Yoshikawa et al. 2012).

In sum, advances in schooling coverage in Latin America are welcome, but are clearly not enough. The interplay of deficiencies generated at an early age with poor-quality teachers and at times inadequate facilities and content implies that children in Latin America, particularly poor ones, are not learning enough and enter the labor market with substantial disadvantages relative to their peers in other

[3] The Colombian program was designed on the basis of a similar program in Jamaica. An efficacy trial of the Jamaica intervention showed that children randomly assigned to receive visits by health paraprofessionals who worked with mothers on early stimulation had cognitive scores that were approximately 0.4 standard deviations higher than children randomly assigned to the control group. Almost two decades after the intervention ended, participants had better performance on tests of math and reading, higher levels of completed schooling (about one-third more years), lower levels of depression and involvement in violent criminal activity, and better labor market outcomes (Grantham-McGregor, Walker, Chang, and Powell 1997; Walker, Grantham-McGregor, Powell, and Chang 2000; Walker, Chang, Powell, and Grantham-McGregor 2005; and Walker, Chang, Vera-Hernández, and Grantham-McGregor 2011). More recent results following up the children in this sample suggest the intervention also raised employment and earnings in early adulthood (Chang et al. 2012).

parts of the world. This is clearly a weak platform from which to improve productivity and reduce the intergenerational transmission of inequality.[4]

Social Insurance

Social insurance aims to protect households against risks—ill health, unemployment, disability, death, or poverty in old age (the latter associated with uncertainty about longevity)—and to contribute to intertemporal consumption smoothing. The central feature of social insurance in most of Latin America is that both the provision and the financing are a function of labor status: in particular, whether a worker is salaried (having a boss and receiving payment in the form of wages) or nonsalaried (self-employed, working on a piece-rate basis, or working in a family firm). As a result, social insurance is intertwined with the functioning of the labor market, with broad implications for the efficacy of these programs, as well as for domestic savings and for productivity (Levy 2008; Ferreira and Robalino 2011).

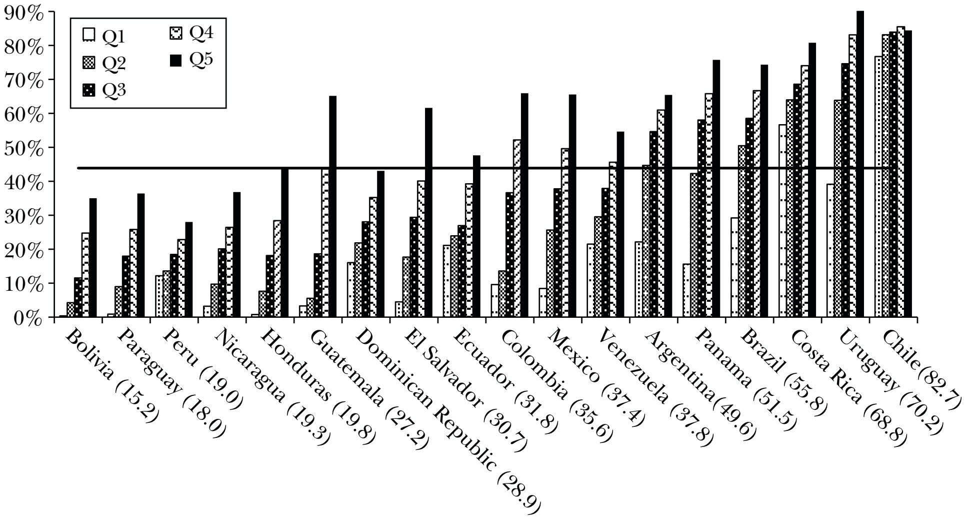
In most counties in Latin America, salaried workers are entitled to a bundle of benefits including, among others, health, work-risk, death and disability insurance, and retirement pensions, and sometimes other benefits like child allowances (Argentina), labor training services (Colombia), or housing and daycare services (Mexico). These benefits are paid from wage taxes. The bundle and the method of financing are usually referred to as "contributory social insurance." In addition, salaried workers are protected against loss of employment through legal requirements for severance pay and related indemnities.[5] In what follows, contributory social insurance should be understood as encompassing employment protection regulations, and health, pensions, and other benefits, as firms and workers must internalize the costs of all these items.

In practice, because regulations are imperfectly enforced, firms sometimes evade and hire salaried workers illegally; and because not all workers participate

[4] "Inequality of opportunity" can be defined as the proportion of total inequality that is explained by predetermined characteristics unrelated to individual effort, like parental education, race, gender, or place of residence. Brunori, Ferreira, and Peragine (2013) argue that inequality of opportunity is higher in Latin America than in other regions. Latin America also has particularly low levels of educational intergenerational mobility—the education of parents is a stronger predictor of the education of children in Latin America than elsewhere (Hertz, Jayasundera, Piraino, Selcuk, Smith, and Verashchagina 2007).

[5] In Latin America, protection against loss of employment typically takes the form of a one-time payment at the time of dismissal, rather than a flow of pre-payments into an insurance fund. Severance pay should in principle be fully internalized by workers in the form of lower wages, with no inefficiency involved. But in practice this is not so, because in Latin America a distinction is made between "just" and "unjust" dismissals (with output adjustment by a firm not considered a just cause for dismissal). This makes severance pay subject to uncertainty, high transaction costs, and delays (an important consideration in the presence of liquidity constraints). In Mexico, for example, trials for severance pay take an average of three years, and it is estimated that lawyers keep about 30 percent of payments (Kaplan, Sadka, and Silva-Mendez 2008; Kaplan and Sadka 2011).

Figure 4
Coverage of Contributory Social Insurance, by Country and Income Quintile



Source: Authors' calculations based on data from Rofman and Oliveri (2012).
Notes: Each bar represents the percentage of employed workers aged 20 and older who are currently contributing to social security; the number in parenthesis after the country name is the average coverage in that country (in percent). The thick horizontal line corresponds to the population-weighted Latin American average.

in the market as salaried employees, contributory social insurance only covers a subset of the labor force, in what is commonly referred to as "formal employment." Figure 4 shows that the share of the labor force covered by contributory social insurance is below 50 percent in most countries in the region, despite the fact that these programs have been obligatory for over half a century.

Until relatively recently, nonsalaried and illegally hired salaried workers—in what is commonly referred to as "informal employment"—were not covered by social insurance in Latin America. Many countries had some publicly provided health care for all, regardless of salaried status, but it was largely limited to basic interventions for maternal and child health. However, over the last two decades, governments across Latin America have created or expanded health, pension, and related programs that are paid from general revenues and thus are referred to as "noncontributory social insurance."

These noncontributory social insurance programs have grown rapidly in terms of budget and coverage. Table 1 lists 13 countries in the region with programs of noncontributory pensions, at an average cost of 0.56 percent of GDP. A number of countries have also introduced noncontributory health programs. The two largest are in Colombia and Mexico, although similar schemes are

Table 1
Noncontributory Pensions and Conditional Cash Transfer Programs in Latin America, 2011

	Noncontributory pensions (NCP)						*Conditional cash transfer programs (CCT)*				
			Coverage		*Transfers*			*Coverage*		*Transfers*	
Country	*NCP*	*Age*	*People (thousand)*	*% of elderly*	*$US (monthly)*	*% GDP*	*CCT*	*Households (thousand)*	*% of households*	*$US*** (monthly)*	*% GDP*
Argentina	√	70+	41	1.4	248	0.03	√	1,876	21.1	162	0.49
Bolivia	√	60+	899	100	28	1.25	√	972	40.2	5	0.23
Brazil	√*	60+	7,340	32.5	328	1.16	√	13,352	28.2	45	0.41
Chile	√	65+	842	53.5	136	0.55	√	264	5.9	104	0.13
Colombia	√	57+	768	15.6	33	0.09	√	2,438	23	33	0.22
Costa Rica	√	65+	93	30.6	146	0.4	√	143	12.7	74	0.23
Dominican Republic							√	831	34.7	25	0.24
Ecuador	√	65+	536	58.2	35	0.34	√	1,212	34	35	0.71
El Salvador	√	60+	20	3.4	50	0.05	√	95	7.1	17	0.15
Guatemala							√	873	37.4	29	0.24
Honduras							√	412	27.7	40	0.32
Mexico	√**	70+	2,149	44.9	40	0.09	√	5,827	24.2	72	0.46
Panama	√	70+	85	56.5	100	0.34	√	74	10	50	0.15
Paraguay	√	65+	25	7.4	92	0.11	√	94	7.1	38	0.13
Peru	√	65+	26	1.5	46	0.01	√	474	7.1	36	0.13
Uruguay	√	65+	33	7.1	238	0.2	√	207	24.7	83	0.48
Latin America			**12,858**	**33.4**	**178**	**0.56**		**29,143**	**22.6**	**64**	**0.37**

Source: Authors' calculation based on data from official records.
* Includes two noncontributory pensions: Benefício de Prestação Continuada and Previdência Rural.
** Includes only the *70 y Más* program, and not the 18 state and municipal pension noncontributory pension programs.
*** Averages based on the demographic structure of household in the poorest quintile.

found in Peru and Nicaragua and, with some variations, in Argentina, Bolivia, and Ecuador. Colombia introduced its *Régimen Subsidiado en Salud* in the early 1990s, initially providing benefits of lower quality to informal workers relative to those offered to formal workers through the *Régimen Contributivo.* Mexico had offered free health services to informal workers through various programs since the 1980s, but in 2003 launched a major effort to expand coverage through the *Seguro Popular.*

These noncontributory programs have substantially expanded the coverage of pensions and health insurance. In Brazil's *Previdência Rural* pension program, coverage among those eligible increased by 40 percentage points within a decade, from a pre-reform level of about 13 percent (Bosch, Cobacho, and Pages forthcoming; Carvalho Filho 2008). Argentina's *Moratorium* increased pension coverage by 27 and 16 percentage points for women and men from a pre-reform level of 30 and 55 percent, respectively (Bosch and Guajarro 2012). Mexico's *Seguro Popular* expanded rapidly: by 2010, it covered more than 43 million affiliates, according to the program's administrative data. And Colombia's *Régimen Subsidiado* increased the coverage of health insurance from levels below 30 percent to more than 90 percent today (Camacho, Conover, and Hoyos 2012).

By extending coverage, noncontributory social insurance programs provide some protection against risks to households who would otherwise be unprotected. In Mexico, the *Seguro Popular* decreased participants' catastrophic health expenditures by 23 percent (King et al. 2009), and the *70 y Más* program, which provides pensions for those 70 years or older, reduced the poverty gap among recipients from 0.61 to 0.46 (Galiani and Gertler 2009). Papers that have modeled the likely effect of the 2008 pension reform in Chile, which introduced a noncontributory pension for the poorest 60 percent of the population, also conclude that it will substantially reduce poverty (Attanasio, Meghir, and Otero 2011; Todd and Joubert 2011). In other cases, the declines in income poverty are smaller because noncontributory pensions allow a larger share of the elderly to stop working (for evidence from Brazil, see Carvalho Filho 2008; for evidence from Mexico, see Galiani and Gertler 2009; Juárez and Pfutze 2012).

Without minimizing these accomplishments, noncontributory social insurance programs do raise various concerns. First, noncontributory programs can be expensive. For example, in the Brazilian *Previdência Rural* program, the value of the pension is large (equivalent to roughly one-third of per capita GDP), and the eligibility age low (60 years for men, 55 for women). As a result, the cost is high, about 0.89 percent of GDP, even though only workers in rural areas are eligible for the program. The Argentinean *Moratorium* cost between 1.5 and 2 percentage points of GDP in the year it was implemented. Moreover, because Latin America's population is aging rapidly, the costs of noncontributory programs could increase substantially in the coming decades. The fraction of the population 65 and over in the Latin America is projected to increase from 7.6 percent of total population in 2010, to 21 percent in 2050; by 2050 there are projected to be three working-age individuals for every person age 65 and older, compared to 8.5 in 2010 (ECLAC 2011). These programs can therefore impose large future liabilities. A related political economy concern is that noncontributory programs face ongoing pressure for increases in coverage and benefits, because unlike contributory programs, benefits and contributions are not directly linked. For example, in 2007 Mexico's *70 y Más* program only covered elderly above age 70 living in towns with less than 2,500 inhabitants; in 2008, it was expanded to include towns with less than 20,000; in 2009, to towns with less than 30,000; in 2012, an election year, eligibility was extended to all, regardless of place of residence; and in 2013, the eligibility age was reduced to 65. A similar situation occurred in Colombia, where the health package financed by the *Régimen Subsidiado* (for informal workers) was initially less generous than that offered by the *Régimen Contributivo* (for formal workers). In 2008, however, the Constitutional Court required that the package of health benefits in the two programs be equated.

The provision of noncontributory social insurance programs can also discourage formal employment, because informal workers receive benefits without paying for them in the form of a wage tax. In 2009, Argentina extended to informal workers the subsidy for minors less than 18 years of age, previously paid only to formal workers, through a program known as the *Asignación Universal por Hijo*.[6] By 2011, the program covered 29 percent of all eligible minors at a cost of 0.64 percent of GDP. Garganta and Gasparini (2012) estimate that this program reduced the probability that informal workers would enter into formal employment by 45 percent for women and 30 percent for men. Bosch and Campos (2010) analyze the impact of Mexico's *Seguro Popular*, comparing the growth rates of firms registered with the Mexican Social Security Institute in municipalities where *Seguro Popular* was implemented early, in 2002–2003, with those where it was implemented later, in 2006–07. They show that, four years after it started, *Seguro Popular* had translated into a 5 percent reduction in formal employment in municipalities that received the program earlier. On the basis of this and similar calculations, Bosch, Cobacho, and Pages (forthcoming) estimate that between 2002 and 2010 *Seguro Popular* resulted in a relocation of between 0.4 and 1 percentage points of total employment from formal to informal jobs, which is equivalent to between 160,000 and 400,000 workers, or between 8 and 20 percent of the total formal jobs created during that period. The analysis by Camacho et al. (2012) of Colombia's *Regimen Subsidiado* suggests qualitatively similar, but substantially larger effects on informality, about 4 percentage points.

More broadly, the segmentation of social insurance into contributory and noncontributory components has three important implications: 1) it reduces the overall efficacy of insurance; 2) it may reduce domestic savings; and 3) it misallocates resources, which can have negative impacts on productivity and growth. We will say a few words about each effect.

The efficacy of insurance is reduced because, across Latin America, transitions from formal to informal employment and back are frequent. In any given year in Argentina and Brazil, 25 and 47 percent of informal workers transit into a formal job, respectively, while 9 and 7 percent transit in the other direction (Ribe, Robalino, and Walker 2012). Because workers only accumulate pension benefits when they are formally employed, contribution densities to pension plans are low. Even in Argentina, Chile, and Uruguay, three of the countries with the highest coverage of contributory social insurance in the region (as shown earlier in Figure 4), mean contribution densities are low at 55 percent, 47 percent, and 58 percent, respectively (Forteza, Luchetti, and Pallares 2009). As a result, the replacement rate, the amount of before-retirement income replaced by the pension, will be low, and contributory pensions will do a poor job helping individuals to smooth consumption between work and retirement. Also, transits between formal and informal employment result

[6] Argentina's *Asignación Universal por Hijo* can also be thought of as a conditional cash transfer program of the sort discussed in the next section. The program was means-tested, was limited to households with children, and benefits were conditioned on school enrollment and use of preventive health services by young children. However, unlike other conditional cash transfer programs, the *Asignación Universal por Hijo* extended to informal workers a benefit that formal workers already received, with the key difference that the benefit for formal workers was paid for by a wage tax while that for informal workers was paid for out of general revenues.

in erratic coverage against risks that are only covered by contributory social insurance (like death, disability, and loss of employment).

In terms of effects on saving, many countries in Latin America reformed their pension systems in the 1990s, often replacing defined benefit with defined contribution programs (Uthoff 2011). One goal of these reforms was to increase the supply of long-term domestic savings denominated in local currency. Although mandating increased pension savings can in principle be offset by reduced private saving, the balance of the available evidence from Latin America shows that this offset has not occurred on a one-to-one basis, implying that forced savings through contributory pensions have increased domestic savings (Aguila 2011; Carpio 2008; Cerda 2008; Quintanilla 2011). However, because the coverage of contributory pensions is so limited, the contribution of these programs to national saving is also limited. The extension of noncontributory pension programs is likely to further undermine individual incentives to save through the contributory system.

Finally, the segmentation of social insurance may be a factor behind the stagnation of productivity growth in the region. On the one hand, if benefits from contributory social insurance are not fully valued by workers, contributory programs act like a tax on formal employment.[7] On the other hand, noncontributory social insurance programs function as a subsidy to informal employment because informal workers receive at least some of the same benefits as formal workers, with the critical difference that they do not pay for them from foregone wages. Moreover, when contributory social insurance and regulations are imperfectly enforced, firms may hire salaried workers illegally but remain inefficiently small to minimize the chance of detection. Taken together, these taxes and subsidies distort the price of labor towards more small firms with salaried labor, more family firms with nonsalaried labor, and more informal employment. In Mexico, Busso, Fazio, and Levy (2012) find that 90 percent of the 3.6 million firms captured in the Census have up to five workers, 96 percent up to ten, and only 1 percent more than 50; but less than three in four firms are registered with the Social Security Institute (a prerequisite for formality).[8] An emerging literature shows that productivity in the informal sector is lower than in the formal one, so any resource movements from the former to the latter will tend to lower aggregate productivity (Fajnzylber, Maloney, and Montes-Rojas 2009, 2011; Busso, Fazio, and Levy 2012; Jung and Tran 2012; Pagés 2010).

[7] The evidence on this point for Latin America is mixed. Important references include Almeida and Carneiro (2012), Bergolo and Cruces (2012), Cruces, Galiani, and Kidyba (2010), Gruber (1997), Kugler and Kugler (2009), and Levy (2008).

[8] Moreover, this count probably overstates the share of larger firms, because the Census only accounts for about 50 percent of private employment. The rest of private employment is either self-employment, or occurs in small firms not captured by the Census, like street stands or street markets. Firms in Latin America may choose to be inefficiently small (from a social point of view) for a variety of reasons, in addition to the incentive to avoid social insurance regulations. For example, many countries have special tax regimes whereby smaller firms face substantially lower rates of value-added or income taxes (Pagés 2010).

In sum, over a half century after they were introduced, contributory social insurance programs across Latin America provide only low levels of coverage, especially among the poor. The noncontributory social insurance programs introduced over the last two decades have extended a measure of protection against some risks to those who would otherwise be uncovered, and they have also led to substantial reductions in poverty among the elderly. However, the interplay of these two systems distorts incentives in the labor market and generates costly trade-offs between extending the coverage of social insurance, on the one hand, and productivity, savings, and fiscal considerations, on the other.

Inequality and Social Assistance

Countries in Latin America have experienced highly unequal distributions of income since at least the middle of the nineteenth century, if not before.[9] Inequality in the region appears to have been particularly high in the 1980s—a result of numerous macroeconomic crises and the adjustment processes that followed (Gasparini and Lustig 2011; Gasparini, Cruces, and Tornarolli 2011; López-Calva and Lustig 2010). But the last decade has seen substantial reductions in inequality. Moreover, inequality fell in countries with different political orientations, with relatively large and small governments, with large and small shares of social spending, and with historically high and low levels of inequality. For this reason, the search for causes of the inequality decline in Latin America has focused on broad trends cutting across the region.

Decompositions suggest there are two main explanations for the decline in inequality in the region in the last decade: A reduction in the wage premium for skilled labor, and increases in the coverage of social programs. The rate of return to schooling in the region increased in the 1980s and 1990s, but fell in the 2000s. As a result, the share of labor income at the bottom of the distribution increased, which accounts for between one-third and one-half of the inequality decline. Meanwhile, the expansion in pension coverage (mainly noncontributory pensions, although data limitations often make it difficult to separate contributory and noncontributory pensions in household surveys) accounts for maybe 5 percent of the decline, and programs that make cash transfers to the poor account for roughly one-quarter of the decline.[10]

[9] The question of just when Latin America began to exhibit high inequality of incomes is under dispute. Robinson and Sokoloff (2004) and Sokoloff and Engerman (2000) argue that Latin America has been unequal for centuries, a result of the European conquest, the abundant natural resources (in particular minerals), and the comparative advantage of the region in the production of crops such as sugar. Williamson (2009) argues that "historical persistence in Latin American inequality is a myth" and that the region only became unusually unequal in the second half of the nineteenth century.

[10] This estimate is based on our own calculations, using the data in Azevedo, Inchautse, and Sanfelice (2012). See also Lustig, López-Calva, and Ortiz-Juárez (2013).

Given the importance of labor incomes, in particular for the poor, on the distribution of income, a number of studies have attempted to isolate the role that shifts in demand and supply have played in the changes in returns to schooling (Aedo and Walker 2012; Gasparini, Galiani, Cruces, and Acosta 2011; Manacorda, Sánchez-Páramo, and Schady 2010). These papers conclude that changes in the supply of workers with different amounts of education explain only a modest fraction of the changes in the skill premium observed in Latin America in the last three decades. Demand-side changes, in particular skill-biased technological change, substantially increased the wage premium for workers with more education (especially university education) in Latin America in the 1990s.[11] However, for reasons that are insufficiently understood, the effects of these demand-side shocks appear to have petered out by the 2000s. Institutional changes in the labor market were also important. The real minimum wage has increased substantially in some countries—by more than 50 percent in Brazil and by 200 percent in Uruguay between 2004 and 2010. The minimum wage has been shown to be an important determinant of the distribution of earnings in some countries in the region (Aedo and Walker 2012; Bosch and Manacorda 2010; Maloney and Núñez Mendez 2004).

Of the various programs that directly redistribute resources to poor households in Latin America, the best known and most studied are targeted cash transfer programs, which were pioneered in the region and have become popular since the late 1990s. In these programs, eligibility is generally determined not by income directly, but by a composite measure of household characteristics, assets, and access to social services that are correlated with consumption or income. Some of these programs, but not all, are "conditional cash transfers" that require households to comply with a number of conditions in return for the cash—generally, preventive health check-ups for children and pregnant mothers and attendance at school for school-aged children. These conditional cash transfer programs are in some cases quite large. Table 1 shows that in 16 countries, average coverage is one out of every four households. Careful randomized evaluations of conditional cash transfer programs in Latin America have shown that they have substantially increased school enrollment and attendance, and preventive health care utilization (Fiszbein and Schady 2009, and the references therein). Children in beneficiary households complete more schooling—for example, Behrman, Parker, and Todd (2011) conclude that three years of *PROGRESA* transfers in Mexico result in approximately 0.3 more years of completed schooling (relative to no transfers). Conditional cash transfers have also had substantial effects on poverty and inequality.

While conditional cash transfers have clearly had overall positive effects, there are three main concerns. First, while cash transfer programs have increased school enrollment, the evidence on whether the additional schooling results in better learning outcomes for children who were brought into school by these programs

[11] Skill-biased technological change may have been transmitted through trade (Sánchez-Páramo and Schady 2003). Other references include Acosta and Gasparini (2007), Behrman, Birdsall, and Székely (2007), and Galiani and Sanguinetti (2003).

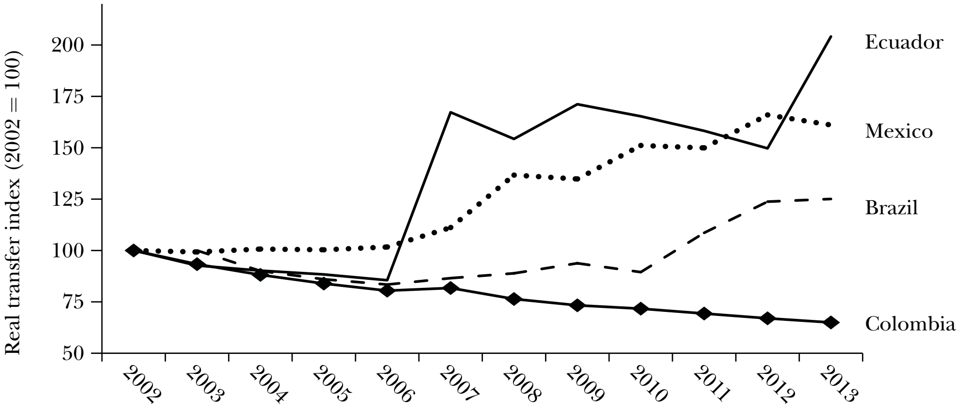
is mixed. In Nicaragua, Barham, Marcours, and Maluccio (2013) find that boys (but not girls) whose families received transfers from the *Red de Protección Social* program when they were between 9 and 11 years of age have test scores that are approximately 0.2 standard deviations higher ten years later; in Mexico, Behrman, Parker, and Todd (2009) find that children who received *PROGRESA* transfers do not have higher test scores than comparable children who did not receive them. Although the reasons for this finding are unclear, the poor quality of education and the fact that many of the children who were brought into school are drawn from the lower end of the distribution of ability are probably part of the explanation.[12] More generally, while conditional cash transfers have increased the utilization of health and educational services, impacts on final human capital outcomes have been more limited. As a result, the effect that conditional cash transfers have on reducing the intergenerational transmission of poverty—a key objective of these programs—may be limited (Levy 2007).

Second, the transfers may in some cases be so large that they can have a negative effect on incentives to work. Figure 5 presents the evolution of transfers for four of the biggest cash transfer programs in Latin America. In Ecuador and Mexico, transfers have increased substantially in magnitude. In Mexico's *PROGRESA*, since renamed *Oportunidades*, transfers are now equivalent to over 40 percent of household pretransfer income in the lowest quintile of the distribution. Transfer income also represents a sizeable share of total income in Ecuador, where the program is also very large in scope, covering almost 40 percent of the population. If leisure is a normal good, we might expect that the income effect of transfers of this magnitude would reduce labor supply.[13] Moreover, many of the cash transfer programs in Latin America periodically "recertify" beneficiaries to ensure that they are still poor, to qualify for continued eligibility. This obviously introduces an incentive for households to continue to be (or at least appear to be) poor. Camacho and Conover (2011) show that, once the exact formula used to calculate the proxy means test that determined eligibility for Colombia's *Familas en Acción* program was made public, there was substantial heaping of households just below the cutoff value. In Chile (and in some other countries), having a household member with a physical

[12] The evidence on this point from developing countries outside the region is also mixed. Filmer and Schady (2009) analyze the effect of a conditional cash transfer-like program in Cambodia. They find no effect of transfers on test scores, in spite of large effects on school enrollment and years of completed schooling. Baird, McIntosh, and Ozler (2011) analyze a pilot program in Malawi which randomly assigned children to conditional transfers, unconditional transfers, and a control group. The conditional transfer had a larger effect on school enrollment than the unconditional transfer, and the conditional transfer (but not the unconditional transfer) had a positive, but modest effect on test scores.

[13] Results from early evaluations of conditional cash transfers in Latin America suggest that adults in recipient households generally did not reduce labor supply in response to transfers (Alzúa, Cruces, and Ripani forthcoming; Skoufias and Di Maro 2008; Parker and Skoufias 2000). It is not clear, however, whether these results hold for programs that have been in place for a decade or longer, and given that transfer amounts have increased. Indeed, the evidence on noncontributory pensions in Latin America, discussed above, as well as the ample literature on welfare programs in the United States, all suggest that reductions in labor supply are a real possibility.

Figure 5
The Evolution of Transfers in Cash Transfer Programs in Latin America



Source: Authors' calculations based on eligibility criteria from conditional cash transfers and household surveys from Latin American countries—Mexico: ENIGH (Encuesta Nacional de Ingresos y Gastos de los Hogares); Ecuador: ENEMDU (Encuesta Nacional de Empleo, Desempleo y Subempleo); Brazil: PNAD (Pesquisa Nacional de Amostro de Domicílios); Colombia: GEIH (Gran Encuesta Integrada de Hogares).

or mental disability increases the value on the proxy means, and this is well known. Herrera, Larrañaga, and Telias (2010) show that, among the poorest households, almost 80 percent report having a household member with a disability on the *Ficha de Protección Social,* the survey that is used to construct the proxy means. In comparison, on the national CASEN household survey, which does not determine eligibility for transfers, about 20 percent of the poorest households report having a member with a disability.

Third, conditional cash transfers can also favor informal over formal employment (in addition to the effects of contributory and noncontributory social insurance programs discussed earlier). The Uruguayan *PANES* program explicitly disqualified recipients if their formal sector earnings increased above a predetermined value. Amarante, Manacorda, Vigorito, and Zerpa (2011) show that *PANES* substantially reduced formal employment among men and that these effects persisted at least two years after the program ended.

Despite the amount of attention from academics and policymakers, neither cash transfers nor noncontributory pensions are always the largest programs that seek to redistribute resources to the poor in Latin America. In many countries, transfers to *all* households through subsidies to energy prices, or exemptions on consumption or value-added taxes on particular goods (usually food, medicines, or transport) absorb a larger share of the budget than either cash transfers or noncontributory social insurance, or both. In Venezuela, energy subsidies represent 6.9 percent of GDP; in Ecuador 6.4 percent; in the Dominican Republic

5.5 percent, and in Argentina, 1.8 percent (International Energy Agency 2011). Exemptions from value-added taxes cost around 2 percent of GDP in Costa Rica, Mexico, Colombia, and Guatemala, 1.6 percent in Peru and Ecuador, and 0.8 percent in Argentina (Corbacho, Cibils, and Lora 2013). Because richer households spend more in absolute value on electricity, cooking oil, and gasoline as well as on tax-exempt goods like food and medicine, these subsidies and exemptions disproportionately benefit better-off households. In Mexico, the residential electricity subsidy is larger than the budget of *Oportunidades*, yet 57 percent of the electricity subsidy goes to households in the top two income quintiles, compared with about 6 percent for the lowest one. Universal subsidies and tax exemptions are very inefficient ways of redistributing resources to the poor.

In sum, about half of the reduction in inequality in Latin America in the last decade has occurred for reasons other than the expansion in the coverage of social programs. Declines in the returns to education and the resulting increase in the share of total labor income among the poor have been particularly important. However, social programs, especially conditional cash transfers, have also played a role. That said, and recognizing the variations across countries, the scope for additional redistribution through cash transfers or noncontributory pensions appears limited, because increasing the value of the transfers is likely to have increasingly negative effects on the labor supply of recipients, accentuate distortions between the informal and formal labor market, or affect incentives to save. This argument does not imply that the region should abandon efforts to further reduce poverty and inequality. Instead, it suggests that other policies are likely to be more effective in the future. At present, personal income taxes in Latin America are among the lowest in the world, collecting only 1.4 percent of GDP versus an average of 8.4 percent of GDP in developed countries (Corbacho, Cibils, and Lora 2013). Increasing tax revenues, particularly from higher-income households, would contribute to lower inequality. In parallel, the region could benefit from redirecting some of the resources now channeled through generalized energy subsidies and exemptions from value-added taxes to invest more in early childhood development, design effective labor training programs, or fund health and other components of social insurance for all workers, regardless of their labor status. These are all measures that would increase incomes, particularly incomes of poor households, via the route of higher productivity.

Conclusions

Although many factors have been at play, increased social spending and new social programs have helped to reduce poverty and inequality across Latin America. What accounts for these changes in social policy in the last two decades? We highlight three important explanations. First, the emergence of more democratic regimes in the 1990s renewed political pressures to respond to unacceptable levels of poverty and inequality, in particular after the “lost” decade of the 1980s. Second, greater macroeconomic stability facilitated growth, providing fiscal revenues to increase spending (aided in some cases by favorable international conditions); it also allowed policymakers to focus on issues other than the latest adjustment program with the International Monetary Fund. Third, there was a recognition that traditional social programs had had only limited success. In contexts of high informality, contributory social insurance had failed to protect the majority of households from risks; in countries with high income inequality (that is, most of Latin America), generalized subsidies were mostly captured by higher income groups. Lessons from the past were converted into pioneering initiatives to provide households in the informal sector with health services (like Colombia's *Régimen Subsidiado*) and pensions (like Brazil's *Previdência Rural*); or to focus income transfers on the poor while turning them into investments in their human capital (like Mexico's *PROGRESA*).

As a result of faster growth, more social spending and new programs, millions of Latin Americans are now eating better, attaining more schooling, enjoying improved access to health services, and having higher incomes during old age. The region's "middle class"—those living in households with income per capita between $10 and $50 (in US dollars) per day—increased from 20 to 30 percent of the population between 1990 and 2010, and their share in national income increased from 40 to 50 percent (Ferreira, Messina, Rigolini, López-Calva, Lugo, and Vakis 2012). These economic and social developments offer a more fertile ground where the rule of law and stronger institutions can develop deeper roots.

Has a turning point been reached? Has Latin America put in place policies to sustain faster rates of growth and higher real incomes, based more on increasing productivity and less on the gains derived from macroeconomic stability and favorable Asian winds? Put differently, can the region in the next decades reach incomes per capita on par with, say, Korea, or rather will it be caught in a "middle income trap"—more stable and less poor, but not truly prosperous and equitable? On this key question, the jury is still out.

There is much that Latin America can and should do to build truly prosperous and equitable societies. Critically, in our view, the region needs to recognize that faster growth requires improved productivity, not just factor accumulation and favorable international conditions. In parallel, the region needs a deeper understanding of the factors limiting productivity growth and the policy reforms required to stimulate it. We believe that at present there is no consensus across Latin America that accelerating productivity growth is essential to achieving faster growth, and even less consensus as to what policies are needed to accomplish it. This situation stands in contrast with the consensus that emerged two decades ago that macroeconomic stability was essential to resuming growth, and that sound monetary policy and prudent fiscal management were needed to deliver stability.

What role does social policy play in all this? First and foremost, it should be concerned with raising social welfare. More educated citizens will result in a better informed electorate and, more generally, in improved institutions. Broader protections against risks and wider opportunities will help to reverse disparities that, sadly, have for too long been Latin America's trademark. But social policy, in our view, should also contribute to productivity growth, or at least not hinder it. In the end, one cannot sustain a welfare state on stagnant productivity, particularly since the costs of that welfare state will increase rapidly as the region's population ages and its epidemiological profile evolves towards more costly pathologies. We want the humanistic and civic value of education, but Latin America also needs competent engineers, nurses, and computer programmers; we want protection against risks, but Latin America also need firms and workers that pay taxes, invest in labor training and innovate; we want less poverty, but Latin America also needs to avoid permanent welfare dependency and unduly reduced participation rates. Higher average per capita incomes are needed if only to provide the revenue base from which social programs can be financed—a key point in a region that in the past suffered much from unsustainable fiscal deficits.

Social policy needs to change because, as we have argued, at present it is only partly effective in reaching its direct objective of higher welfare, but also because it is making an insufficient contribution to productivity growth (and in some cases, is working against it). More particularly, social policy in Latin America needs to address four challenges. First, it needs to increase abilities among children and young adults, to create the skilled workforce that can support sustained productivity growth. Of particular relevance are interventions that seek to ensure that poor children do not fall behind at early ages in terms of their nutrition and cognitive and noncognitive development. Put differently, there is not enough "pre-distribution" (to borrow a term from Heckman 2012) in the form of investments early on. These early investments need to be followed with education that is of higher quality. The poor quality of teachers in many countries in Latin America requires better pre-service and in-service training. But, we believe, it also requires greater flexibility to reward good teachers and dismiss poor ones. In a recent paper on the United States, Rockoff and Staiger (2010) argued in this journal that, while teachers can improve learning outcomes, the observable characteristics of teachers (including degrees, test scores and experience, at least after the first two to three years) explain very little of the difference in their effectiveness. On this basis, Rockoff and Staiger propose a teacher selection system that requires few upfront teaching-specific investments, recruits widely, but gives tenure to only a very small minority of teachers. The conclusions of their analysis seem to us to be especially relevant for Latin America, where many teachers are ineffective but, thanks to very powerful unions in many countries, teachers often receive automatic tenure upon entering the public education system.

Second, the coexistence of contributory and noncontributory social insurance is only partly effective in protecting workers against risks and does a poor job in smoothing consumption between work and retirement. Moreover, the way in which social insurance is financed, including the use of cumbersome mechanisms to protect against the risks of employment loss, distorts firm and worker behavior, and may also decrease long-term saving. Much research remains to be done in this area, but these distortions could have substantial, negative consequences for productivity in Latin America. Correspondingly, reforms in the design of these programs could offer substantial benefits. Attention needs to be focused on unifying the source of financing for health insurance; more generally, risks that are common to any form of employment should be financed from the same revenue source. Inefficient labor protection regulations should be replaced by proper unemployment insurance. A clearer distinction should be made between the two objectives of pensions: Avoiding old-age poverty (which can be ensured with a modest noncontributory pension that does not depend on whether a worker is in formal or informal employment, and may or may not be means-tested), and smoothing consumption between work and retirement (which will likely require reforms that seek to increase the coverage of the contributory pension system). Insofar as possible, pension programs should not distort labor–leisure and formal–informal choices, and should not discourage long-term saving in local currency.

Third, while direct income transfers are clearly part of an effective social policy, transfer programs have their limits. Indeed, over-reliance on cash transfers may hurt those that they are intended to help, and also depress productivity, if it lowers labor market participation rates, increases informality, and results in long-term welfare dependency. Conditional cash transfers have also been used as a way of protecting households from temporary shocks, even though it is not clear that they are the right instrument for this purpose. Many of the challenges faced by the current beneficiaries of conditional cash transfer programs are better addressed by policies that improve the quality of services and the functioning of labor markets. Given the coverage and transfer amounts in place in many countries today, further redistribution may be better accomplished by reforming personal income taxes, or by redirecting generalized subsidies embedded in energy prices and special tax regimes into investments in early childhood development, in labor training, or in improving the quality of health and education services.

Finally, Latin America needs to ensure that its social programs, including pensions, health care, cash transfers, and expenditures on education are fiscally sustainable in the long run, and are not vulnerable to the vagaries of international commodity prices. Governments should also make more explicit the link between contributions and benefits in all social insurance programs, as this will help address the political economy considerations that have driven the growth of benefits in noncontributory systems (Antón, Hernández, and Levy 2012).

The solutions to these problems are technically complex, and may need to be implemented in a global economic context that is less favorable than it has been in the recent past. But these complexities are dwarfed by the political challenges, as the needed reforms touch core fibers of the region's social fabric: the relations between parents, teachers, and the government; subsidies and taxes; and the interactions between firms and workers in the labor market. The societal consensuses that, after many painful crises, were built around the importance of prudent macroeconomic policies need to be extended to areas where vested interests and strongly held beliefs make such consensuses more difficult to reach.

■ *The authors are with the Inter-American Development Bank, but their opinions do not necessarily reflect those of the institution they are affiliated with. We thank Yessenia Loayza for excellent research assistance, and Jere Behrman, Mariano Bosch, Guillermo Cruces, Francisco Ferreira, Luis Felipe López Calva, Eduardo Lora, Nora Lustig, Bill Maloney, Angel Melguizo, and Marcos Robles for many useful comments. We also thank the editors of The* Journal of Economic Perspectives *for their constructive suggestions.*

References

Acosta, Pablo, and Leonardo Gasparini. 2007. "Capital Accumulation, Trade Liberalization, and Rising Wage Inequality: The Case of Argentina." *Economic Development and Cultural Change* 55(4): 793–812.

Aedo, Christian, and Ian Walker. 2012. *Skills for the 21*st *Century in Latin America and the Caribbean.* World Bank, Washington, DC.

Aguila, Emma. 2011. "Personal Retirement Accounts and Savings." *American Economic Journal: Economic Policy* 3(4): 1–24.

Almeida, Rita, and Pedro Carneiro. 2012. "Enforcement of Labor Regulation and Informality." *American Economic Journal: Applied Economics* 4(3): 64–89.

Alzúa, María Laura, Guillermo Cruces, and Laura Ripani. Forthcoming. "Welfare Programs and Labor Supply in Developing Countries. Experimental Evidence from Latin America." *Journal of Population Economics.*

Amarante, Verónica, Marco Manacorda, Andrea Vigorito, and Mariana Zerpa. 2011. "Social Assistance and Labor Market Outcomes: Evidence from the Uruguayan PANES." Inter-American Development Bank, Technical Note IDB-TN-453.

Antón, Arturo, Fausto Hernández, and Santiago Levy. 2012. *The End of Informality in Mexico? Fiscal Reform for Universal Social Insurance.* Inter-American Development Bank, Washington, DC.

Araujo, Caridad, Yyannú Cruz Aguayo, and Norbert Schady. 2012. "The Effects of Teacher Characteristics on Learning Outcomes among Young Children in Ecuador." *Paper presented at the 17*th *Annual Meeting of the Latin American Economic Association, Lima, Peru.*

Attanasio, Orazio, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina. 2012. "Stimulation and Early Childhood Development in Colombia: The Impact of a Scalable Intervention." Paper presented at "Promises for Preschoolers: Early Childhood Development and Human Capital Accumulation" Conference, University College London.

Attanasio, Orazio, Costas Meghir, and Alejandro Otero. 2011. "Pensions, Work, and Informality: The Impact of the 2008 Chilean Pension Reform." Unpublished paper, University College London.

Azevedo, Joao Pedro, Gabriela Inchautse, and Viviane Sanfelice. 2012. "Decomposing the Recent Inequality Decline in Latin America." Unpublished paper, World Bank.

Baird, Sarah, Craig McIntosh, and Berk Ozler. 2011. "Cash or Condition? Evidence from a Cash Transfer Experiment." *Quarterly Journal of Economics* 126(4): 1709–53.

Barham, Tania, Karen Macours, and John Maluccio. 2013. "More Schooling *and* More Learning? Effects of a 3-Year Conditional Cash Transfer Program in Nicaragua after 10 Years." Unpublished paper.

Behrman, Jere R., Nancy Birdsall and Miguel Székely. 2007. "Economic Policy Changes and Wage Differentials in Latin America." *Economic Development and Cultural Change* 56(1): 57–97.

Behrman, Jere R., Maria C. Calderon, Samuel H. Preston, John Hoddinott, Reynaldo Martorell, and Aryeh D. Stein. 2009. "Nutritional Supplementation in Girls Influences the Growth of Their Children: Prospective Study in Guatemala." *American Journal of Clinical Nutrition* 90(5): 1372–79.

Behrman, Jere R., Susan W. Parker, and Petra E. Todd. 2009. "Medium-Term Impacts of the Oportunidades Conditional Cash Transfer Program on Rural Youth in Mexico." In *Poverty, Inequality and Policy in Latin America*, edited by Stephan Klasen and Felicitas Nowak-Lehman, 219–70. Cambridge, MA: MIT Press.

Behrman, Jere R., Susan W. Parker, and Petra E. Todd. 2011. "Do Conditional Cash Transfers for Schooling Generate Lasting Benefits? A Five-Year Followup of PROGRESA/Oportunidades." *Journal of Human Resources* 46(1): 93–122.

Bergolo, Marcelo, and Guillermo Cruces. 2012. "Work and Tax Incentive Effects of Social Insurance Programs: Evidence from an Employment Based Benefit Extension." Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2154274.

Berlinski, Sebastian, Sebastian Galiani, and Paul Gertler. 2009. "The Effect of Pre-Primary Education on Primary School Performance." *Journal of Public Economics* 93(1–2): 219–34.

Berlinski, Sebastian, Sebastian Galiani, and Marco Manacorda. 2008. "Giving Children a Better Start: Preschool Attendance and School-Age Profiles." *Journal of Public Economics* 92(5–6): 1416–40.

Bosch, Mariano, and Raymundo Campos. 2010. "The Trade-offs of Social Assistance Programs in the Labor Market: The Case of the Seguro Popular Program in Mexico." *Serie documentos de trabajo del Centro de Estudios Económicos* 2010-12.

Bosch, Mariano, Belen Cobacho, and Carmen Pages. Forthcoming. "Taking Stock of Eight Years of Implementation of Seguro Popular in Mexico." In *Social Insurance and Labor Markets: How to Protect Workers while Creating Good Jobs,* edited by Markus Frölich, David Kaplan, Carmen Pages, David Robalino, and Jamele Rigolini. Oxford University Press.

Bosch, Mariano, and Jarret Guajarro. 2012. "Labor Market Impacts of Non-Contributory Pensions: The Case of Argentina's *Moratorium.*" IInter-American Development Bank Working Paper IDB-WP-366

Bosch, Mariano, and Marco Manacorda. 2010. "Minimum Wages and Earnings Inequality in Urban Mexico. *American Economic Journal: Applied Economics* 2(4): 128–49.

Brunori, Paolo, Francisco H. G. Ferreira, and Vito Peragine. 2013. "Inequality of Opportunity, Income Inequality, and Economic Mobility: Some International Comparisons." World Bank Policy Research Working Paper 6304.

Busso, Matías, Victoria Fazio, and Santiago Levy. 2012. "(In)Formal and (Un)Productive: The Productivity Costs of Excessive Informality in Mexico." Inter-American Development Bank Working Paper 341.

Camacho, Adriana, and Emily Conover. 2011. "Manipulation of Social Program Eligibility." *American Economic Journal: Economic Policy* 3(2): 41–65.

Camacho, Adriana, Emily Conover, and Alejandro Hoyos. 2012. "Effects of Colombia's Social Protection System on Workers' Choice between Formal and Informal Employment." Unpublished paper, Universidad de los Andes.

Carpio, Miguel Angel. 2008. "The Effects of Social Security Privatization on Consumption, Saving and Welfare: Evidence from Peru." Ph.D. Dissertation, Department of Economics, Universitat Pompeu Fabra.

Carvalho Filho, Irineu Evangelista. 2008. "Old-Age Benefits and Retirement Decisions of Rural Elderly in Brazil." *Journal of Development Economics* 86(1): 129–46.

Cerda, Rodrigo. 2008. "Social Security and Wealth Accumulation in Developing Economies: Evidence from the 1981 Chilean Reform." *World Development* 36(10): 2029–44.

Chang, Susan, Sally Grantham-McGregor, Paul Gertler, James Heckman, Rodrigo Pinto, Christel Vermeersch, Suzan Walker, and Ariana Zanolini. 2012. "Labor Market Returns to Early Childhood Stimulation: A 20 Year Follow-up to the Jamaican Study." Paper presented at "Promises for Preschoolers: Early Childhood Development and Human Capital Accumulation" Conference, University College London.

Corbacho, Ana, Vincente Fretes Cibils, and Eduardo Lora, eds. 2013. *More than Revenue: Taxation as a Development Tool.* Inter-American Development Bank and Palgrave Macmillan.

Cruces, Guillermo, Sebastian Galiani, and Susana Kidyba. 2010. "Payroll Taxes, Wages and Employment: Identification through Policy Changes." *Labour Economics* 17(4): 743–49.

Daude, Christian, and Eduardo Fernández-Arias. 2010. "Productivity and Factor Accumulation in Latin America and the Caribbean: A Database." Inter-American Development Bank, Washington, DC. http://www.iadb.org/research/pub_desc.cfm?pub_id=DBA-015.

Economic Commission for Latin America and the Caribbean (ECLAC). 2011. *Long-Range Population Projections.* Demographic Observatory, Year 6, No. 11. Santiago de Chile. Pdf document: http://www.eclac.cl/cgi-bin/getProd.asp?xml=/publicaciones/xml/1/46771/P46771.xml&xsl=/celade/tpl/p9f.xsl&base=/celade/tpl/top-bottom.xslt. (There is also an Excel spreadsheet, which is sometimes updated: http://www.eclac.cl/celade/proyecciones/basedatos_BD.htm.)

Fajnzylber, Pablo, William Maloney, and Gabriel Montes-Rojas. 2009. "Releasing Constraints to Growth or Pushing on a String? Policies and Performance of Mexican Micro-Firms." *Journal of Development Studies* 45(7): 1027–47.

Fajnzylber, Pablo, William Maloney, and Gabriel Montes-Rojas. 2011. "Does Formality Improve Micro-firm Performance? Evidence from the Brazilian Simples Program." *Journal of Development Economics* 94(2): 262–76.

Ferreira, Francisco H. G., Julian Messina, Jamele Rigolini, Luis-Felipe López-Calva, Maria Ana Lugo, and Renos Vakis. 2012. *Economic Mobility and the Rise of the Latin American Middle Class.* Washington DC: World Bank.

Ferreira, Francisco H. G., and David Robalino. 2011. "Social Protection in Latin America: Achievements and Limitations." In *The Oxford Handbook of Latin American Economics*, edited by Jose Antonio Ocampo and Jaime Ros, 836–62. Oxford University Press.

Filmer, Deon, and Norbert Schady. 2009. "School Enrollment, Selection, and Test Scores." World Bank Policy Research Working Paper 4998.

Fiszbein, Ariel, and Norbert Schady. 2009. *Conditional Cash Transfers: Reducing Present and Future Poverty*. Washington, DC: World Bank.

Forteza, Alvaro, Leonardo Luchetti, and Montserrat Pallares. 2009. "Measuring the Coverage Gap." In *Closing the Coverage Gap: The Role of Social Pensions and Other Retirement Income Transfers*, edited by Robert Holzmann, David Robalino, and Noriyuki Takayama, 23–40. Washington, DC: World Bank.

Galiani, Sebastian, and Paul Gertler. 2009. "Primer Seguimiento a la Evaluación de Impacto del Programa de Atención a Adultos Mayores de 70 Años y Más en Zonas Rurales (Programa 70 y Más)." Unpublished paper, SEDESOL.

Galiani, Sebastian, and Pablo Sanguinetti. 2003. "The Impact of Trade Liberalization on Wage Inequality: Evidence from Argentina." *Journal of Development Economics* 72(2): 497–513.

Garganta, Santiago, and Leonardo Gasparini. 2012. "El Impacto de un Programa Social sobre La Informalidad Laboral: El Caso de La AUH en Argentina." CEDLAS Working Paper 133, Universidad Nacional de La Plata.

Gasparini, Leonardo, Guillero Cruces, and Leopoldo Tornarolli. 2011. "Recent Trends in Income Inequality in Latin America." *Economía* 11(2): 147–90.

Gasparini, Leonardo, Sebastian Galiani, Guillermo Cruces, and Pablo Acosta. 2011. "Educational Upgrading and Returns to Skills in Latin America: Evidence from a Supply-Demand Framework, 1990–2010." World Bank Policy Research Working Paper 5921.

Gasparini, Leonardo, and Nora Lustig. 2011. "The Rise and Fall of Income Inequality in Latin America." Working Paper 1110, Tulane University, Department of Economics.

Grantham-McGregor, Sally, Susan P. Walker, Susan M. Chang, and Christine A. Powell. 1997. "Effects of Early Childhood Supplementation with and without Stimulation on Later Development in Stunted Jamaican Children." *American Journal of Clinical Nutrition* 6(2): 247–53.

Gruber, Jonathan. 1997. "The Incidence of Payroll Taxation: Evidence from Chile." *Journal of Labor Economics* 15(3, Part 2): S72–S101.

Hanushek, Eric A., and Ludger Woessmann. 2012. "Schooling, Educational Achievement, and the Latin American Growth Puzzle." *Journal of Development Economics* 99(2): 497–512.

Heckman, James. 2012. "Promoting Social Mobility." Available at http://www.bostonreview.net/BR37.5/ndf_james_heckman_social_mobility.php.

Herrera, Rodrigo, Osvaldo Larrañaga, and Amanda Telias. 2010. "La Ficha de Protección Social." In *Las Nuevas Políticas de Protección Social en Chile,* edited by Osvaldo Larrañaga and Dante Contreras, 265–96. Santiago, Chile: Uqbar Editores.

Hertz, Tom, Tamara Jayasundera, Patrizio Piraino, Sibel Selcuk, Nicole Smith, and Alina Verashchagina. 2007. "The Inheritance of Educational Inequality: International Comparisons, and Fifty-Year Trends." *B.E. Journal of Economic Analysis and Policy* 7(2): 1–46.

Hoddinott, John, John A. Maluccio, Jere R. Behrman, Rafael Flores, and Reynaldo Martorell. 2008. "Effect of a Nutrition Intervention during Early Childhood on Economic Productivity in Guatemalan Adults." *Lancet* 371(9610): 411–16.

International Energy Agency. 2011. *World Energy Outlook 2011: Energy Subsidies.* http://www.iea.org/weo/subsidies.asp. Database accessed on October 7, 2012.

Juárez, Laura, and Tobias Pfutze. 2012. "The Effects of a Non-Contributory Pension Program on Labor Participation: The Case of *70 y Más* in Mexico." Unpublished paper, Instituto Tecnológico Autónomo de México and Oberlin College.

Jung, Juergen, and Chung Tran. 2012. "The Extension of Social Security Coverage in Developing Countries." *Journal of Development Economics* 99(2): 439–58.

Kaplan, David S., Joyce Sadka, and Jorge Luis Silva-Mendez. 2008. "Litigation and Settlement: New Evidence from Labor Courts in Mexico." *Journal of Empirical Legal Studies* 5(2): 309–350.

Kaplan, David S., and Joyce Sadka. 2011. "The Plaintiff's Role in Enforcing a Court Ruling: Evidence from a Labor Court in Mexico." InterAmerican Development Bank, Working Paper IDB-WP-264.

King, Gary, Emmanuela Gakidou, Kosuke Imai, Jason Lakin, Ryan T. Moore, Clayton Nall, Nirmala Ravishankar, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila, and Héctor Hernández Llamas. 2009. "Public Policy for the Poor? A Randomised Assessment of the Mexican Universal Health Insurance Programme." *Lancet* 373(9673): 1447–54.

Kugler, Adriana, and Maurice Kugler. 2009. "Labor Market Effects of Payroll Taxes in Developing Countries: Evidence from Colombia." *Economic Development and Cultural Change* 57(2): 335–58.

Levy, Santiago. 2007. *Progress against Poverty: Sustaining Mexico's Progresa-Oportunidades Program.* Washington, DC: Brookings Institution Press.

Levy, Santiago. 2008. *Good Intentions, Bad Outcomes: Social Policy, Informality and Economic Growth in Mexico.* Washington, DC: Brookings Institution Press.

López-Calva, Luis Felipe, and Nora Lustig. 2010. *Declining Inequality in Latin America: A Decade of Progress?* Brookings Institution Press, Washington, DC.

Lustig, Nora, Luis Felipe López-Calva, and Eduardo Ortiz-Juárez. 2013. "Declining Inequality in Latin America in the 2000s: The Cases of Argentina, Brazil, and Mexico." *World Development* 44(April): 129–141.

Maloney, William F., and Jairo Núñez Mendez. 2004. "Measuring the Impact of Minimum Wages: Evidence from Latin America." In *Law and Employment: Lessons from Latin America and the Caribbean,* edited by James J. Heckman and Carmen Pages, 109–130. University of Chicago Press.

Maluccio, John. A., John Hoddinott, Jere R. Behrman, J. Reynaldo Martorell, Agnes Quisumbing, and Aryeh D. Stein. 2009. "The Impact of Experimental Nutritional Interventions on Education into Adulthood in Rural Guatemala." *Economic Journal* 119(537): 734–763.

Manacorda, Marco, Carolina Sánchez-Páramo, and Norbert Schady. 2010. "Changes in Returns to Education in Latin America: The Role of Demand and Supply of Skills." *Industrial and Labor Relations Review* 63(2): 307–26.

Mashburn, Andrew J., Jason T. Downer, Bridget K. Hamre, Laura M. Justice, and Robert C. Pianta. 2010. "Consultation for Teachers and Children's Language and Literacy Development during Pre-Kindergarten." *Applied Developmental Science* 14(4): 179–96.

Metzler, Johannes, and Ludger Woessman. 2012. "The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation." *Journal of Development Economics* 99(2): 486–96.

Pagés, Carmen, ed. 2010. *The Age of Productivity: Transforming Economies from the Bottom Up.* Inter-American Development Bank and Palgrave Macmillan, New York.

Pagés, Carmenk. 2011. *Social Strategy for Equity and Productivity in Latin America and the Caribbean.* Washington, DC: Inter-American Development Bank.

Pagés, Carmen. 2012. *The World of Forking Paths: Latin America and the Caribbean in the Face of Global Economic Risks.* Washington, DC: Inter-American Development Bank.

Parker, Susan W., and Emmanuel Skoufias. 2000. "The impact of PROGRESA on Work, Leisure, and Time Allocation." International Food Policy Research Institute. http://www.ifpri.org/publication/impact-progresa-work-leisure-and-time-allocation.

Pianta, Robert C. 2011. "Teaching Children Well: New Evidence-Based Approaches to Teacher Professional Development and Training." Center for American Progress.

Pianta, Robert. C., and Bridgett Hamre. 2009. "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity." *Educational Researcher 38*(2): 109–119.

Quintanilla, Ximena. 2011. "The Effect of the Chilean Pension Reform on Wealth Accumulation." Working Paper 47, Superintendencia de Pensiones, Santiago, Chile.

Ribe, Helena, David Robalino, and Ian Walker. 2012. *From Right to Reality: Incentives, Labor Markets, and the Challenge of Universal Social Protection in Latin America and the Caribbean.* Washington, DC: World Bank.

Robinson, James, and Kenneth Sokoloff. 2004. "Historical Roots of Latin American Inequality." In *Inequality in Latin America and the Caribbean: Breaking with History?* edited by David de Ferranti, Guillermo Perry, Francisco H. G. Ferreira, and Michael Walton. Washington, DC: World Bank.

Rockoff, Jonah E., and Douglas Staiger. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24(3): 97–118.

Rofman, Rafael, and Maria Laura Oliveri. 2012. "Pension Coverage in Latin America: Trends and Determinants." Social Protection & Labor Discussion Paper 1217, World Bank.

Sánchez-Páramo, Carolina, and Norbert Schady. 2003. "Off and Running? Technology, Trade and the Rising Demand for Skilled Workers in Latin America." World Bank Policy Research Working Paper 3015.

Schady, Norbert. 2012. "El Desarrollo Infantil Temprano en América Latina y el Caribe: Acceso, Resultados y Evidencia Longitudinal de Ecuador." In *Educación para la Transformación*, edited by Marcelo Cabrol and Miguel Székely, 53–92. Washington, DC: Inter-American Development Bank.

Schady, Norbert, Jere Behrman, Maria Caridad Araujo, Rodrigo Azuero, Raquel Bernal, David Bravo, Florencia Lopez-Boo, Karen Macours, Daniela Marshall, Christina Paxson and Renos Vakis. 2012. "Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries." Paper presented at "Promises for Preschoolers: Early Childhood Development and Human Capital Accumulation" Conference, University College London, June 25.

Skoufias, Emmanuel, and Vincenzo Di Maro. 2008. "Conditional Cash Transfers, Adult Work Incentives, and Poverty." *Journal of Development Studies* 44(7): 935–60.

Sokoloff, Kenneth L., and Stanley L. Engerman. 2000. "Institutions, Factor Endowments, and Paths of Development in the New World." *Journal of Economic Perspectives* 14(3): 217–32.

Todd, Petra, and Clement Joubert. 2011. "The Impact of Chile's 2008 Pension Reform on Labor Force Participation, Pension Savings, and Gender Equity." Unpublished paper, University of Pennsylvania.

Uthoff, Andras. 2011. "Social Security Reforms in Latin America." In *The Oxford Handbook of Latin American Economics,* edited by Jose Antonio Ocampo and Jaime Ros, pp. 863–85. Oxford University Press.

Vegas, Emiliana, and Lucrecia Santibañez. 2010. *The Promise of Early Childhood Development in Latin America*. World Bank, Washington, DC.

Walker, Susan P., Susan M. Chang, Christine A. Powell, and Sally Grantham-McGregor. 2005. "Effects of Early Childhood Psychosocial Stimulation and Nutritional Supplementation on Cognition and Education in Growth-Stunted Jamaican Children: Prospective Cohort Study." *Lancet* 366(9499): 1804–1807.

Walker, Susan P., Susan M. Chang, Marcos Vera-Hernández, and Sally Grantham-McGregor. 2011. "Early Childhood Stimulation Benefits Adult Competence and Reduces Violent Behavior." *Pediatrics* 127(5): 849–57.

Walker, Susan. P., Sally Grantham-McGregor, Christine A. Powell, and Susan M. Chang. 2000. "Effects of Growth Restriction in Early Childhood on Growth, IQ, and Cognition at Age 11 to 12 Years and the Benefits of Nutritional Supplementation and Psychosocial Stimulation." *Journal of Pediatrics* 137(1): 36–41.

Williamson, Jeffrey G. 2009. "History without Evidence: Latin American Inequality since 1491." NBER Working Paper 14766.

Yoshikawa, Hiro, D. Leyva, Catherine Snow, E. Treviño, M. C. Barata, C. Weiland, C. and M. C. Arbour. 2012. "Interim Impacts on Classroom Quality of an Initiative to Improve the Quality of Preschool Education in Chile: A Cluster-Randomized Trial." Unpublished paper, Harvard University.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 219–238

The Investment Strategies of Sovereign Wealth Funds[†]

Shai Bernstein, Josh Lerner, and Antoinette Schoar

Sovereign wealth funds have emerged as major investors in corporate and real resources worldwide. Estimates of their size are difficult, because disclosure regulations and practices differ widely from country to country. But in 2012, the Sovereign Wealth Fund Institute estimated that total assets of these funds were more than $5 trillion: that is, more than double the $2.1 trillion managed by hedge funds (as estimated by Hedge Funds Research Inc., accessed July 21, 2012), although it is only 2.3 percent of the $212 trillion in total global financial assets (as estimated by McKinsey Global Institute 2011).

At first blush, sovereign wealth funds might seem an excellent opportunity for nations with high variance in public revenues to ensure steady cash flow levels and provide resources for long-term investments: for example, countries relying on commodity trade that occasionally encounter windfalls of natural resources. Such countries, without a fund to direct investments, could otherwise fall prey to the "Dutch disease" and squander short-lived windfalls from natural resources in a way that weakens the economy's long-run potential. But sovereign wealth funds also have limitations, since they may create economic distortions. For example, there are concerns about lack of transparency and political capture: funds with political leaders on their boards may be tempted to shore-up domestic firms as they succumb to political pressure, passing up on high net present value investments in other

■ *Shai Bernstein is Assistant Professor of Finance, Graduate School of Business, Stanford University, Stanford, California. Josh Lerner is the Jacob H. Schiff Professor of Investment Banking, Harvard Business School, Harvard University, Boston, Massachusetts. Antoinette Schoar is Michael Koerner '49 Professor of Entrepreneurial Finance, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts. Lerner and Schoar are affiliates of the National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are shaib@gsb.stanford.edu, josh@hbs.edu, and aschoar@mit.edu.*

[†] To access the Appendix, visit
http://dx.doi.org/10.1257/jep.27.2.219.

doi=10.1257/jep.27.2.219

firms and creating product market distortions by favoring connected or poorly performing firms. Similarly, as the interaction between sovereign wealth funds and political agenda grows, opportunities for nepotism increase, potentially reducing the overall skill of sovereign wealth fund managers relative to professionals and diluting the returns.

Thus, sovereign wealth funds are particularly interesting because of the potential interactions between mission and ownership structure. Their investment charters usually state that the fund seeks to maximize financial returns for the benefit of long-term public policies, such as retiree benefits or economic development needs. But the quasi-public nature of these funds means that they are exposed to political influences, often with more short-term goals.

This article will review several of the central issues that face sovereign wealth funds. After an overview of their magnitude, we will then consider the institutional arrangements under which many of the sovereign wealth funds operate and how such arrangements might influence the effectiveness of their investment policies. We focus on a specific set of agency problems that is of first-order importance for these funds: that is, the direct involvement of political leaders in the management process. We show that sovereign wealth funds with greater involvement of political leaders in fund management are associated with investment strategies that seem to favor short-term economic policy goals in their respective countries at the expense of longer-term maximization of returns. In particular, sovereign wealth funds where political involvement is more prevalent tend to support domestic firms by investing in segments and markets where valuation levels are inflated (as measured by price/earnings ratios), and subsequently see a reversal in these price/earnings ratios. The opposite patterns hold for funds that rely on external managers. While we are not able to disentangle causality with our existing data, the associations are striking.

Sovereign wealth funds face several other issues, like how best to cope with demands for transparency, which can allow others to copy their investment strategies, and how to address the problems that arise with sheer size, like the difficulties of scaling up investment strategies that only work with a smaller value of assets under investment. In the conclusion, we discuss how various approaches cultivated by effective institutional investors worldwide—from investing in the best people to pioneering new asset classes to compartmentalizing investment activities—may provide clues as to how sovereign wealth funds might address these issues.

An Overview of Sovereign Wealth Funds

Depending on how one counts, there are between 40 and 70 different sovereign funds, run by political entities as disparate as New Mexico and Kazakhstan. Table 1 lists the 20 largest sovereign wealth funds and estimates of their holdings: the funds on this list comprise about 90 percent of the total assets of sovereign wealth funds. The wealth within these funds has differing origins. In many of the most visible cases, such as Abu Dhabi, petroleum has been the source of abundant wealth. Other

Table 1
Leading Sovereign Wealth Funds

Country	*Fund Name*	*Assets (billions of dollars)*	*Inception*	*Origin of wealth*
UAE – Abu Dhabi	Abu Dhabi Investment Authority	627	1976	Oil
Norway	Government Pension Fund – Global	593	1990	Oil
China	SAFE Investment Company	568	1997	Non-commodity
Saudi Arabia	SAMA Foreign Holdings	533	N/A	Oil
China	China Investment Corporation	440	2007	Non-commodity
Kuwait	Kuwait Investment Authority	296	1953	Oil
China – Hong Kong	Hong Kong Monetary Authority Investment	293	1993	Non-commodity
Singapore	Government of Singapore Investment Corporation	248	1981	Non-commodity
Singapore	Temasek Holdings	158	1974	Non-commodity
Russia	National Welfare Fund	150	2008	Oil
China	National Social Security Fund	135	2000	Non-commodity
Qatar	Qatar Investment Authority	100	2005	Oil
Australia	Australian Future Fund	80	2006	Non-commodity
UAE – Dubai	Investment Corporation of Dubai	70	2006	Oil
UAE – Abu Dhabi	International Petroleum Investment Company	65	1984	Oil
Libya	Libyan Investment Authority	65	2006	Oil
Kazakhstan	Kazakhstan National Fund	58	2000	Oil
Algeria	Revenue Regulation Fund	57	2000	Oil
UAE – Abu Dhabi	Mubadala Development Company	48	2002	Oil
South Korea	Korea Investment Corporation	43	2005	Non-commodity

Note: This information about the 20 largest sovereign wealth funds is compiled from the Sovereign Wealth Fund Institute, http://www.swfinstitute.org/fund-rankings/ (accessed July 21, 2012).

commodities, from diamonds to copper or phosphates, have been the foundation of other funds, like the Chilean sovereign fund (though none of these funds made it onto the list of the top 20 funds). Still others have been primarily funded from the proceeds from the sale of state-owned properties or businesses. Other funds, such as those of China and Singapore, have their origin in trade surpluses.

Sovereign wealth funds are growing quickly. They increased ten-fold in the last two decades: from $500 billion in 1990 to more than $5 trillion today. Over the past three years, they have achieved a 24 percent annual growth rate. Much of this growth has been driven (not surprisingly) by the rising price of petroleum, and has been concentrated in producer nations such as Norway, the United Arab Emirates, and Kuwait. But other important players include nations such as China that pile up foreign currency because they run persistent, large trade surpluses. These countries less and less often put these reserves "under a mattress"—that is, holding safe but low-return US Treasury bonds—and are instead seeking broader portfolios.

Sovereign funds frequently have multiple goals, which different organizations emphasize to varying extents. There are three distinct roles sovereign wealth funds can play. First, they can serve as a source of capital for future generations, especially in countries where future generations may no longer be able to rely on commodities for a steady stream of revenue. For example, the nation of Kiribati is a collection of islands in the Pacific Ocean (formerly known as the Gilbert Islands) with a population of under 100,000 residents. For many decades, the dominant export from the country was guano, bird droppings used for fertilizer. The island's leaders set up the Kiribati Revenue Equalization Reserve Fund in 1956, and imposed a tax on production by foreign firms. The last guano was extracted in 1979, but the fund remains a key economic contributor. At $600 million, it is ten times the size of the nation's gross domestic product, and the interest generated by the fund represents 30 percent of the nation's revenue. Such a use is similar to that of a university that receives a major bequest: typically, these funds are not spent immediately, but instead added to its endowment so it can benefit many cohorts of students. Second, sovereign wealth funds can play a stabilizing role by reducing the volatility of government revenues. Countries that depend on commodities for the bulk of their exports can be whipsawed by shifts in prices, as, for instance, many oil exporters were in the mid-1980s and late 1990s. Finally, these funds can serve as holding companies, in which the government places its strategic investments. Public leaders may see fit to invest in domestic or foreign firms for strategic purposes, and the sovereign funds provide a way to hold and manage these stakes.

The Mixed Legacy

Many nations have failed to save the wealth created by developing natural resources. Consider, for instance, the experience of Norway in the 1970s and 1980s (for more details, see Pope 1995; Gjedrem 2005). In the oil surge of those years, the government received a tremendous windfall of funds from its numerous rigs in the North Sea. While efforts were made to enact legislation that set aside money for the future, most of the money was spent immediately. Some of the spending benefited physical and social infrastructure: Norway rebuilt its excellent system of roads and bridges and provided free health care and higher education to all residents. But other expenditures were less beneficial for long-term growth. For example, minimum wages were set extremely high, which rendered a number of economic sectors uncompetitive in global markets, and industries were subsidized. Much of the funding for industry was earmarked for dying sectors, such as shipbuilding. This support allowed facilities to remain open for a few years more, but could not reverse the inexorable decline of such industries. Much of the funding for new ventures went to friends or relatives of parliamentarians or of the bureaucrats responsible for allocating the funds. Moreover, Norway's policy of aggressively spending the government's petroleum revenues brought chaos to public and private finances when oil prices plunged in the mid-1980s. The government's oil revenue dropped from about $11.2 billion in 1985—or about 20 percent of Norway's gross domestic product—to $2.4 billion in 1988. The resulting retrenchment of public spending and tightening of credit led numerous banks to fail, as well as bringing an unprecedented wave of bankruptcies by private citizens.

Nor was Norway the first nation to struggle with the influx of wealth. Back in the 1970s, *The Economist* magazine coined the term "Dutch Disease" to describe the economic malaise that gripped the Netherlands when it experienced an influx of natural gas royalties during the 1960s. An example much further back in time, documented by historian David Landes (1998), would be the corrosive effects that the tremendous wealth generated by Spain's overseas conquests had on that nation's economy.

Sovereign wealth funds can address these downsides of a sudden accumulation of natural wealth in two ways. First, by not spending the gains from natural resources (or other sources) immediately, but rather preserving them for future generations, the distorting impact of the windfall is reduced. Had the Norwegian government kept public spending in check during the 1970s and 1980s, it is unlikely that the disruptions in subsequent years would have been as severe. Second, earmarking a percentage of windfall revenues into an investment fund may reduce the risk that government officials will spend these revenues in an unwise or corrupt manner—assuming, of course, the sovereign fund is run in a professional manner. In an ideal world, a soundly managed sovereign fund can address some of the macroeconomic problems that an influx of funding may cause, such as inflation and exchange rate overvaluation (see the discussion in Ang 2010 for an exploration of these issues).

But the structure of sovereign wealth funds can face two serious agency problems. First, the political process can introduce short-run pressures on sovereign wealth funds to financially support local firms or subsidize industrial policies within the country. There are two opposing views of the consequences of these investment pressures. Advocates for government-directed investments often argue that financial markets in these countries can be underdeveloped or myopic or both, and thus leave profitable investment opportunities on the table (Atkinson and Stiglitz 1980; Stiglitz 1993). The opposing, less-sanguine view of politically directed investments suggests that political involvement can either lead to misguided policy attempts to prop up inefficient firms or industries or engage in investment activities in industries, sectors, or geographies that are "hot" (Shleifer and Vishny 1994; Banerjee 1997; Hart, Shleifer, and Vishny 1997).

This conceptual framework suggests some testable implications. If the benevolent view of sovereign wealth funds is accurate, we would expect to find that government investments in local firms are directed at industries that face financial constraints and subsequently perform very well. If the latter view is true, we would predict the opposite: investments would be disproportionately directed to local firms, follow a pro-cyclical trend, and subsequently perform poorly. In addition, if sovereign wealth funds are run by politically connected but financially inexperienced managers, we might expect that not only would they make poor choices in their home and foreign investments, but would also display poorer stock-picking ability even looking solely at the international portfolio of the fund.

Political Involvement and Investment Distortions

There has been relatively little empirical analysis of agency problems at sovereign funds, largely due to data restrictions.[1] Recent papers by Gompers and Metrick (2001), Lerner, Schoar, and Wongsunwai (2007), and Hochberg and Rauh (2011) have highlighted the heterogeneity in investment strategies, and ultimately returns, across different types of institutional investors.[2] Because we are interested in understanding the extent to which the investment behavior of sovereign wealth funds is shaped by short-term political considerations, we focus on the funds' long-term investments—acquisitions, purchases of private equity, and structured equity positions in public firms—on the grounds that these distortions should be most evident in these areas.

Descriptive Statistics

To analyze the investment strategies of sovereign wealth funds, we combine data from a number of publicly available sources. Here, we offer an overview of the sources for this data: for details, please see the online Appendix available with this article at http://e-jep.org.

First, we look at information on the funds themselves, starting with profiles of the funds published by J.P. Morgan (Fernandez and Eschweiler 2008) and Preqin (Friedman 2008). The key variables collected at the fund level are assets under management, the presence of politicians in the managing bodies of the funds, reliance on external managers, and whether the stated investment goals are "strategic." By "strategic," we mean that the investments are related to the country's long-term industry development strategy rather than simply aiming to maximize the financial returns of the portfolio. We categorize a fund as "strategic" if its stated investment goals are the management of the government's physical assets, the acquisition of strategic assets, or domestic development. We categorize a fund as "nonstrategic" if its stated goals are investment of oil/commodity revenues, currency reserve management, or pension funding. These measures of the characteristics of the funds are admittedly crude characterizations of organizational structures: these are recorded

[1] Several papers conduct event-study analyses of how the stock market reacts when sovereign wealth funds make investment announcements. The reactions are usually positive, at least in the short term (Kotter and Lel 2008; Dewenter, Han, and Malatesta 2010; Bortolotti, Fotak, Megginson, and Miracky 2010; Knill, Lee, and Mauck 2010). Chhaochharia and Laeven (2009) show that sovereign wealth funds largely invest in countries that share the same ethnicity, language, and religion. Fernandes (2011) and Dyck and Morse (2011), rather than exploring transactions, focus on holdings of sovereign wealth funds (that is, the stock rather than the flow of investments). The latter paper, which is most complementary to the analysis below, finds that many holdings by these funds can be explained by financial return maximization or state planning motives, demonstrating the tension between the two objectives.

[2] While the Santiago principles of the International Working Group of Sovereign Wealth Funds state that "relevant financial information regarding the SWF should be publicly disclosed" (http://www.iwg-swf.org/pubs/gapplist.htm; accessed December 22, 2012), consistent return data for most sovereign funds is hard to come by. For instance, the Abu Dhabi Investment Authority in recent annual reports has reported its aggregate returns over 20- and 30-year time horizons; aggregate returns over shorter horizons have not been disclosed, much less those of individual asset classes.

as binary variables, rather than as continuous variables that we might be able to analyze more carefully. Moreover, these measures are reported as of 2008: we do not have a time series on the governance structure or types of advisors involved in the funds.

Second, we examine the direct investments that the funds made, relying on reporting from Dealogic's M&A Analytics, SDC's Platinum M&A, and Bureau van Dijk's Zephyr. Transactions included in the database encompass outright acquisitions, venture capital and private equity investments, and structured minority purchases in public entities (frequently called PIPEs, or "private investments in public entities"). The databases do not include investments into hedge, mutual, or private equity funds or open market purchases of minority stakes in publicly traded firms.

Finally, we want to look at the investment climate around the time of the transaction and to measure investment performance. Because many investments are in private firms, price/earnings ratios determined in public equity markets are not available. As a proxy, we use the price/earnings ratios of firms traded in stock markets in the target company's industry and nation, where the price/earnings ratios are weighted by the size of the firms in the industry. We construct this price/earnings measure both for the time when the sovereign wealth fund first makes the acquisition, and for a year later, which give us an admittedly approximate performance measure for each deal.

The result of this process is a sample of 29 sovereign wealth funds that carried out 2,662 transactions between January 1984 and December 2007. The assets of these funds, $3.1 trillion, represent about 60 percent of the assets of sovereign wealth funds according to the Sovereign Wealth Fund Institute. The bulk of sovereign wealth funds that are not included are very new, very small, or have traditionally eschewed private equity investing (for example, the Norwegian Government Pension Fund and China's SAFE Investment Company).

Table 2 presents descriptive statistics for this sample. Panel A of Table 2 sorts the funds into three regions: seven funds in the Asian group, 15 funds in the Middle Eastern group, and seven funds in the Western group. The Western group includes funds from North America, Australia, and Europe. The sample of 2,045 transactions by the Asian funds is substantially larger than the 533 observations in the Middle Eastern group and the 84 of the Western group.[3] While the sample consists of transactions between the years 1984 and 2007, transactions are more common

[3] One possible explanation for these differences in sample size is that we have only partial coverage of the deals. We believe, however, that this can only explain part of the differences. More important, we believe, are the differences in fund sizes and the willingness to engage in direct investments. To estimate the coverage of our sample, we compare the aggregate transaction value of our sample to the estimate in a J.P. Morgan publication (Fernandez and Eschweiler 2008). They estimate outstanding investments by sovereign wealth funds in alternatives investments like hedge funds and private equity at the end of 2007 as $316 billion. In our sample, the aggregate transaction value in the years 2003–2007 (excluding the public investments) is $198 billion (expressed in 2008 US dollars). Given that we include direct private equity investments but exclude private equity partnerships and hedge fund investments while they include all three, the comparison suggests we have reasonable sample coverage.

Table 2
Descriptive Statistics

Panel A: Groups

	Funds	*Transactions*	*External managers (%)*	*Politicians (%)*	*Average fund size in 2008 (billions of dollars)*
Asia group	7	2045	42.85	57.14	132.7
Middle East group	15	533	13.33	13.33	124.76
Western group	7	84	42.85	14.28	40.874

Panel B: Transactions

	N	*Mean*	*Median*	*Std. Dev*
Acquisition stake (%)	1,998	56.59	50.00	39.01
Average deal size (million 2008$)	1,743	158.23	67.50	256.24
Home investment (%)	2,662	33.92	0.00	47.35
Region Investment (%)	2,662	29.70	0.00	45.70
P/E Level	2,642	25.60	21.46	13.48
P/E Change (%)	2,632	–1.17	–0.01	11.19
Market-adjusted Return	543	4.67	13.20	42.82

Continued

in recent years: more than half of the Asian group transactions, 60 percent of the Middle Eastern group transactions, and 90 percent of the Western group transactions happened in the most recent five years of the sample.

Panel B shows that the average transaction size is $158 million (in 2008 US dollars), although the median is much lower at $67 million. The average stake acquired by the sovereign wealth funds is a majority interest of 56.6 percent. The average price/earnings level in the industry-country-year of the target of a transaction is 25.6, and the typical investment segment experiences a drop of –1.2 percent in the mean price/earnings ratio in the year after the investment. For approximately 20 percent of the investments that occurred in publicly traded firms, we also examine the market-adjusted returns in the six months after the transaction (as discussed further below). Sovereign wealth funds have played an important role in private equity investing. [4]

Panel C reports on the funds according to their governance structure. About 24 percent of the funds (20 percent of transactions) have politicians involved in the fund, and 28 percent of the funds (10 percent of transactions) rely primarily on

[4] Over the years 2003 through 2007, the aggregate value of private equity transactions by sovereign wealth funds in our sample was $198 billion (excluding investments by sovereign wealth funds in private equity partnerships). Based on estimates of Stromberg (2008) and the Private Equity Council, investments by sovereign wealth funds account for approximately 9.5 percent of the aggregate value of global private equity deals over a similar time period.

Table 2—continued

Panel C: Politicians and external managers

	Politicians are involved in the management of the fund			*Politicians are not involved in the management of the fund*			*External managers are involved in the management of the fund*			*External managers are not involved in the management of the fund*		
	N	*Mean*	*Median*	*N*	*Mean*	*Median*	*N*	*Mean*	*Median*	*N*	*Mean*	*Median*
Acquisition stake (%)	366	49.16	40.03	1,625	58.35	50.75	203	42.19	20.00	1,788	58.31	51.00
Average deal size (million 2008$)	378	190.07	72.49	1,367	147.86	66.01	219	236.41	85.48	1,526	145.62	64.82
Home investment (%)	508	44.80	0.00	2,146	31.39	0.00	275	8.20	0.00	2,379	36.89	0.00
Region Investment (%)	508	31.80	0.00	2,146	29.32	0.00	275	44.02	0.00	2,379	28.17	0.00
P/E Level	506	25.29	21.66	2,128	25.70	20.51	272	19.68	21.52	2,362	25.57	21.46
P/E Change (%)	502	−2.62	−0.01	2,122	−0.82	−0.01	269	2.48	0.00	2,355	−0.82	−0.01

Panel D: Funds' stated investment objectives (Preqin 2008)

	Investment of oil/commodity revenues	*Currency reserve management*	*Pension funding*	*Management of government physical assets*	*Acquisition of strategic assets*	*Domestic development*
Number of funds	5	8	2	7	4	3
Number of transactions	73	1833	34	460	722	178
Home investment (%)	11	32	32	59	33	32

Notes: The sample consists of 2,662 investments by 29 sovereign wealth funds. It excludes transactions that were withdrawn or rejected. All descriptive statistics are equally weighted. *P/E Level* is the average of the price/earnings ratios of publicly traded firms in the industry, country, and year of the transaction. *Region Investment* equals 1 (100 percent) if investment was at the same region (Asia, Middle East, or Western countries) but not at home country. *P/E Change* is the change in the average of the price/earnings ratios of publicly traded firms in the industry and country of the transaction in the year after the deal. *Market-adjusted Return* is the difference between the return of the target in the six months after the transaction and the return of the corresponding benchmark over the same period. The deal size and the price/earnings ratio variables are winsorized.

outside managers. Both funds with political leaders and external managers tend to make larger investments. Interestingly, when politicians are involved, funds invest more in firms headquartered in the home country (45 percent of the deals in the sample) relative to funds without their involvement (only 31 percent of the transactions). Funds with primarily external managers invest less in the home country (only 8 percent) relative to funds that do not rely on external managers (these invest 37 percent in the home country).

The final panel of Table 2 reports the stated fund objectives. Currency reserve management is the objective associated with most funds and most transactions. Funds whose stated goal is the management of government physical assets have the largest share of domestic investments; those whose goal is the investment of oil/commodity revenues, the fewest.

Propensity to Invest at Home

We now document more systematically how the governance structures of sovereign wealth funds are associated with differences in their investment strategies. In particular, we investigate whether the involvement of external managers or that of politicians in investment management is correlated with outcomes. We analyze investment strategies of sovereign wealth funds looking at their propensity to invest at home, the industry-country price/earnings levels at the time of the investments, the subsequent changes in the price/earnings ratios, and the size of the acquisition stakes of their investments.

One of the important governance-related problems in the investment policies of sovereign wealth funds might be that their money is used to bail out underperforming firms or industries. To analyze how funds vary in their allocation of investments between the home nation and outside, we estimate a probit model. The dependent variable is a home investment dummy, which equals one if the target investment is made within the home nation of the sovereign wealth fund and zero otherwise. In Table 3, we regress the home dummy on indicator variables for the presence of political leaders in the management of the fund and the reliance on external managers. (We cluster the standard errors at the level of the country where the sovereign wealth fund is based.) The displayed coefficients are marginal effects. In the specifications where year dummies are added, the sample only includes transactions from 1991 onward.[5]

In the base specification, we employ no controls. In the second and subsequent regressions, we control for the geographic location of the sovereign wealth fund (Asian, Middle Eastern, and Western). The results in the first column show that in cases where political leaders are involved with the management of the funds, domestic investments are more common, while involvement of external managers is associated with fewer domestic investments. The magnitude of the effects is large: the coefficient on the politician dummy reflects a 41.3 percent increase in the likelihood of investing at home when politicians are involved. In comparison, the coefficient on the external manager dummy is equivalent to a 27.3 percent lower share of domestic investments when external managers are employed.

[5] Regressions are weighted by winsorized transaction sizes (expressed in 2000 US dollars). Since we only have sizes for 67 percent of our transactions, we impute missing weights by constructing the fitted values from a regression of deal sizes on fixed effects for the investment year, target industry, target region, and fund. After adding imputed observations, we winsorize the deal size variable at the 5 percent and 95 percent level to reduce the effect of extreme observations.

Table 3
Sovereign Wealth Fund Behavior

	Dependent variable			
	Home Dummy		*P/E Levels*	*P/E Change*
	(1)	*(2)*	*(3)*	*(4)*
Politicians	0.413*** (0.107)		4.153*** (1.300)	–0.042*** (0.004)
External Managers	–0.273*** (0.058)		–4.562*** (1.307)	0.023*** (0.007)
Home P/E		0.006** (0.003)		
Outside P/E		–0.005* (0.003)		
Home Investment			–5.607* (2.625)	–0.017 (0.013)
Year dummies	No	No	Yes	Yes
Sovereign wealth fund region dummies	No	Yes	Yes	Yes
Target region dummies	No	No	Yes	Yes
R^2	0.128	0.097	0.167	0.128
N	2,618	2,618	2,533	2,524

Notes: The sample consists of 2,662 investments by 29 sovereign wealth funds. It excludes transactions that were withdrawn or rejected. The dependent variable *Home Dummy* is a dummy denoting whether the investment target was based in the same nation as the sovereign wealth fund; the dependent variable *P/E Levels* is the weighted (by firm value) average of the price/earnings ratios of publicly traded firms in the industry, country, and year of the transaction; the dependent variable *P/E Change* denotes a one year percentage change in the value of *P/E Levels* from the year of the transaction. *External Managers* is equal to 1 if external managers are involved in the management of the fund, zero otherwise. The *Politicians* variable is a dummy equal to 1 if politicians are involved in the management of the fund. The *Home P/E* variable is the country-level P/E ratio of home country. The *Outside P/E* variable is equal to the target country P/E ratio if investment is not in the sovereign wealth fund's home nation. If investment is at home, *Outside P/E* is equal to the average (weighted by the total transaction sizes of the sovereign wealth fund deals in the sample) P/E ratios of all other countries in which investments were made by sovereign wealth funds. *Home Investment* is a dummy variable which equals one if the target is based in the same country as the sovereign wealth fund. We include dummy variables for different regions, set equal to 1 when a fund or target is based in Asia or the Middle East. The estimation method in the first two regressions is a weighted probit model and in the second pair is weighted ordinary least squares, using in both cases as weights winsorized transaction sizes (converted to 2000 US dollars). The displayed coefficients are marginal effects. Standard errors are clustered at the sovereign wealth fund country. When year dummies are added, the sample only includes transactions from 1991 onward.
***,**, and * indicate levels of significance of 1, 5, and 10 percent, respectively.

In column 2 of Table 3, we repeat the regression from column 1, but add measures of the price/earnings level of the sovereign wealth fund's nation (*Home P/E Level*) and of the price/earnings level of the country in which the fund invests (*Outside P/E Level*). The results show that there is a significant correlation between higher price/earnings levels in the other countries and a lower propensity to invest at home. An increase in one standard deviation of *Outside P/E* decreases the likelihood of investing at home by 3.11 percent, while that of *Home P/E* increases it by 4.5 percent.

The cross-sectional results suggest that sovereign wealth funds invest more at home if their local equity markets have relatively high price-to-earnings levels and similarly they are less likely to invest at home if foreign markets are valued highly. One possible explanation for this pattern might be that sovereign funds try "correctly" to invest in markets that have high option values, high price-to-value levels.[6] But an alternative interpretation would be that they choose investments that are overvalued. Given the return dynamics which we present in the next section, it rather appears that the results are more consistent with sovereign wealth funds engaging in "trend chasing," that is, they gravitate to markets where equity values have already been bid up highly.

Valuation Levels

In a second step, we examine whether there are significant differences in the market timing of the transactions undertaken by sovereign wealth funds that have involvement of politicians compared to those run by professional managers. In the third column of Table 3, we rerun the same regression as before but the dependent variable is the weighted average (by firm value) of the price-to-earnings ratios of publicly traded firms in the industry, country, and year of the transaction. We find that having politicians involved is strongly associated with investments in higher-priced sectors (a premium of three-to-four times earnings), while external managers are associated with investments in lower-valued sectors.

Investment Performance

To understand the propensity of sovereign wealth funds with political involvement to invest in industries with high valuations as measured by price/earnings ratios, we now look at the later performance of these industries. On the one hand, investments in high price-to-earnings industries could be a sign that politicians favor industries with attractive investment opportunities as argued, for example, in Gordon (1959) and Bekaert, Harvey, Lundblad, and Siegel (2007). On the other hand, investments in industries with high price-to-earnings ratios might suggest that sovereign wealth funds engage in trend chasing and buy into inflated valuations, as discussed in Lakonishok, Shleifer, and Vishny (1994). If the first interpretation is true, we should see that sovereign wealth funds outperform in home investments, while the opposite would hold under the second explanation.

The regression in the fourth column in Table 3 is structured to be parallel to the first three columns, but now the dependent variable is the percentage change in the average price-to-earnings ratio of firms in that country and industry in the year following the investment. By looking at the subsequent performance

[6] High price to value means that the market values the company much higher than its assets in place. The only reason that is rational is if the market expects this firm to have great returns in the future. This is exactly the option value that is priced into the firm's stock.

of the sector, we can address some of the interpretative challenges highlighted above. As in the previous section, we use a transaction size-weighted ordinary least squares specification.

We see here that sovereign wealth funds where political leaders play a role select sectors with significant drops in price-to-earnings ratios going forward (−4.2 percent). This is in contrast to the case when external managers are involved, where price-to-earnings values increase in the year following the investment (+2.3 percent).[7] The analysis suggests that sovereign wealth funds with politician involvement do not select high price-to-earnings sectors because they have better private information about investment opportunities (as the finding of home bias in investments might initially suggest). Rather, it seems to reflect a willingness to trend chase and overpay for investments. The analysis suggests, at least weakly, that these effects are stronger when it comes to domestic investments.

In unreported regressions, we verify that these results also hold if we use data at the deal level for the subset of firms that were publicly traded at the time of investment. We obtain the information from Datastream for all target companies that were publicly traded and calculate the cumulative abnormal returns relative to the local market benchmark in the six months after the transaction. We find once again that in the basic specifications, politician-influenced sovereign wealth funds are associated with lower returns. These transactions significantly underperform, generating 16 percent lower returns in the six months after the investments. The home investment dummy now has a significantly negative coefficient, suggesting underperformance among domestic investments. While the sample of publicly traded transactions is considerably smaller, the similarity to the results reported in Table 3 is reassuring.[8]

Overall, our results lend support to the hypothesis that funds exposed to political influences show major deviations from long-run return maximization. Sovereign wealth funds with politician involvement are more likely to invest domestically, while those sovereign wealth funds where external managers play an important role are more likely to invest internationally. Politically influenced sovereign wealth funds also concentrate their funds in sectors that both have high price-to-earnings levels

[7] When interactions with home investments are added in unreported regressions, the interaction term between politician influence and home investments is negative and significant, reflecting a decline of 6.8 percent in returns when investing at home.

[8] In an unreported regression, we also consider a benchmark that matches to the type of security. We use as the dependent variable the percentage change in the weighted (by firm value) average EBITDA/assets ratio of all publicly traded firms if the target is public, or if the target is private, all privately held firms in the corresponding three-digit SIC industry, country, and year of the target in the transaction. We determine the ratios for the corresponding firms from the 2009 edition of the Orbis database from Bureau van Dijk, which includes financial information about private firms for many nations. The important advantage of Orbis is that it includes data on both public and private firms (in fact, most of the firms in this database are private). Unfortunately, in many cases, the information is quite scanty, so we can only obtain a ratio for the corresponding industry, country, and year for 796 firms—far fewer than for the price-to-earnings ratio, where we have a benchmark for 2,553 firms. The results are quite weak. In the basic regressions, the *Politicians* variable retains a negative coefficient and the *External Managers* a positive one, but neither are statistically significant.

and then experience a drop in these levels, especially in their domestic investments, patterns that do not hold in funds that rely on external managers. Political pressures seem to force these sovereign wealth funds to use their funds to support underperforming local industries rather than build a savings buffer for the long run. The performance gap between domestic and international investments when more political appointees are on the board also supports the interpretation that politically connected managers are not purely making poor decisions when investing but that there is a strategic component.

Stated Investment Objective

Some sovereign wealth funds profess a desire to focus on more short-term strategic objectives, such as the acquisition of useful companies or domestic industrial development. Others aim more at the long-term return goals that are akin to those of a university endowment.

In Table 4, we repeat the analyses of Table 3, but look specifically at the role that investment objectives play. Recall that we define funds' objectives to be "strategic" if stated goals include management of government physical assets, acquisition of strategic assets, or domestic development. We consider the rest of the objectives as "nonstrategic" (investment of oil/commodity revenues, currency reserve management, or pension funding).[9] We employ the same sample, number of observations, and dependent variables as those reported in Table 3. The independent variables change slightly across the four regressions and we add the independent variable *Strategic Objectives* and as well as the interaction of *Strategic Objectives* with *Politicians* (*Politicians* × *Strategic Objectives*). In the regression analyses of the decision to invest at home, we find that when political leaders are involved, those funds that have strategic objectives show a significantly higher probability of investing at home. Meanwhile, the coefficients on *Strategic Objectives* or *Politicians* as separate variables are either insignificant or of reduced statistical significance. In the other two regressions, the interaction between the strategic objective measure and politicians are insignificant. As before political leader–influenced investments are associated with high prices and subsequent underperformance regardless of their stated strategic objectives.

Robustness

One could be worried that our results might be driven either by some of the smaller deals or the valuation trends in the years immediately before the financial crisis. Alternatively, one might worry that there is a sample selection bias, which is doubtless a greater problem among the smaller transactions. To verify that our

[9] Most funds include multiple goals, which typically fall under the same broad category. In 220 investments, fund goals included both strategic and nonstrategic objectives. We included all these transactions in the nonstrategic group, and verified that results are similar when these are included in the strategic group instead.

Table 4
Investment Objectives

	Dependent variable			
	Home Dummy		*P/E Levels*	*P/E Change*
	(1)	*(2)*	*(3)*	*(4)*
Politicians	0.192	0.199*	3.630***	–0.038***
	(0.128)	(0.103)	(0.900)	(0.012)
Strategic Objectives	–0.086	–0.069	2.261*	–0.015
	(0.077)	(0.154)	(1.249)	(0.009)
Politicians × Strategic Objectives	0.477***	0.454*	–0.993	0.014
	(0.176)	(0.245)	(2.504)	(0.013)
External Managers	Yes	Yes	Yes	Yes
Home P/E	No	Yes	No	No
Outside P/E	No	Yes	No	No
Home Investment	No	No	Yes	Yes
Year dummies	No	No	Yes	Yes
Sovereign wealth fund region dummies	No	Yes	Yes	Yes
Target region dummies	No	No	Yes	Yes
R^2	0.142	0.014	0.169	0.128
N	2,618	2,618	2,533	2,524

Notes: The four regressions are very similar to those reported in Table 3. The main changes are the addition of *Strategic Objectives* and the interaction of *Strategic Objectives* with *Politicians* (*Politicians* × *Strategic Objectives*). Robust standard errors, allowing for data clustering by the countries in which the sovereign wealth funds are based, are shown in parenthesis.
***,**, and * indicate levels of significance of 1, 5, and 10 percent levels, respectively.

results are robust to these concerns, we undertake a number of additional tests that examine different subsets of the data.

We repeated all the regressions presented in this paper using two subsamples, one which includes the largest 75 percent of the deals, and the other with the largest 50 percent of the deals. Even after removing the smaller half of the transactions, the remaining transactions maintain the same distribution across the groups. And in both subsamples, the results remained similar to the ones reported in the paper. We also run the regressions without winsorizing the data (without trimming the outliers). We repeat our analysis excluding either the last two years or the last year of the sample and find that the results remain unchanged. Finally, we conduct simple weighted mean tests and run unweighted regressions to explore the robustness of the results. The results exhibit similar patterns to the ones described in the analyses above.

In short, our results lend support to the hypothesis that funds, which are exposed to political influences, show major distortions from long-run return maximization. Sovereign wealth funds with politician involvement are more likely to invest domestically, while those funds where external managers play an important role are more likely to invest internationally. Politically influenced sovereign wealth funds also concentrate their funds in sectors that both have high price/earnings levels and then experience a drop in these levels, especially in their domestic investments, while these patterns do not hold in funds that rely on external managers. Funds that have stated strategic goals are more likely to invest at home but only if politicians are involved.

Other Challenges: Transparency and Managing Growth

Although sovereign wealth funds have existed for six decades, they are facing increased political scrutiny in many nations both because of their accelerating growth and because of highly public transactions that drew them into the global spotlight, such as the $7.5 billion investment in Citigroup in November 2007 by the Abu Dhabi Investment Authority. The controversies surrounding investments by sovereign funds are not new—witness the 1987 row over the Kuwait Investment Office's purchase of a 20 percent stake in British Petroleum—yet the intensity of scrutiny in recent years has been unprecedented and seems unlikely to subside. In part, these concerns can be attributed to intense anxiety in many established economies about globalization and the changing global balance of power. But these fears can also be understood as a reaction to the intense secrecy that surrounds some of the activities of sovereign wealth funds. Greater visibility—publicizing the size of the pools, investment strategies, and particular investments—could help dispel at least part of the worries over sovereign funds. While the International Working Groups of Sovereign Wealth Fund's 2009 Generally Accepted Principles and Practices (GAPP) ("Santiago Principles") spoke of the desirability of transparency along a variety of dimensions, actual compliance with these principles has been quite limited.

The reluctance of sovereign wealth funds regarding disclosure may have two roots. First, too much disclosure can have real costs, since it can lead to increased imitation by other investors. The experience of American university endowments offers a useful object lesson here. In the past, a substantial lag typically occurred between the time a few university endowments first began investing in an asset class and the time other institutions followed. For instance, many of the Ivy League schools began investing in venture capital in the early 1970s, but most corporate and public pensions did not follow until the 1980s and 1990s, respectively. More recently, such lags are much shorter. Within a couple of years of Harvard's initiating a program to invest in forestland, for instance, many other institutions adopted similar initiatives. In general, an investment by a prominent institution can trigger a rush of capital seeking to gain access to the same type of investment, thus, making it much harder for the investor to continue what might otherwise have been a successful strategy (for a further discussion of these issues, see World Economic Forum 2011).

Furthermore, even an aggressive policy of encouraging transparency will not solve all of the challenges that sovereign wealth funds face. Investment decisions that would seem unremarkable when made by an individual or institutional investor can become political hot potatoes when undertaken by a sovereign fund. Consider, for instance, the experience of Norway's Government Pension Fund when the fund trimmed its portfolio of firms using child labor and thus sold $400 million of Wal-Mart stock (based on reports that Wal-Mart was selling goods that had been produced in other countries for the firm using child labor). This decision triggered a diplomatic row with the American ambassador, who accused Norway of passing "essentially a national judgment on the ethics of the [company]" (as reported in Landler 2007). The fund pointed out that when it had shared with Wal-Mart its draft report presenting evidence about the company's labor practices, Wal-Mart ignored it. (For an overview of the dispute, see Pozen 2007). Similarly, when Norway's Government Pension Fund, along with many hedge funds, sold short the shares of Icelandic banks in 2006, it triggered a major diplomatic row with that nation (as reported in *The Economist*, 2008).

Another major challenge that sovereign wealth funds must address is how to ensure attractive investment returns as they grow. Strategies that work for a modest-sized institution may be difficult to scale up into a larger organization. For instance, it may be possible for a billion-dollar endowment to generate attractive returns from investments of $10 million apiece in private equity funds or in developing markets. If a sovereign fund with 100 times the capital were to pursue a similar strategy, it would probably 1) be unable to find enough attractive investments to have a return that significantly boosts that of the overall fund; or 2) find that purchases of larger blocks of stock affect the market price to the extent that the strategy is far less profitable. For similar reasons, many university endowments have struggled to maintain their success as they have become larger. Thus, for the larger sovereign funds, generating attractive returns is by no means simple.

Sovereign wealth funds have adopted a range of approaches to deal with this issue. At one extreme is the Norwegian Sovereign Wealth Fund, which allocates almost no capital into alternative assets (private equity, hedge funds, or real estate investment) or illiquid markets. Instead the fund mainly invests into liquid and very transparent investments—like public debt and equity markets —outside of the home country.

This strategy minimizes the requirements on specialized knowledge of the investment staff and might allow the fund to maintain returns as it grows in size. At the other extreme are sovereign wealth funds like Temasek from Singapore that have heavily invested in private deals either via allocations to private equity or through direct investments in companies, often in other Asian economies. This latter strategy places much higher requirements on the investment office. Particularly in asset classes such as private equity and real estate, where funds and strategies often do not scale well, such strategies might ultimately be more difficult to continue as a fund grows in size.

Is there a way to overcome the diseconomies of scale that can drag down the returns of large institutional investors? One approach that the Government Investment Corporation of Singapore has tried has been to build an organizational structure in which a number of subsidiaries are managed separately. In this way, managers can make smaller investments. Such separate funds can also serve as "laboratories": successful approaches can be emulated by the other funds, while mistakes can be less costly since they affect only one subsidiary. It is an open research question whether such approaches allow sovereign wealth funds to continue to invest successfully as they grow.

■ *We thank Harvard Business School's Division of Research for financial assistance and various seminar audiences and Timothy Taylor for helpful comments. Parts of this essay are adapted from Lerner (2009). Chris Allen and Jacek Rycko provided excellent research assistance.*

References

Ang, Andrew. 2010. "The Four Benchmarks of Sovereign Wealth Funds." http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1680485.

Atkinson, Anthony B., and Joseph E. Stiglitz. 1980. *Lectures on Public Economics.* London: McGraw Hill.

Bekaert, Geert, Campbell R. Harvey, Christian Lundblad, and Stephan Siegel. 2007. "Global Growth Opportunities and Market Integration." *Journal of Finance* 62(3): 1081–1137.

Banerjee, Abhijit V. 1997. "A Theory of Misgovernance." *Quarterly Journal of Economics* 112(4): 1289–1332.

Bortolotti, Bernardo, Veljko Fotak, William L. Megginson, and William Miracky. 2010, "Sovereign Wealth Fund Investment Patterns and Performance." Unpublished paper.

Chhaochharia, Vidhi, and Luc A. Laeven. 2009. "The Investment Allocation of Sovereign Wealth Funds." http://ssrn.com/abstract=1262383.

Dewenter, Katherine L., Xi Han, and Paul H. Malatesta. 2010. "Firm Values and Sovereign Wealth Fund Investments." *Journal of Financial Economics* 98(2): 256–78.

Dyck, I. J. Alexander, and Adair Morse. 2011. "Sovereign Wealth Fund Portfolios." Chicago Booth Research Paper no. 11-15. http://ssrn.com/abstract=1792850.

Economist, The. 2008. "Asset-backed Insecurity." January 17. http://www.economist.com/node/10533428.

Fernandes, Nuno G. 2011. "Sovereign Wealth Funds: Investment Choices and Implications around the World." http://ssrn.com/abstract=1341692.

Fernandez, David G., and Bernhard Eschweiler. 2008. *Sovereign Wealth Funds: A Bottom-Up Primer.* New York: J.P. Morgan Research.

Friedman, Tim. 2008. *Preqin Sovereign Wealth Fund Review.* London: Preqin Limited.

Gjedrem, Svein. 2005. "The Management of Petroleum Wealth." Lecture at the Polytechnic Association, November 8. http://www.bis.org/review/r051116b.pdf.

Gompers, Paul A., and Andrew Metrick. 2001. "Institutional Investors and Equity Prices." *Quarterly Journal of Economics* 116(1): 229–59.

Gordon, Myron J. 1959. "Dividends, Earnings, and Stock Prices." *Review of Economics and Statistics* 41(2, Part 1): 99–105.

Hedge Funds Research Inc. n.d. http://www.hedgefundresearch.com. Accessed July 21, 2012.

Hart, Oliver, Andrei Shleifer, and Robert W. Vishny. 1997. "The Proper Scope of Government: Theory and an Application to Prisons." *Quarterly Journal of Economics* 112(4): 1127–62.

Hochberg, Yael, and Joshua Rauh. 2011. "Local Overweighting and Underperformance: Evidence from Limited Partner Private Equity Investments." NBER Working Paper 17122.

Knill, April M., Bong-Soo Lee, and Nathan Mauck. 2010. "Is Sovereign Wealth Fund Investment Destabilizing?" http://ssrn.com/abstract=1328045.

Kotter, Jason, and Ugur Lel. 2008. "Friends or Foes? The Stock Price Impact of Sovereign Wealth Fund Investments and the Price of Keeping Secrets." International Finance Discussion Paper 940, Federal Reserve Board.

McKinsey Global Institute. 2011. *Mapping Global Capital Markets 2011.* August. http://www.mckinsey.com/insights/mgi/research/financial_markets/mapping_global_capital_markets_2011.

Lakonishok, Josef, Andrei Shleifer, and Robert W. Vishny. 1994. "Contrarian Investment, Extrapolation, and Risk." *Journal of Finance* 49(5): 1541–78.

Landes, David S. 1998. *The Wealth and Poverty of Nations: Why Some Are So Rich and Some So Poor.* New York: Norton.

Landler, Mark. 2007. "Norway Backs Its Ethics With Its Cash." *New York Times*, May 4. http://query.nytimes.com/gst/fullpage.html?res=9E01E1DB113EF937A35756C0A9619C8B63&sec=&spon.

Lerner, Josh. 2009. *Boulevard of Broken Dreams: Why Public Efforts to Boost Entrepreneurship and Venture Capital Have Failed—And What to Do about It.* Princeton: Princeton University Press.

Lerner, Josh, Antoinette Schoar, and Wan Wongsunwai. 2007. "Smart Institutions, Foolish Choices? The Limited Partner Performance Puzzle." *Journal of Finance* 62(2): 731–64.

Pope, Kyle. 1995. "Uneasy Boom: Norway's Oil Bonanza Stirs Fears of a Future When Wells Run Dry—As Output Climbs, Many Say Money is Being Wasted and a Slump Lies Ahead—A Town's Varying Fortunes." *Wall Street Journal*, October 3.

Pozen, Robert C. 2007. "Norway Sells Wal-Mart." Harvard Business School Case No. 9-308-019.

Shleifer, Andrei, and Robert Vishny. 1994. "Politicians and Firms." *Quarterly Journal of Economics* 109(4): 995–1025.

Stiglitz, Joseph E. 1993. "The Role of the State in Financial Markets." In *Proceedings of the World Bank Annual Conference on Economic Development.* Washington, International Bank for Reconstruction and Development/World Bank, pp. 19–56.

Stromberg, Per. 2008. "The New Demography of Private Equity." Unpublished working paper, Swedish Institute for Financial Research.

World Economic Forum. 2011. *The Future of Long-Term Investment.* Geneva: World Economic Forum. http://www3.weforum.org/docs/WEF_FutureLongTermInvesting_Report_2011.pdf.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 239–246

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., Saint Paul, Minnesota, 55105.

Smorgasbord

Nicholas Bloom, Erik Brynjolfsson, Lucia Foster, Ron Jarmin, Itay Saporta-Eksten, and John Van Reenen, offer a first glimpse of results from a recent Census Bureau study of "Management in America." "[T]here is enormous dispersion of management practices across America: 18% of establishments adopt at least 75% of structured management practices for performance monitoring, targets and incentives; while 27% of establishments adopt less than 50% of these practices." "[M]ore structured management practices are tightly linked to better performance: establishments adopting more structured practices for performance monitoring, target setting and incentives enjoy greater productivity and profitability, higher rates of innovation and faster employment growth." US Census Bureau, Center

■ *Timothy Taylor is Managing Editor,* Journal of Economic Perspectives, *based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

http://dx.doi.org/10.1257/jep.27.2.239. doi=10.1257/jep.27.1.239

for Economic Studies, CES 13-01, January 2013. At ftp://ftp2.census.gov/ces/wp/2013/CES-WP-13-01.pdf. In the Winter 2010 issue of this journal, Bloom and Van Reenen offered some international evidence based on a similar methodology in "Why Do Management Practices Differ across Firms and Countries?"

Pankaj Ghemawat and Steven A. Altman have authored the *DHL Global Connectedness Index 2012*. "Furthermore, while 20% (or even 30%) of goods and services being traded across borders is far more than the same ratio mere decades ago, it is still far short of the 90% or more that one would expect if borders and distance did not matter at all. . . . Borders and distance still matter a great deal, implying that even the most connected countries have substantial headroom available to participate more in international trade." "The global connectedness patterns traced in this report also highlight how distance, far from being dead, continues to depress connectedness of all types. While the distance between a randomly selected pair of countries is roughly 8,500 km, the average distance traversed by merchandise trade, foreign direct investment flows, telephone calls, and human migration all cluster in the range from 3900 km to 4750 km. This accords with the finding that most international flows take place within rather than between continental regions." "While the growth of international internet bandwidth implies that we can just as easily read foreign news websites as domestic ones, people still overwhelmingly get their news from domestic sources when they go online: news page views from foreign news sites constitute 1% of the total in Germany, 3% in France, 5% in the United Kingdom and 6% in the United States (and are in single digits everywhere else sampled—as low as 0.1% in China). Furthermore, news coverage by domestic sources itself tends to be very domestic." At http://www.dhl.com/content/dam/flash/g0/gci_2012/download/dhl_gci_2012_complete_study.pdf.

Elroy Dimson, Paul Marsh, and Mike Staunton lay out their predictions of "The low-return world" in the first section of the *Credit Suisse Global Investment Returns Yearbook 2013*. Until a decade ago, it was widely believed that the annualized equity premium relative to bills was over 6%. This was strongly influenced by the Ibbotson Associates Yearbook. In early 2000, this showed a historical US equity premium of 6¼% for the period 1926–99. Ibbotson's US statistics appeared in numerous textbooks and were applied worldwide to the future as well as the past. It is now clear that this figure is too high as an estimate of the prospective equity premium. First, it overstates the long-run premium for the USA. From 1900–2012, the premium was a percentage point lower at 5.3%, as the early years of both the 20th and 21st centuries were relatively disappointing for US equities. Second, by focusing on the USA—the world's most successful economy during the 20th century—even the 5.3% figure is likely to be an upwardly biased estimate of the experience of equity investors worldwide. . . . To assume that savers can confidently expect large wealth increases from investing over the long term in the stock market—in essence, that the investment conditions of the 1990s will return—is delusional. . . . While a low-return world imposes stresses on investors and savers in an over-leveraged world recovering from a deep financial crisis, it provides essential relief for borrowers. The danger here is that if this continues too long, it creates "zombies"—businesses kept alive by low interest rates and a reluctance to write off bad loans. This can suppress creative destruction and rebuilding, and can prolong the downturn." At http://www.investmenteurope.net/digital_assets/6305/2013_yearbook_final_web.pdf.

Claudio Borio asks "The Financial Cycle and Macroeconomics: What Have We Learnt?" "The financial crisis that engulfed mature economies in the late 2000s has prompted much soul searching. Economists are now trying hard to incorporate financial factors into standard macroeconomic models. However, the prevailing, in fact almost exclusive, strategy is a conservative one. It is to graft additional so-called financial "frictions" on otherwise fully well behaved equilibrium macroeconomic models . . . The main thesis is that macroeconomics without the financial cycle is like Hamlet without the Prince. In the environment that has prevailed for at least three decades now, just as in the one that prevailed in the pre-WW2 years, it is simply not possible to understand business fluctuations and their policy challenges without understanding the financial cycle." Bank of International Setttlements, Working Paper #395, December 2012. At http://www.bis.org/publ/work395.pdf.

Daniel W. Sacks, Betsey Stevenson, and Justin Wolfers discuss "The New Stylized Facts about Income and Subjective Well-Being." From the abstract: "In recent decades economists have turned their attention to data that asks people how happy or satisfied they are with their lives. Much of the early research concluded that the role of income in determining well-being was limited, and that only income relative to others was related to well-being. . . . Our research suggests that absolute income plays a major role in determining well-being and that national comparisons offer little evidence to support theories of relative income. We find that well-being rises with income, whether we compare people in a single country and year, whether we look across countries, or whether we look at economic growth for a given country. Through these comparisons we show that richer people report higher well-being than poorer people; that people in richer countries, on average, experience greater well-being than people in poorer countries; and that economic growth and growth in well-being are clearly related. Moreover, the data show no evidence for a satiation point above which income and well-being are no longer related." The paper is available as IZA Discussion Paper No. 7105, released in December. At http://ftp.iza.org/dp7105.pdf.

Nicholas Lardy and Nicholas Borst offer "A Blueprint for Rebalancing the Chinese Economy." "For the past several years China's top leadership has repeatedly described the country's current economic model as 'uncoordinated, unsteady, imbalanced, and unsustainable.' This language is in sharp contrast to what has been a decade of apparent success: high-speed economic growth and emergence into the ranks of middle-income countries. What accounts for this discontinuity between rhetoric and record? Chinese policymakers have correctly assessed that the country's economic growth over the past decade has been based on superelevated levels of investment and systematic suppression of private consumption. The resulting capital-intensive growth model has not generated adequate gains in consumption and employment and instead has built up significant distortions in the economy." "The imbalances in the Chinese economy were created by distortions to three of the most fundamental prices in the economy: interest rate, exchange rate, and price of energy. An underdeveloped social safety net and high levels of income inequality have exacerbated these imbalances. Rebalancing policies should focus on allowing these key price to be more market-determined, and the government should increase social transfers and work towards a more equitable distribution of income." Peterson Institute for International Economics, Policy Brief PB 13-02. February 2013. At http://www.piie.com/publications/pb/pb13-2.pdf. This article can serve as an accompaniment to the five-paper symposium on various aspects of China's economy in the Fall 2012 issue of this journal.

The March 2013 issue of *Finance & Development* includes seven articles looking ahead at economic and social issues in the Middle East region. As one example, Vali Nasr writes: "The Arab population today numbers 400 million, which will double to 800 million by 2050. Population growth makes aggressive economic growth an urgent imperative. Even to tread water and maintain current living standards, the Arab economies would need to grow at 'tiger-economy' rates of 9 to 10 percent for a decade or more. That is a daunting task, one the public sector cannot accomplish alone. Growth must come from the private sector, and that requires reform of the economy: removing regulations, relaxing government control, promoting trade, and bolstering the rule of law. . . . Middle-class entrepreneurs represent the best hope for betterment of their countries—and the most potent weapon against extremism and for democracy. Until now the Arab world's tiny middle class has relied on state salaries and entitlements, with few ties to free markets. The growth of local entrepreneurship on the back of burgeoning capitalism—and integration with the world economy—could help change that. " At http://www.imf.org/external/pubs/ft/fandd/2013/03/index.htm.

In "Wayward Sons: The Emerging Gender Gap in Labor Markets and Education," David Autor and Melanie Wasserman point out: "Over the last three decades, the labor market trajectory of males in the U.S. has turned downward along four dimensions: skills acquisition; employment rates; occupational stature; and real wage levels." "[W]e argue first that sharp declines in the earnings power of non-college males combined with gains in the economic self-sufficiency of women—rising educational attainment, a falling gender gap, and greater female control over fertility choices—have reduced the economic value of marriage for women. This has catalyzed a sharp decline in the marriage rates of non-college U.S. adults—both in absolute terms and relative to college-educated adults—a steep rise in the fraction of U.S. children born out of wedlock, and a commensurate growth in the fraction of children reared in households characterized by absent fathers. The second part of the hypothesis posits that the increased prevalence of single-headed households and the diminished child-rearing role played by stable male parents may serve to reinforce the emerging gender gaps in education and labor force participation by negatively affecting male children in particular." Third Way. March 2013. At http://content.thirdway.org/publications/662/Third_Way_Report_-_NEXT_Wayward_Sons-The_Emerging_Gender_Gap_in_Labor_Markets_and_Education.pdf. (Full disclosure: David Autor is Editor of the *Journal of Economic Perspectives.*)

The Kaiser Family Foundation has pulled together a list of about 130 possible *Policy Options to Sustain Medicare for the Future.* The five listed options that would have the largest fiscal effect are: 1) Increase Medicare payroll tax by 1 percentage point for all workers (worth $651 billion over 10 years); 2) Increasing premiums for Part B and Part D: for example, raise Part B premiums by 2% per year until they cover 35% of total Part B expenses ($231 billion over 10 years); 3) Set Federal contributions per beneficiary at the average plan bid in a given area, including traditional Medicare as a plan, weighted by enrollment ($161 billion over 10 years); 4) Require manufacturers to pay a minimum rebate on drugs covered under Medicare Part D for beneficiaries receiving low-income subsidies ($137 billion over 10 years); and 5) Raise the age of Medicare eligibility from 65 to 67 ($113 billion over 10 years). January 2013. At http://www.kff.org/medicare/upload/8402.pdf.

Reinhilde Veugelers investigates "The World Innovation Landscape: Asia Rising?" "The United States is by far the biggest spender on R&D ($402 billion in 2009), accounting for about 32 percent of the global total. But the US share (not volume) is in decline, having stood at 38 percent in 1999. The country making the most spectacular inroad is China, which by 2009 was the second biggest spender ($154 billion), accounting for about 12 percent of the global total. Its R&D expenditure is now similar to that of Germany, France and Italy combined. Japan has been pushed into third place, at 11 percent ($138 billion). . . . Although other countries, such as South Korea, are also catching up, the Chinese emergence in science is uniquely rapid, particularly in engineering, chemistry and physics. . . . It would be wrong to discount the Chinese innovation potential on the basis of current performance. China clearly has the ambition to become a world-leading innovator, creating and capturing high-tech value added, particularly in targeted areas." Bruegel Policy Contribution, Issue 2013-02, February 2013, at http://www.bruegel.org/publications/publication-detail/publication/766-the-world-innovation-landscape-asia-rising/#.UWYLSjeOXKU.

From Federal Reserve Banks

David Luttrell, Harvey Rosenblum, and Jackson Thies have written an essay on "Understanding the Risks Inherent in Shadow Banking: A Primer and Practical Lessons Learned." "Through the 2007–09 financial crisis, the term 'shadow banking' appeared in headlines and descriptions of the contagion in money and capital markets. This paper provides a narrative of the role and inherent risks of the shadow banking system, describing its basic functioning and development, rise to prominence, and precipitous decline. . . . While working to ensure the current reform effort has a chance to end bailouts, eliminate TBTF ["too big to fail"], and promote financial stability, we should remember the lessons learned from Minsky about boom times: *The transition from hedge to speculative to Ponzi financing is a slippery slope of greed, perhaps accompanied by a generous dose of willful blindness—a human tendency to see what we want to see, or are conditioned to see or overlook.* In addition to stronger regulatory standards for bank capital and liquidity, the broader financial system encompassing banks, shadow banks, and capital markets requires greater market discipline and changes to institutional incentives to lift the veil of obfuscation and opacity that leads to mispriced risk. *Currently, the drivers of systemic risk remain largely intact, and shadow banking appears poised to grow considerably, and dangerously, if it does not acquire the necessary market discipline to shape risk-taking activities.*" Federal Reserve Bank of Dallas, Staff Papers No. 18, November 2012 At http://www.dallasfed.org/assets/documents/research/staff/staff1203.pdf.

Juan M. Sánchez and Emircan Yurdagul address the question "Why Are Corporations Holding So Much Cash?" "In 2011, cash holdings [of all firms] amounted to nearly $5 trillion, more than for any other year in the series, which starts in 1980. . . . There are two main reasons why firms find it beneficial to hold cash: precautionary motive and repatriation taxes. The first motive is very simple: Firms hold cash and equivalent liquid assets because they provide the flexibility that firms need in their transactions. Two factors interact directly with this proposed explanation: uncertainty and credit constraints. . . . The second motive is present for multinational firms and is due to repatriation taxes. . . . In particular, taxes due to the U.S. government from corporations operating abroad are determined by the difference between the taxes already paid abroad and the taxes that U.S. tax rates would imply. Importantly, such taxation only takes place when earnings are repatriated. Therefore, firms may have incentives to keep foreign earnings abroad. As a consequence, in times of limited foreign investment opportunities and high profitability, these funds are likely to be held abroad in the form of cash." *Regional Economist,* Federal Reserve Bank of St. Louis, January 2013, pp. 5–8. At http://www.stlouisfed.org/publications/pub_assets/pdf/re/2013/a/cash.pdf.

Discussion Starters

John B. Shoven discusses "Efficient Retirement Design." "[T]he vast majority of people start their Social Security almost immediately upon reaching 62 or retiring. They start collecting Social Security as soon as possible. . . . Well, workers could separate the decision to retire from the decision to commence Social Security. They could delay collecting Social Security and this might make sense once they learn that monthly Social Security benefits are higher the later they are started. In fact, monthly benefits go up for each month of delay from age 62 to age 70. Defined contribution assets could be used to finance the deferral of Social Security instead of financing a supplement to Social Security. It turns out that deferral is a better strategy for most people. . . . Our conclusion is that most people should be using at least a substantial part of their retirement savings to defer Social Security rather than supplement it. Almost no one is getting it right." Shoven presents examples showing that retiring, living on savings for a time, and postponing when you begin drawing on Social Security can have lifetime benefits in excess of $200,000 in plausible cases. Stanford Institute for Economic Policy Research Policy Brief, March 2013. At http://siepr.stanford.edu/publicationsprofile/2549.

Bruce Everett makes a case for "Back to Basics on Energy Policy." "In June 1973, President Richard Nixon addressed the emerging energy crisis, saying that 'the answer to our long-term needs lies in developing new forms of energy.' He asked Congress for a five-year, $10 billion budget to 'ensure the development of technologies vital to meeting our future energy needs.' With this speech, the federal government set out to engineer a fundamental transformation of our energy supply. All seven subsequent presidents have endorsed Nixon's goal, and during the past 40 years, the federal government has spent about $150 billion (in 2012 dollars) on energy R&D, offered $35 billion in loan guarantees, and imposed numerous expensive energy mandates in an effort to develop new energy sources. During this time, many talented and dedicated people have worked hard, done some excellent science, and learned a great deal. Yet federal energy technology policy has failed to reshape the U.S. energy market in any meaningful way." *Issues in Science and Technology,* Fall 2012. At http://www.issues.org/29.1/bruce.html.

John Schmitt reviews of the controversies over the minimum wage in "Why Does the Minimum Wage Have No Discernible Effect on Employment?" "This report examines the most recent wave of this research—roughly since 2000—to determine the best current estimates of the impact of increases in the minimum wage on the employment prospects of low-wage workers. The weight of that evidence points to little or no employment response to modest increases in the minimum wage. The report reviews evidence on eleven possible adjustments to minimum-wage increases that may help to explain why the measured employment effects are so consistently small. The strongest evidence suggests that the most important channels of adjustment are: reductions in labor turnover; improvements in organizational efficiency; reductions in wages of higher earners ("wage compression"); and small price increases. Given the relatively small cost to employers of modest increases in the minimum wage, these adjustment mechanisms appear to be more than sufficient to avoid employment losses, even for employers with a large share of low-wage workers." Center for Economic and Policy Research. February 2013. http://www.cepr.net/documents/publications/min-wage-2013-02.pdf.

The "Third Grade Follow-up to the Head Start Impact Study: Final Report" has been published by the Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services. Basically, the report shows that Head Start provides short-term gains to preschool children, but those gains have faded to essentially nothing by third grade. The findings are summarized in this way: "In summary, there were initial positive impacts from having access to Head Start, but by the end of 3rd grade there were very few impacts found for either cohort in any of the four domains of cognitive, social-emotional, health and parenting practices. The few impacts that were found did not show a clear pattern of favorable or unfavorable impacts for children." December 2012. At http://www.acf.hhs.gov/sites/default/files/opre/head_start_report.pdf. The paper in this issue by Greg J. Duncan and Katherine

Magnuson, "Investing in Preschool Programs," offers a broader perspective on these issues.

Troy R. Hawkins, Bhawna Singh, Guillaume Majeau-Bettez, and Anders Hammer Strømman present a "Comparative Environmental Life Cycle Assessment of Conventional and Electric Vehicles." We find that EVs [electric vehicles] powered by the present European electricity mix offer a 10% to 24% decrease in global warming potential (GWP) relative to conventional diesel or gasoline vehicles assuming lifetimes of 150,000 km. However, EVs exhibit the potential for significant increases in human toxicity, freshwater eco-toxicity, freshwater eutrophication, and metal depletion impacts, largely emanating from the vehicle supply chain. Results are sensitive to assumptions regarding electricity source, use phase energy consumption, vehicle lifetime, and battery replacement schedules. Because production impacts are more significant for EVs than conventional vehicles . . . [a]n assumption of 100,000 km decreases the benefit of EVs to 9% to 14% with respect to gasoline vehicles and results in impacts indistinguishable from those of a diesel vehicle. Improving the environmental profile of EVs requires engagement around reducing vehicle production supply chain impacts and promoting clean electricity sources in decision making regarding electricity infrastructure." *Journal of Industrial Ecology,* February 2013, vol. 17, no. 1, pp. 53–64. At http://onlinelibrary.wiley.com/doi/10.1111/j.1530-9290.2012.00532.x/pdf.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 247–252

Correspondence

Are Cognitive Functions Localizable?

The Fall 2011 issue of this journal published a two-paper section on "Neuroeconomics." One paper, by Ernst Fehr and Antonio Rangel, clearly and concisely summarized a small part of the fast-growing literature. The second paper, "It's about Space, It's about Time, Neuroeconomics, and the Brain Sublime," by Marieke van Rooij and Guy Van Orden, is beautifully written and enjoyable to read, but misleading in many critical ways. A number of economists and neuroscientists working at the intersection of the two fields shared our reaction and have signed this letter, as shown below. Some of the paper's descriptions of empirical findings and methods in neuroeconomics are incomplete, badly out of date, or flatly wrong. In studies the authors describe in detail, their skeptical interpretations have often been refuted by published data, old and new, that they overlook.

In the first part of their paper, van Rooij and Van Orden argue that neuroimaging studies are based on a faulty model of how the brain works because brain functions cannot be spatially localized to particular brain regions or networks. Skepticism about spatial localization is not new. Fifteen years ago, Van Orden and Paap (1997) also attacked the strong spatial modularity view. But 15 years is a long time ago in neuroscience. Skepticism about spatially locating brain circuitry might have been cautiously reasonable in 1997, but is clearly a minority view in neuroscience now. During those intervening years, the focus on localization of multiple regions ("circuits") has actually proved to be very useful. For example, in their textbook *Neuroscience*, Purves et al. (2008, p. 22) note:

> When used in combination with functional imaging, well-designed behavioral tasks can facilitate identification of brain networks devoted to specific complex functions, including language skills, mathematical and musical ability, emotional responses, aesthetic judgments, and abstract thinking.

In addition, there are *many* examples of fairly localized functional specialization. Examples include language areas and pathways, somatosensory cortex (for example, perception of touch), face recognition, and visual areas clearly corresponding to distinct steps in visual processing. For higher-order cognition, it is certainly true that a single region of the brain identified by one study is likely to be active in a variety of tasks. However, van Rooij and Van Orden substantially overrepresent the case against functional localization in their Appendix by only listing a highly selective set of studies with the widest range of different interpretations.

Van Rooij and Van Orden did not describe the many methods that are actively used now to check whether localized spatial regions predict common functional activity across tasks. For example, cross-method studies combine fMRI, causal manipulation of activity in targeted regions of the brain (using transcranial stimulation with magnetism or direct current), and behavior of patients with focal lesion damage in certain regions. Results from these types of studies will simply not fit together if there is no functionally reliable localization. Another tool is "activation likelihood estimation," a meta-analytic method of combining results from many different studies. Regions that appear in tasks with a common functional component repeatedly are picked out by this approach, and regions with study-specific activity disappear. This method has been used since Turkeltaub, Eden, Jones, and Zeffiro (2002); the latest version is described in Eickhoff, Bzdok, Laird, Kurth, and Fox (2012). Yarkoni, Poldrack, Nichols, Van Essen, and Wager (2011) offer other methods of computational meta-analysis of neuroimaging data.

Van Rooij and Van Orden offer two specific examples in this part of their discussion: neuroimaging studies of ultimatum bargaining and trust games, and of the ambiguity-risk distinction. Both discussions are misleading.

Sanfey et al. (2003) first used fMRI during ultimatum bargaining. Van Rooij and Van Orden note that:

> If we take the results concerning fairness and generosity from the trust game, together with the previous results concerning unfairness and punishment from the ultimatum game, it would appear that responses to fairness and unfairness are formulated in different parts of the brain.

However, recent studies corroborate many of the conclusions of Sanfey et al. (2003) quite well. For instance, Chang, Smith, Dufwenberg, and Sanfey (2011) find substantial overlap between the neural systems involved in decision-making in the trust game and the regions identified by Sanfey et al. in the ultimatum game. Furthermore, if Sanfey et al. (2003) had wrongly interpreted the roles of the brain regions identified in their study, then disrupting activity in those brain regions using transcranial magnetic stimulation should have no effect on responder behavior in ultimatum games. But it does, as Fehr and Rangel point out (citing Knoch, Pascual-Leone, Meyer, Treyer, and Fehr 2006).

In their example of the neural distinction between responses to risky and ambiguous gambles, van Rooij and Van Orden compare results from Smith, Dickhaut, McCabe, and Pardo (2002) and Hsu, Bhatt, Adolphs, Tranel, and Camerer (2005). But these papers are not comparable, because the tasks are quite distinct. Smith et al. use variants of Ellsberg colored-ball tasks, choosing between an ambiguous gamble and a risky gamble (in some trials). Hsu et al. had subjects choose between certain amounts or individual gambles. Thus, Smith's contrasts *do not* directly measure the difference between computation of ambiguous gamble valuation compared to risky gamble valuation; Hsu et al.'s analyses *do* measure that difference (using a conjunction of activity in all three tasks).

Moreover, the Hsu et al. (2005) finding of stronger lateral activity in the orbitofrontal cortex in response to ambiguity has been corroborated in two ways. In their original paper, they predict that people with brain damage in that area of the brain would be ambiguity-neutral, and they test and confirm this hypothesis in their paper. In addition, their findings were closely corroborated by Levy, Snell, Nelson, Rustichini, and Glimcher (2010, fig. S5). Van Rooij and Van Orden do not mention either of these corroborations.

In addition to their limited discussion of the neuroeconomic literature, van Rooij and Van Orden misrepresent the statistical methodology of neuroimaging. For example, they write that "the spatial approach to studying the brain assumes that the brain can be treated as the sum of its parts . . . this approach underlies what is often called the General Linear Model of the brain." There is no so-called "General Linear Model of the Brain." The phrase "General Linear Model" in the context of neuroimaging refers simply to the statistical technique of multiple linear regression. The term General Linear Model is *never* used by neuroscientists to describe a model of how the whole brain works; for example, the term does not appear in any widely used neuroimaging textbook.

Another misrepresentation involves the discussion of multiple tests of statistical hypotheses. Van Rooij and Van Orden say "any contrast using brain images can be counted on to make 'discoveries'." Neuroimaging researchers are well aware of the potential for false positives. Every widely used neuroimaging textbook contains detailed discussions of the multiple comparisons problem and methods for addressing it (for example, Ashby, 2011, chap. 6; Poldrack, Mumford, and Nichols, 2011, chap. 7; Huettel, Song, and McCarthy, 2009, chap. 12). The best standard of practice in neuroimaging research is to describe and account for the multiple testing problem (Poldrack, Fletcher, Henson, Worsley, Brett, and Nichols 2008).

To summarize what you have just read: the criticism that spatial identification of brain regions and circuits cannot be identified with functions is an old criticism. It has largely been disproven, is no longer widely believed in neuroscience, can be tested with various methods (cross-method and meta-analysis), is largely disproven for one example they discuss (ultimatums) by newer studies, and is disproven for the other example (ambiguity) by data in the same paper they cite.

More broadly, the description of the general methodology of neuroeconomic research by van Rooij and Van Orden is misleading. Even given that articles in this journal are not meant to be comprehensive literature reviews, their sourcing on general neuroeconomics is surprisingly thin, ignoring the leading edited compilation (Glimcher, Camerer, Fehr, and Poldrack 2009), several reviews (for example, Fehr and Camerer 2007; Loewenstein, Rick, and Cohen 2008), recent papers from inside economics (such as Bernheim 2009; Rustichini 2009), and a thoughtful recent book (Glimcher, 2011). In their online Appendix, they strangely, but clearly, misplotted many brain areas.

The more interesting part of the van Rooij and Van Orden paper, which is closely linked to Van Orden's own research, is about why attention to the detailed time course of neural activity is important (as a supplement to spatial understanding). There is no disagreement here, since virtually all neuroscientists who use methods with good time resolution do so to understand fine-grained (subsecond) temporal dynamics, neuroeconomists included. Indeed, the Fehr and Rangel paper in the same issue of this journal discusses drift diffusion models, which predict choices, response times, and other temporal features.

For more than 10 years neuroeconomists have thought about the criticisms and methodological imperfections described by Van Orden and Papp (1997) and have figured out how to respond to those criticisms. It is unfortunate that the paper by van Rooij and Van Orden does not reflect this progress, and therefore misleads readers about the state-of-the-art, rather than educating them.

Colin Camerer, Caltech
Alec Smith, Caltech

Signers (ordered by timing of reply):
Camelia M. Kuhnen, Northwestern
Donald T. Wargo, Temple
Gregory Samanez-Larkin, Vanderbilt
Read Montague, Virginia Tech
Dino J. Levy, NYU
David Smith, Duke
Dar Meshi, Freie Universitaet Berlin
Peter H. Kenning, Zeppelin University, Germany
John Clithero, Caltech
Bernd Weber, Bonn
Todd Hare, Zurich
Scott Huettel, Duke
Camilla Josephson, Linköping University
Mathieu d'Acremont, Caltech
Daria Knoch, Basel
Ian Krajbich, Zurich
Benedetto De Martino, University College London
Peter N. C. Mohr, Freie Universität Berlin
Jan Barton, Emory
Marja-Liisa Halko, Aalto University
Christina F. Chick, Cornell University
Lorena Gianotti, Basel
Hauke R. Heekeren, Freie Universität Berlin

References

Ashby, Gregory F. 2011. *Statistical Analysis of fMRI Data.* Cambridge, MA: MIT Press.

Bernheim, B. Douglas. 2009. "On the Potential of Neuroeconomics: A Critical (But Hopeful) Appraisal." *American Economic Journal: Microeconomics* 1(2): 1–41.

Chang, Luke, Alec Smith, Martin Dufwenberg, and Alan G. Sanfey. 2011. "Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion." *Neuron* 70(3): 560–72.

Eickhoff, Simon B., Danilo Bzdok, Angela R. Laird, Florian Kurth, and Peter T. Fox. 2012. "Activation Likelihood Estimation Meta-Analysis Revisited." *NeuroImage* 59(3): 2349–61.

Fehr, Ernst, and Colin F. Camerer. 2007. "Social Neuroeconomics: The Neural Circuitry of Social Preferences." *Trends in Cognitive Sciences* 11(10): 419–27.

Glimcher, Paul W. 2011. *Foundations of Neuroeconomic Analysis.* Oxford University Press.

Glimcher, Paul W., Colin F. Camerer, Ernst Fehr, and Russell A. Poldrack, eds. 2009. *Neuroeconomics: Decision Making and the Brain.* San Diego: Academic Press.

Gold, Joshua I., and Michael N. Shadlen. 2000. "Representation of a Perceptual Decision in Developing Oculomotor Commands." *Nature* 404(6776): 390–94.

Hsu, Ming, Meghana Bhatt, Ralph Adolphs, Daniel Tranel, and Colin F. Camerer. 2005. "Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making." *Science* 310(5754): 1680–83.

Huettel, Scott A., Allen W. Song, and Gregory McCarthy. 2009. *Functional Magnetic Resonance Imaging*, 2nd edition. Sunderland, MA: Sinauer Associates, Inc.

Knoch, Daria, Alvaro Pascual-Leone, Kaspar Meyer, Valerie Treyer, and Ernst Fehr. 2006. "Diminishing Reciprocal Fairness by Disrupting the Right Prefrontal Cortex." *Science* 314(5800): 829–32.

Levy, Ifat, Jason Snell, Amy J. Nelson, Aldo Rustichini, and Paul W. Glimcher. 2010. "Neural Representation of Subjective Value under Risk and Ambiguity." *Journal of Neurophysiology* 103(2): 1036–47.

Loewenstein, George, Scott Rick, and Jonathan D. Cohen. 2008. "Neuroeconomics." *Annual Review of Psychology* 59(1): 647–72.

Poldrack, Russell A., Paul C. Fletcher, Richard N. Henson, Keith J. Worsley, Matthew Brett, and Thomas E. Nichols. 2008. "Guidelines for Reporting an fMRI Study." *NeuroImage* 40(2): 409–14.

Poldrack, Russell A., Jeanette A. Mumford, and Thomas E. Nichols. 2011. *Handbook of Functional MRI Data Analysis.* Cambridge University Press.

Purves, Dale, George J. Augustine, David Fitzpatrick, William C. Hall, Anthony-Samuel LaMantia, James O. McNamara, and Leonard E. White, eds. 2008. *Neuroscience*, 4th edition. Sunderland, MA: Sinauer Associates, Inc.

Roskies, Adina L. 2010. "Saving Subtraction: A Reply to Van Orden and Paap." *British Journal for the Philosophy of Science* 61(3): 635–65.

Rustichini, Aldo. 2009. "Neuroeconomics: What Have We Found, and What Should We Search For." *Current Opinion in Neurobiology* 19(6): 672–77.

Sanfey, Alan G., James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen. 2003. "The Neural Basis of Economic Decision-Making in the Ultimatum Game." *Science* 300(5626): 1755–58.

Smith, Kip, John Dickhaut, Kevin McCabe and José V. Pardo. 2002. "Neuronal Substrates for Choice under Ambiguity, Risk, Gains, and Losses." *Management Science* 48(6): 711–18.

Turkeltaub, Peter E., Guinevere F. Eden, Karen M. Jones, and Thomas A. Zeffiro. 2002. "Meta-analysis of the Functional Neuroanatomy of Single-Word

Reading: Method and Validation." *Neuroimage* 16(3, Part A): 765–780.

Uttal, William R., 2001. *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain.* Cambridge, MA: MIT Press.

Van Orden, Guy C., and Kenneth R. Paap. 1997. "Functional Neuroimages Fail to Discover Pieces of Mind in Parts of the Brain." *Philosophy of Science* 64(Supplemental Proceedings): S85–S94.

Yarkoni, Tal, Russell A. Poldrack, Thomas E. Nichols, David C. Van Essen, and Tor D. Wager. 2011. "Large-Scale Automated Synthesis of Human Functional Neuroimaging Data." *Nature Methods* 8(8): 665–70.

Response from Marieke van Rooij and John G. Holden*

Van Rooij and Van Orden outlined several practical and theoretical concerns about research efforts that link specific spatial locations in the brain to specific economic, emotional, and cognitive structures, in the Fall 2011 issue ("It's About Space, It's About Time, Neuroeconomics and the Brain Sublime," pp. 31–56). They offered an alternative view focused on the brain and body's temporal dimension. The alternative proposal is that thought and behavior are supported by flexible, self-organizing, and dynamic pattern formation processes (Van Orden, Holden, and Turvey 2003, 2005). In their letter, Camerer, Smith, and their co-signers question the validity of this critique. Our response is directed at their essential claim that our skepticism regarding spatial modularity of brain function has been superseded and even largely disproven.

Attempts to relate spatial brain coordinates to high-level cognitive functions are motivated by a *subtractive* logic. More active brain regions require more nutrients such as oxygen and glucose than less active regions. Brain imaging methods measure various markers that accompany the resulting metabolic changes in blood flow, either directly or indirectly. For example, fMRI measures the BOLD signal, the ratio of oxyhemoglobin (Hb) to deoxyhemoglobin (dHb) in the cerebral blood flow. Typically, for each volume element (voxel) of the brain, the relative metabolic activity corresponds to one pixel in a corresponding neuroimage. Thus, increased brain activation accompanied by larger BOLD signals in an fMRI scan, map to increased intensity levels in a neuroimage. The location of cognitive function is identified with differences in normalized and averaged brain activity, in contrasts of baseline and experimental images. Thus metabolic activity is statistically subtracted, yielding an image of relative activation differences.

All these statistical operations are subsumed by what is commonly known as the General Linear Model, a statistical model that includes regression and related methods such as analysis of variance. These methods form the basis of *statistical parametric mapping*, for example (for example, Friston, Holmes, Worsley, Poline, Frith, and Frackowiak 1994; Friston et al. 2007). The approach is widely used and cited in the neuroimaging literature (for example, Mumford and Nichols 2009). The standard regression model states that an observable equals a weighted sum of predictors, plus error. Relying on these techniques in imaging analyses is logically equivalent to adopting a general linear model of the brain. As our commenters write, neuroscientists do not use the phrase "general linear model of the brain," because they do not view it as an accurate description of how the brain works. Knowingly or not, neuroimagers using the subtractive logic assume, as van Rooij and Van Orden wrote in the original article, "that the brain can be treated as the sum of its parts."

According to Roskies's (2010) defense of subtraction techniques, they rely on three crucial, a priori (before the fact), assumptions: 1) "brain instantiates mind," 2) "different localized regions of tissue have different and stable functionalities," and 3) "blood flow is a guide to neural activity." Each of these assumptions is questionable.

Regarding assumption 1, debate on the nature of mind-body relations is a historically persistent and unsettled topic in biology, psychology, and metaphysics.

Assumption 2 that "different localized regions of tissue have different and stable functionalities" is best described as a working research hypothesis: *If* the brain is composed of distinct, modular, functional regions that are causally transparent to thought and behavior, *then* differences in spatial brain activation can be functionally interpreted. The potential difficulties with this approach have long been recognized, and remain conspicuously unanswered by advocates of subtraction such as Camerer, Smith, and their co-signers.

Our concern is that much of the neuroimaging enterprise is recapitulating previously established localizationist failures in behavioral methods. In 1896, F. C. Donders, a Dutch ophthalmologist, developed the hypothesis of distinct component sub-operations intervening between a stimulus and response, and devised the subtractive method to investigate these components. The limitation of Donders's method, which persists to this day, is that the hypothesized distinct components and subtasks must be known before the fact for the resulting subtractions to be meaningful (Uttal 2001). For instance, lesions to certain parts of the brain, and how they may affect specific behaviors, were among the earliest localizationist targets in behavioral studies (for example, Broca 1861). Nevertheless, the storied history of language-based dissociation studies has failed to converge on a view of the brain as including fixed modules or even on criteria for determining if such modules are an accurate description of how the brain works (Van Orden, Pennington, and Stone 2001).

The subtraction approach to studying localization of brain activity depends crucially and circularly on both theory and task. As the comment from Camerer and Smith illustrates, the theoretical debates inevitably degenerate into irresolvable disputes about what aspects of task and method are displayed in a specific hypothetical brain function (Van Orden, Pennington, and Stone 2001). Given a vast and variegated set of potential brain studies from which to choose, and the fact that minor task details routinely yield contradictory theoretical narratives, it is unsurprising that they or we or anyone else could reach endlessly contradictory conclusions regarding the same basic human activity, such as gambling.

Camerer, Smith, and their co-signers discuss how cross-method studies seek to resolve such issues. In practice, cross-method studies only amplify the already vast pool of potential contrasts, priorities, and perspectives on measured variables. They also claim that meta-analysis provides a solution, but meta-analysis is rooted in the self-same linear statistical system: "[T]he null hypothesis is that the *n* peak coordinates reported in the set of studies to be analyzed are *randomly and uniformly distributed* throughout gray matter" (Wager, Lindquist, and Kaplan 2007, p. 153, italics added).

By the mid-twentieth century, Saul Sternberg (1969) proposed an alternative method to discover subtask components in behavioral data that relieved the requirement that scientists know, before the fact, the functional components that intervene between a presented stimulus and participant's response. Sternberg proposed an *additive factors* logic. Broadly speaking, this approach used factorial manipulations to examine how components might combine their effects additively.

Sternberg's approach was potentially tractable, and within a few decades, an enormous set of psychological factors were assessed in multifactor behavioral experiments. Regarding the localizationist enterprise, to date, "not one cognitive mechanism exists on which cognitive scientists can agree about its boundaries, its empirical shape, or details about its function" (Van Orden, Holden, and Turvey 2005, p. 121). Instead, complex, contextually embedded chains of interactions among factors are routinely observed across studies (Van Orden, Pennington, and Stone 2001). This outcome is expected in the absence of modularity.

Indeed, Camerer, Smith, and their co-signers implicitly arrive at the same conclusion when they refer to the localization of "multiple regions" and "brain networks." The imaging literature now reports increasingly complex and distributed brain networks, measured during both tasked and untasked conditions (for example, Bullmore and Sporns 2009; Ciuciu, Varoquaux, Abry, Sadaghiani, and Kleinschmidt 2012). Multiple motifs, or repeating connectivity patterns, are nested within a network's connectivity. Different architectures that combine these low-level motifs provide for distinct local network flow patterns. The emerging emphasis is functional *connectivity*; the study of such patterns is sometimes called "connectomics," by analogy to genomics. In stark contrast to subtracted images that identify just a few functional locations, the network approach reveals so many interconnected regions, scientists must wonder what regions are *not* functionally associated with a targeted activity (for example, Anderson 2010).

Assumption 3, that "blood flow is a guide to neural activity," means that neuroimagers typically highlight those brain regions where the relative difference in blood flow between baseline and experimental sessions are at or near maximum. But what is special about a brain region that displays the largest differences in nutrient consumption?

Instead of considering blood flow measures in a brain, consider for a moment traffic flow measures in a city. Imagine that a researcher has data available on vehicles operating with the highest fuel consumption at any given time (that is, maximum nutrient consumption), or vehicles with full gas tanks (that is, maximum Hb to dHb ratios). It would clearly be misguided to conclude that only these vehicles are participating in transportation activities. In reality, the overwhelming bulk of transportation activities in a crowded city corresponds to vehicles that are not at extremes of fuel use or fuel storage—even stopped vehicles, at intersections for instance, crucially support transportation. It seems similarly unwise to assume that the brain areas with high nutrient consumption or BOLD signals are the only parts of the brain involved in a specific act of thinking.

As stated, fMRI statistical practices are rooted in the General Linear Model. Linear statistical methods are designed to discover and distinguish separable sets of constants (or means) shrouded by unsystematic, independent, and homogeneous sources of noise. As the fractal physiology literature cited by van Rooij and Van Orden makes clear, physiological signals rarely conform to such static assumptions. Instead, strong autocorrelation and inherent (fractal) fluctuations are fundamental properties of nearly all physiological signals. Long ago, it was recognized that applying static linear statistical methods to measurements that express intrinsic, long-range fluctuations, such as cardiovascular diffusion and flow, can be counted on to routinely yield spurious differences (as an example, see the Yule-Slutsky effect discussed in Klein, 1997).

The patterns we are discussing are also referred to as "scale-free." In a model of the brain using subtractive logic, in which different tasks are associated with specific locations in the brain, the brain is divided into modules that can be labeled as either contiguous or isolated based on their distance from each other. In a scale-free system, there is no "typical" distance or time scale, in the sense that any chosen size of scale leads to different conclusions about which modules are involved and interconnected. Scale-free behavior is routinely reported in neuronal, fMRI BOLD, EEG, and behavioral signals. On this point, imaging and behavioral studies are in an unusual position of agreement regarding the empirical facts. Ubiquitous scale-free patterns in physiology and behavior implicate underlying flexible, self-organizing, and dynamic pattern formation processes, as discussed by van Rooij and Van Orden.

This exquisite dynamic flexibility has crucial implications. The instantaneous neurophysiologic details of every human thought or act may be sufficiently idiosyncratic that neuroimages reveal largely transient patterns. If so, then imaging practices that seek to determine a specific location in the brain associated with a specific cognitive function, are analogous to attempting to infer a theory of lightning by tracing the paths of reams of individual lightning bolt images. As Uttal (2001) famously claimed, subtractive brain imaging is effectively a new phrenology.

Marieke van Rooij
John G. Holden
University of Cincinnati
Cincinnati, Ohio

*Note: Guy Van Orden passed away May 11th, 2012. Thus, John G. Holden, one of Guy's long-time collaborators, served as co-author on this response.

References

Anderson, Michael L. 2010. "Neural Reuse: A Fundamental Organizational Principle of the Brain." *Behavioral and Brain Sciences* 33(4): 245–313.

Broca, P. Paul. 1861. "Perte de la Parole, Ramollissement Chronique, et Destruction Partielle du Lobe Antérieur Gauche du Cerveau." *Bulletin de la Société Anthropologique* 2: 235–38.

Bullmore, Edward, T., and Olaf Sporns. 2009. "Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems." *Nature Reviews Neuroscience* 10(3): 186–98.

Ciuciu, Philippe, Gaël Varoquaux, Patrice Abry, Sepideh Sadaghiani, and Andreas Kleinschmidt. 2012. "Scale-Free and Multifractal Time Dynamics of fMRI Signals during Rest and Task." *Frontiers in Physiology* 3(June): Article 186.

Friston, Karl, J., Andrew P. Holmes, Keith J. Worsley, Jean-Baptiste Poline, Christopher D. Frith, and Richard S. Frackowiak. 1994. "Statistical Parametric Maps in Functional Imaging: A General Linear Approach." *Human Brain Mapping* 2(4): 189–210.

Friston, Karl J., et al. 2007. "Multiple Sparse Priors for the M/EEG Inverse Problem." *NeuroImage* 39(3): 1104–20.

Klein, Judy L. 1997. *Statistical Visions in Time: A History of Time Series Analysis, 1662–1938*. Cambridge University Press.

Mumford, Jeanette, A., and Thomas E. Nichols. 2009. "Simple Group fMRI Modeling and Inference." *NeuroImage* 47(4): 1469–75.

Roskies, Adina, L. 2010. "Saving Subtraction: A Reply to Van Orden and Paap." *British Journal for the Philosophy of Science* 61(3): 635–65.

Sternberg, Saul. 1969. "The Discovery of Processing Stages: Extensions of Donders' Method." *Acta Psychologica* 30(C): 276–315.

Uttal, William R. 2001. *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*. Cambridge, MA: MIT Press.

Van Orden, Guy, John G. Holden, and Michael T. Turvey. 2003. "Self-Organization of Cognitive Performance." *Journal of Experimental Psychology: General* 132(3): 331–50.

Van Orden, Guy, John G. Holden, and Michael T. Turvey. 2005. "Human Cognition and 1/f Scaling." *Journal of Experimental Psychology: General* 134(1): 117–23.

Van Orden, Guy, Bruce F. Pennington, and Gregory O. Stone. 2001. "What Do Double Dissociations Prove?" *Cognitive Science* 25(1): 111–72.

Wager, Tor D., Martin A. Lindquist, and Lauren, A. Kaplan. 2007. "Meta-Analysis of Functional Neuroimaging Data: Current and Future Directions." *Social Cognitive and Affective Neuroscience* 2(2): 150–58.

Journal of Economic Perspectives—Volume 27, Number 2—Spring 2013—Pages 253–254

Notes

For additional announcements, check out the continuously updated JEP online Bulletin Board, *(http://www.aeaweb.org/bulletinboard.php). Calls for papers, notices of professional meetings, and other announcements of interest to economists should be submitted to Ann Norman, at jep@jepjournal.org, in one or two paragraphs containing the relevant information. These will be posted at the JEP online Bulletin Board. Given sufficient lead time (at least one month before an issue goes online), we will also print a shorter, one-paragraph version of your notice in the "Notes" section of the* Journal of Economic Perspectives. *We reserve the right to edit material received.*

The Annual Meeting of the American Economic Association will be held in Philadelphia, Pennsylvania, January 3–5, 2014. The headquarters is the Philadelphia Marriott Downtown. Information and procedures for employers and job seekers are in the registration material at www.vanderbilt.edu/AEA. For additional information, please go to www.vanderbilt.edu/AEA.

John Bates Clark Medal. The American Economic Association is pleased to announce that Raj Chetty was awarded the John Bates Clark Medal for 2013.

2013 Distinguished Fellows. The Association is pleased to announce that the Distinguished Fellows for 2013 are Harold Demsetz, Stanley Fischer, Jerry Hausman, and Paul Joskow.

The 2013 *American Economic Journal* (AEJ) "Best Paper Prize" highlights the best paper published in each of the *American Economic Journals: Applied Economics*, *Economic Policy*, *Macroeconomics*, and *Microeconomics* over the last three years. Nominations are provided by AEA members, and winners are selected by the journals' editors, co-editors, and boards of editors. Complimentary full-text articles are available at the web addresses below.

AEJ: Applied Economics: "The Short- and Long-Term Career Effects of Graduating in a Recession," Philip Oreopoulos, Till von Wachter, and Andrew Heisz, vol. 4, no. 1 (January 2012), pp. 1–29, http://www.aeaweb.org/articles.php?doi=10.1257/app.4.1.1

AEJ: Economic Policy: "Does State Fiscal Relief During Recessions Increase Employment? Evidence from the American Recovery and Reinvestment Act," Gabriel Chodorow-Reich, Laura Feiveson, Zachary Liscow, and William Gui Woolston, vol. 4, no. 3 (August 2012), pp. 118–45), http://www.aeaweb.org/articles.php?doi=10.1257/pol.4.3.118

AEJ: Macroeconomics: "How Much Does Immigration Boost Innovation?" Jennifer Hunt and Marjolaine Gauthier-Loiselle, vol. 2, no. 2 (April 2010), pp. 31–56, http://www.aeaweb.org/articles.php?doi=10.1257/mac.2.2.31

AEJ: Microeconomics: "Information Disclosure and Unraveling in Matching Markets," Michael Ostrovsky and Michael Schwarz, vol. 2, no. 2 (May 2010), pp. 34–63, http://www.aeaweb.org/articles.php?doi=10.1257/mic.2.2.34

Resources for Economics (RFE), a guide sponsored by the AEA, lists more than 2,000 resources in 97 sections available at http://www.aeaweb.org/rfe/index.php. The guide is designed for economics students of all levels, as well as academics, practitioners, and those simply interested in economics.

***EconLit* for Members** is a searchable bibliographic database. Compiled and abstracted in an easily searchable format, *EconLit* is a comprehensive index of journal articles, books, book reviews, collective volume articles, working papers, and dissertations. It contains all the *Econlit* records, but does not link to full-text journal articles and books available on library and institutional installations. It is primarily a tool for members who do not enjoy institutional access to *EconLit*. To find out more go to http://www.aeaweb.org/econlit/efm/index.php.

Retired faculty available for part-time or temporary teaching. JOE (Job Openings for Economists) now lists retired economists interested in teaching on either a part-time or temporary basis. Individuals can add or delete their name any time.

Listings are deleted on November 30; the service is closed during December and January, re-opening February 1. Register at http://www.aeaweb.org/joe/available_faculty/.

Publication updates on Twitter. Want to see forthcoming article previews as soon as they are available for the *American Economic Review*, *American Economic Journals*, and the *Journal of Economic Literature*? Follow us on Twitter! https://twitter.com/AEAjournals/.

Call for papers. The 15th annual conference of the **National Business & Economics Society** will be held March 12–15, 2014, at the Mauna Lani Bay Resort on the Kohala Coast of Hawaii's Big Island. NBES is a multidisciplinary academic association that focuses on promoting research of both a theoretical and practical nature. See the NBES website at www.nbesonline.com. Interested authors should submit a full paper or a 1–2 page abstract to info@nbesonline.com. **Submission deadline: August 31, 2013.**

American Institute of Indian Studies Fellowship Competition. The American Institute of Indian Studies invites applications from scholars from all disciplines who wish to conduct their research in India. Junior fellowships are given to doctoral candidates to conduct research for their dissertations in India for up to eleven months. Senior long-term (six to nine months) and short-term (four months or less) fellowships are available for scholars who hold the PhD degree. Scholarly/Professional development fellowships are available to scholars and professionals who have not previously worked in India. Eligible applicants include 1) US citizens; and 2) citizens of other countries who are students or faculty members at US colleges and universities (this rule does not apply to US citizens). Applications can be downloaded from the website: www.indiastudies.org. Inquiries should be directed to (773) 702-8638 or aiis@uchicago.edu. **Application deadline is July 1, 2013.**

Call for papers: ***Indian Journal of Economics and Business*** **(IJEB)** and Serials Publications, New Delhi, are happy to announce the next meeting: the International Conference on Economic and Business Issues, on December 19 and 20, 2013, at Hotel Grand Ramee, Apte Road, Pune 411004, India. The paper and abstract **submission deadline is June 30, 2013**, but early submissions are encouraged. The registration deadline is October 31, 2013. For more information contact Dr. Kishore G. Kulkarni, Distinguished Professor of Economics and Editor, *Indian Journal of Economics and Business*, (visit www.ijeb.com), Metropolitan State University of Denver at kulkarnk@msudenver.edu or 303-556-2675 (phone) or 303-556-3966 (fax).

Call for papers: ***Academic Journal of Management Sciences*** **(AJMS)** is scholarly academic journal based in Pakistan, published quarterly by International Society of Universal Research in Sciences (EyeSource). AJMS is an online open access journal. It invites both academicians and practitioners to publish their core research ideas, having significant impact on business, government policies, and implications for the practical world. Researchers are welcome to challenge or extend the existing body of knowledge in the following research areas: management, organizational behavior, entrepreneurship, economics, accounting and finance, production and operations management, human resource management, strategic management, marketing, government and public policy, disaster management, supply chain management, and e-management practices. For more information please visit: http://isurs.org/view jc.php?id=j2.

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary *JEP* for your e-reader, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $20.00, $30.00, or $40.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

Founded in 1885

The Journal of

Economic Perspectives

Spring 2013, Volume 27, Number 2

Symposia

The Growth of the Financial Sector

Early and Later Interventions

Articles

AMERICAN
ECONOMIC
ASSOCIATION