The Journal of Economic Perspectives

A journal of the
American Economic Association

Summer 2014

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.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Summer 2014, Vol. 28, No. 3. 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 28 • Number 3 • Summer 2014

Symposia

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.

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.

The Journal of

Economic Perspectives

Summer 2014, Volume 28, Number 3

Symposia

Entrepreneurship

Classic Ideas in Development

Academic Production

Articles

AMERICAN ECONOMIC ASSOCIATION

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 3–24

The Role of Entrepreneurship in US Job Creation and Economic Dynamism[†]

Ryan Decker, John Haltiwanger, Ron Jarmin, and Javier Miranda

The United States has long been viewed as having among the world's most entrepreneurial, dynamic, and flexible economies. It is often argued that this dynamism and flexibility has enabled the US economy to adapt to changing economic circumstances and recover from recessions in a robust manner. While the evidence provides broad support for this view, the outcomes of entrepreneurship are more heterogeneous than commonly appreciated and appear to be evolving in ways that could raise concern. Evidence along a number of dimensions and a variety of sources points to a US economy that is becoming less dynamic. Of particular interest are declining business startup rates and the resulting diminished role for dynamic young businesses in the economy.

We begin by describing how the concept of entrepreneurship is reflected in existing data on firm age and size. The recent addition of firm age to official statistics represents a dramatic improvement in the information available to entrepreneurship researchers. We then turn to a discussion of the role of startup firms in job creation. Business startups account for about 20 percent of US gross (total) job creation while high-growth businesses (which are disproportionately young)

■ *Ryan Decker is a PhD student in Economics, University of Maryland, College Park, Maryland. John Haltiwanger is the Dudley and Louisa Dillard Professor of Economics and Distinguished University Professor, University of Maryland, College Park, Maryland, and also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Ron Jarmin is Assistant Director for Research and Methodology, US Census Bureau, Washington, DC. Javier Miranda is a Principal Economist, US Census Bureau, Washington, DC. The authors' email addresses are decker@econ.umd.edu, haltiwan@econ.umd.edu, ron.s.jarmin@census.gov, and Javier.miranda@census.gov. Haltiwanger is the corresponding author.*

[†] To access the data Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.3.3

doi=10.1257/jep.28.3.3

account for almost 50 percent of gross job creation. Startups and young businesses are small, the underlying reason many commentators described small businesses as the engine of US job growth prior to availability of data by firm age (for summaries, see Haltiwanger, Jarmin, and Miranda 2013; Haltiwanger 2012).

The contribution of startups and young businesses to job creation involves rich dynamics. Most business startups exit within their first ten years, and most surviving young businesses do not grow but remain small. However, a small fraction of young firms exhibit very high growth and contribute substantially to job creation. These high-growth firms make up for nearly all the job losses associated with shrinking and exiting firms within their cohort. The implication is that each entering cohort of startups makes a long-lasting contribution to net job creation.

The contribution of startups and young firms to job creation is part of an overall rapid pace of reallocation of productive resources across firms in the US economy. Young firms exhibit rich post-entry dynamics: specifically, low-productivity young firms contract and exit, while high-productivity young firms rapidly expand. In addition, young firms appear to play a critical role in innovative activity that also contributes to productivity growth (including within-firm productivity growth).

An optimal pace of business dynamics—encompassing the processes of entry, exit, expansion, and contraction—would balance the benefits of productivity and economic growth against the costs to firms and workers associated with reallocation of productive resources. It is difficult to prescribe what the optimal pace should be, but evidence accumulating from multiple datasets and methodologies suggests that the rate of business startups and the pace of employment dynamism in the US economy has fallen over recent decades and that this downward trend accelerated after 2000 (Haltiwanger, Jarmin, and Miranda 2011; Reedy and Litan 2011). A critical factor in accounting for the decline in business dynamics is a lower rate of business startups and the related decreasing role of dynamic young businesses in the economy. For example, the share of US employment accounted for by young firms has declined by almost 30 percent over the last 30 years.

These trends suggest that incentives for entrepreneurs to start new firms in the United States have diminished over time. We do not identify all the factors underlying these trends in this paper but offer some clues based on the empirical patterns for specific sectors and geographic regions. We conclude with reflections on ripe research topics in this area.

Measuring Entrepreneurship

Measuring entrepreneurship and its economic effects is difficult. Available government data on US firms do not have a specific entry for "entrepreneurs." These data have traditionally contained information about the size of firms, and thus some observers have written or spoken as if small businesses are synonymous with entrepreneurs. However, we believe entrepreneurial activity is better represented by *new* businesses—that is, by age rather than by size. Indeed, using recently

available data with information on both firm *size* and firm *age*, Haltiwanger, Jarmin, and Miranda (2013) show that most of the job-creating prowess of small businesses is due to the contribution of startups and young businesses, which are also small at that stage in their lifecycle, as we discuss in the next section.

When thinking about "new" businesses, the distinction between new *firms* and new *establishments* is critical. Many US firms operate multiple establishments (that is, specific locations of business activity). New establishments of existing firms can take many forms, including simply replacing outdated existing establishments, thus new establishments often do not conform to standard notions of entrepreneurial behavior. For this reason, entrepreneurial research should focus on startups and young firms. It should exclude "new" businesses emerging from reorganizations such as mergers and acquisitions.

The Business Dynamics Statistics and the Longitudinal Business Database maintained by the US Census Bureau track the universe of employer firms—that is, firms in the private, nonagricultural sector with at least one employee.[1] The Longitudinal Business Database includes annual observations beginning in 1976 and currently runs through 2011. It provides information on detailed industry, location, and employment for every establishment in the private sector in a nationally comprehensive and integrated manner. With these data, it is possible to distinguish between true "startups," new establishments of existing businesses, and "new" firms formed by combining pre-existing establishments through merger and acquisition activity. The Longitudinal Business Database identifies the relationship between firms and establishments both cross-sectionally and over time. Thus, researchers can link establishment and firm information, compute characteristics such as firm size and firm age from the establishment information, and track patterns over time.

However, even within the category of startups, we should expect to find various types of entrepreneurs. Schoar (2010) argues for distinguishing between "subsistence" entrepreneurs and "transformational" entrepreneurs. Her distinction was intended primarily for emerging economies where many entrepreneurs have limited prospects for growth, but we think this distinction is useful for the US economy as well. Subsistence entrepreneurs can be thought of as those that create small businesses that provide employment for the entrepreneur and perhaps a few others (often family members), which do not usually grow. For example, Hurst and Pugsley (2011) find that many young and small business owners in the US economy state they do not have aspirations for high growth, but rather often started businesses for nonpecuniary reasons like time flexibility or personal goals. Transformational

[1] The Business Dynamic Statistics are recently available public domain data derived from the Longitudinal Business Database, and can be found at http://www.census.gov/ces/dataproducts/bds/. A fuller description of these data and the measures we use are provided in the online Appendix available with this paper at http://e-jep.org. We note that the Longitudinal Business Database employment and job creation numbers track closely those of the County Business Patterns and Statistics of US Business programs of the US Census Bureau (Haltiwanger, Jarmin, and Miranda 2009), as they all share the Census Bureau's Business Register as their source data.

entrepreneurs, on the other hand, create small startup businesses with the intention to innovate and grow, thus creating employment for other workers and value added for the economy. Clearly, only a subset of these transformational entrepreneurs are likely to succeed in the marketplace and grow. When people discuss the importance of entrepreneurs in job creation and productivity growth, they are envisioning transformational entrepreneurs, not subsistence entrepreneurs.

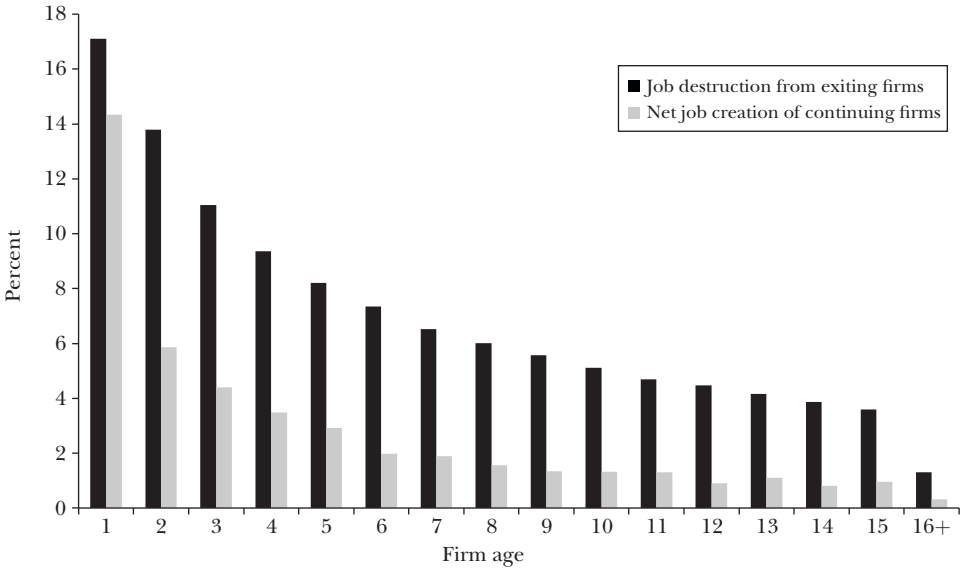
Other useful sources of data on startups include the Business Employment Dynamics from the US Bureau of Labor Statistics, which is a rich source of quarterly information on job creation and destruction. These data are more timely than the Census data but are less useful for studying firm size and firm age effects because firms are defined based on all activity operating under a single Employer Identification Number. However, many large multi-establishment firms have multiple Employer Identification Numbers. By contrast the Census Bureau uses survey and administrative data to create an enterprise-wide representation of all the establishments in the US economy that are under common ownership and control. Other important sources of data describing firm dynamics for the US economy include COMPUSTAT (which tracks publicly traded firms), the National Establishment Time Series (based on Dun and Bradstreet data), and the Kauffman Survey tracking a sample of recent entrants.

Startups and US Jobs

Startups and young businesses clearly play an important role in job creation. Between 1980 and 2010, the gross number of jobs created annually by all establishments averaged about 18 percent of the workforce—an average of 16.3 million jobs per year—according to our calculations from the Business Dynamics Statistics. About one-sixth of this amount, an average of 2.9 million jobs annually, can be traced to new firms, and another one-sixth can be traced to new establishments of existing firms. The net job creation statistics are even more striking. For new firms (that is, those with age equal to zero in the Business Dynamics Statistics), "net" and "gross" job creation are the same, because they have no previous jobs to lose, and so their net job creation is also 2.9 million jobs per year. Over these 30 years, average net job creation in the entire US private sector was approximately 1.4 million jobs per year. The implication is that cohorts of firms aged one year or older typically exhibit net job declines.

Taken at face value, these statistics might be interpreted as implying that all net job creation is due to startups. This interpretation is misleading for two reasons. First, gross job creation is an order of magnitude greater than net job creation in any given period. This implies that in any given period there are many different groupings of growing firms that could be used to account for net job creation. Second, it is important to examine the post-entry dynamics of startups. If, for example, all startups failed after a short period of time, then startups would not be making any long-lasting contribution to net job creation.

Figure 1
Up or Out Dynamics for Young Firms



Source: Annual averages of statistics computed from the Longitudinal Business Database from 1992–2011.
Notes: Figure 1 shows patterns of net employment growth of continuing firms and job destruction from firm exit for firms age 1 and older. Startups have firm age equal to zero, so this figure reports on the post-entry dynamics of firms. (See footnotes 2 and 3 and online Appendix for details.)

To explore post-entry dynamics, we need to track firm growth and survival as a function of firm age. We rely here on the methodology developed by Davis, Haltiwanger, Jarmin, and Miranda (2007) and Haltiwanger, Jarmin, and Miranda (2013). Firm age is measured using the age of the oldest establishment in the firm. For startups, all of the establishments of the new organization are entrants so firm age is zero. In this methodology, continuing firms age "naturally," one year at a time, as long as the organization stays in existence.[2] Consistent with this approach, firm exits represent legal entities that cease to exist and in which all of their associated establishments shut down. Thus, firm exits do not reflect legal entities that cease through organizational change or buyout activity and where at least some establishments continue operation in subsequent years.

Using this approach, Figure 1 shows patterns of net employment growth for continuing firms and job destruction from firm exits for firms age 1 and older.[3]

[2] As part of this same methodology for assigning firm age, Haltiwanger, Jarmin, and Miranda (2013) develop a method for capturing firm growth that focuses on organic growth rather than growth from merger and acquisition activity and other related changes in organization. We use that methodology here, and explain it in more detail in the Appendix available with this paper at http://e-jep.org.

[3] Reported statistics in this figure reflect the net employment growth for the cell using the growth rate methodology developed by Davis, Haltiwanger, and Schuh (1996). The growth rate concept at any level

Recall that startups have firm age equal to zero, and so Figure 1 provides insights on the post-entry dynamics of firms. The statistics are based on tabulations of pooled data from 1992–2011 from the Longitudinal Business Database. Conditional on survival, younger firms have much higher rates of job growth than more mature firms. Indeed, even the typical 10 year-old firm has net growth that is 1.4 percentage points higher than the typical 16+ year-old firm. The exit dynamics look different as well. Young firms have a substantially higher exit rate (job destruction from exit is an employment-weighted exit rate): 50 percent of the jobs generated by an entering cohort of firms will have been lost to exits by age five.

Thus, young firms exhibit a strong "up or out" dynamic (Haltiwanger, Jarmin, and Miranda 2013). For any given cohort, jobs lost due to the high failure rate of young firms are almost offset by the growth of the surviving firms. Five years after the entry of a typical cohort, total employment is about 80 percent of the original employment contribution of the cohort—in spite of losing about 50 percent of the original employment to business exits. In this sense, cohorts of startups do make a long-lasting contribution to net job creation. But the high volatility of young firms highlights that the contribution to net job creation from startups can't be simply understood from the immediate contribution upon starting up.

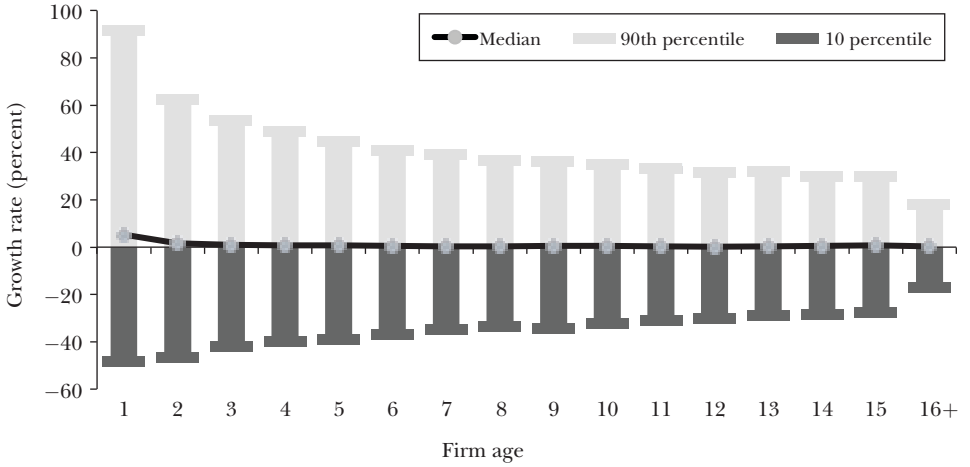
The high mean net employment growth of surviving young firms masks enormous heterogeneity among young surviving firms. Figure 2A shows the 90th, 50th, and 10th percentiles net job growth of surviving firms by firm age. Figure 2B shows the mean and median of net employment growth of surviving firms by firm age. The mean is the aggregate net growth rate for each age group. The 90th, 10th, and median are the employment-weighted percentiles, so they reflect the employment-weighted distribution underlying the aggregate mean. Young firms have very high dispersion of growth and also very high skewness. The skewness is seen in the relative magnitudes of the 90th and 10th percentiles where the growth rates of younger firms are much more skewed to the right (positive) compared to more mature firms. This accounts for the high mean growth rate of young firms relative to older firms that is evident in the lower panel of Figure 2. In short, the typical young firm (as captured by the median) exhibits little or no growth even conditional on survival (many of these are presumably the "subsistence" entrepreneurs discussed earlier), however, among all the young firms, a few do exhibit very high rates of growth which yields a high mean growth rate.

The skewed right tail of the growth rate distribution of young firms is important for understanding the contribution of startups and young firms to overall job creation. Startups account for less than 10 percent of firms and about 20 percent of firm-level gross job creation. But the contribution of startups to job creation does not stop at entry, at least for some firms. High-growth firms—defined here as firms expanding

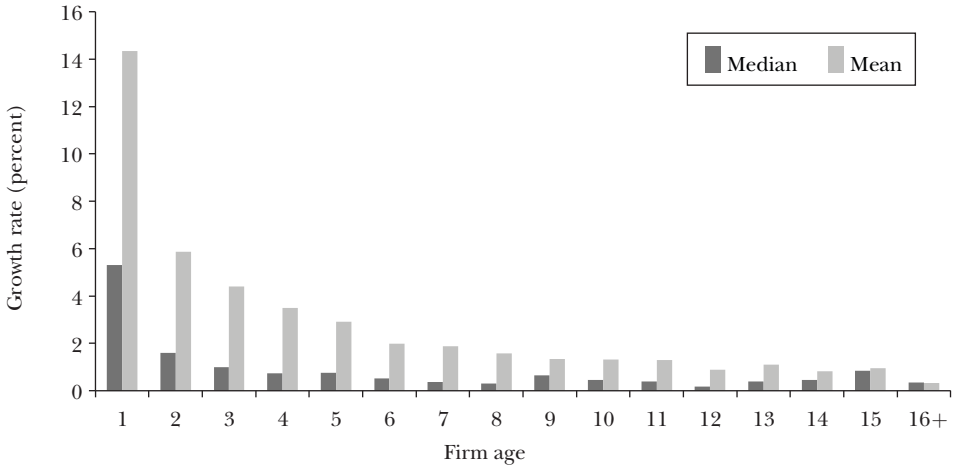
of aggregation is based on the change in the number of jobs for a cell from period $t-1$ to t divided by the average number of jobs in periods $t-1$ and t. This growth rate measure was developed originally by Törnqvist, Vartia, and Vartia (1985). Like log changes, this growth rate measure is symmetric but it has the advantage of accommodating zeroes (entry and exit). It is a second order Taylor approximation to the log first difference. See the online Appendix with this paper at http://e-jep.org for more details.

Figure 2
Net Employment Growth and Growth Rates for Surviving Firms

A: 90th, Median, and 10th Percentiles of Net Employment Growth for Surviving Firms



B: Mean and Median Net Employment Growth Rates for Surviving Firms



Notes: Annual averages of statistics computed from Longitudinal Business Database, from 1992–2011. The 90th, 10th, and median are all based on the employment-weighed firm level growth rate distribution for each firm age cell. The mean is the aggregate net growth rate of the firm age cell which is equivalent to the employment-weighted average of the firm level net growth rate in each cell.

their employment by more than 25 percent per year—account for about 15 percent of firms and 50 percent of firm-level gross job creation. Together, startups and high-growth firms (which are disproportionately young as seen in Figure 2) account for about 70 percent of firm-level gross job creation in a typical year. Balancing this positive contribution is the sharp job loss that occurs for many firms in the first several

Table 1
Estimates of Net Growth for Continuing Firms by Firm Size and Firm Age
(in percent)

	Firm size (number of employees)								
Firm age	*1–4*	*5–9*	*10–19*	*20–49*	*50–99*	*100– 249*	*250– 499*	*500+*	*All sizes*
1–2	6.7	9.0	11.5	12.9	14.9	14.5	14.3	17.9	11.8
3–4	2.0	2.5	4.4	5.6	7.4	6.8	6.7	10.1	4.5
5–6	−0.5	−0.2	1.8	2.9	4.6	3.9	3.7	7.0	3.0
7–8	−2.0	−1.2	0.8	2.1	3.7	3.0	2.9	6.0	2.4
9–10	−3.7	−2.4	−0.2	0.9	2.5	1.8	1.5	4.5	1.8
11–12	−2.1	−0.9	0.9	2.4	3.8	3.0	2.5	5.5	1.9
13–15	−0.8	0.0	3.0	3.5	4.4	4.0	3.2	5.7	1.8
16+	−2.4	−1.7	−0.9	−0.1	0.6	1.3	1.6	0.9	0.7
All ages	0.2	1.2	1.7	2.1	2.5	2.9	3.1	1.4	

Notes: Tabulations from the US Census Bureau. Details of methodology and results can be found in the online appendix of Haltiwanger, Jarmin, and Miranda (2013). The analysis covers the period 1992–2005.

years after startup entry. Overall, the evidence shows that most startups fail, and most that do survive do not grow. But among the surviving startups are high-growth firms that contribute disproportionately to job growth. These high-growth young firms yield the long-lasting contribution of startups to net job creation.

To understand the population of startups and their dynamics, it helps to realize that their size distribution is quite different from that of the overall distribution of firms. Startups are small; more than 90 percent of all startups have fewer than 20 employees, and these small startups account for about half of all startup employment. For purposes of comparison, about 50 percent of employment in the US private sector is accounted for by the less than 1 percent of US firms with more than 500 employees. Large startups are almost nonexistent and account for a very small share of startup activity; however, many small firms are old, accounting for about 30 percent of employment. This image is enriched by examining the average net employment growth rate by firm age and firm size class in Table 1. Table 1 focuses on continuing firms, to help link back to the evidence in Figure 1 regarding the high average net growth rate for young surviving firms. Average net job growth falls monotonically with firm age for all firm size groups. However, the average net growth for young firms is substantially higher for firms that are larger than 20 employees. Such patterns highlight that rapid employment growth among young surviving firms is especially present among larger—or at least not micro-sized—young firms.

Startups, Reallocation, and Productivity

The high pace of labor market reallocation with a critical role for startups and young firms raises a question: What is driving this reallocation? Considerable

evidence suggests that entry, exit, expansion, and contraction of firms are closely related to measures of productivity and profitability (for surveys and summaries, see Syverson 2011; Foster, Haltiwanger, and Krizan 2001; Bartelsman and Doms 2000).

Producers, even those within the same industry, display enormous differences in measured productivity and profitability. For example, within the same four-digit US manufacturing industry, the establishment at the 90th percentile of the productivity distribution produces almost twice as much output (measured by real revenue) with the same measured inputs as the 10th percentile establishment (Syverson 2004, 2011). These large differences are persistent, but not permanent, indicating that firms are subject to ongoing idiosyncratic productivity and profitability shocks.

These large differences in productivity and profitability are connected to the growth and survival dynamics of firms. As such, they underlie the reallocation dynamics we have discussed and documented above. For example, Foster, Grim, and Haltiwanger (2013) estimate, using data for the entire US manufacturing sector for the period 1980–2011, that the probability of exit for the establishment at the 90th percentile of the productivity distribution within an industry is 4 percentage points lower than the establishment at the 10th percentile (the average annual exit rate for manufacturing establishments is 8 percent per year). Conditional on survival and holding initial size constant, estimates from the same data imply that an establishment at the 90th percentile of productivity has a growth rate that is 3 percentage points higher than a plant at the 10th percentile. Finally, estimates from Syverson (2011) also show that entrants tend to start with productivity that is on average similar to incumbents. These estimated effects of productivity on growth and survival are very similar to those found in the literature. As stated in that paper: "[A] robust finding in the literature—virtually invariant to country, time period or industry—is that higher productivity producers are more likely to survive than their less efficient counterparts."[4]

When focusing on the relationship between reallocation of resources across the economy and firm-level measures of productivity, several issues arise. One issue is that much of the micro empirical literature in this area focuses on establishment-level rather than firm-level measures of productivity. Part of the reason for this focus reflects the complexity of measuring firm-level productivity for large, mature firms operating across several sectors. However, when the research looks just at establishments owned by young firms, as in Foster, Grim, and Haltiwanger (2013), the marginal effects of revenue-based measures of productivity on growth and survival (discussed above) are, if anything, larger. This is consistent with the up-and-out dynamics of young firms discussed earlier.

Another issue is that much of the literature in this area measures total factor productivity in terms of revenue. This choice is primarily due to data limitations as most sources of firm- and establishment-level data that are used to measure productivity do not include micro level measures of output and input prices. Instead, real

[4] We discuss the robustness of the findings in the literature further in section II of the online Appendix available with this paper at http://e-jep.org.

output and real inputs are measured as establishment- or firm-level revenues and costs divided by a price deflator. Such revenue-based measures of productivity will thus reflect some combination of the technical efficiency with which inputs are translated into outputs, together with demand and cost effects on prices.

Revenue-based measures of productivity may be a reasonable approximation for measures that adjust for price dispersion across producers. When Foster, Haltiwanger, and Syverson (2008) compare the revenue-based measures of total factor productivity with total factor productivity measured in physical quantities for a sample of manufacturing industries, they find that the correlation between these two measures of productivity is .75, which suggests that broad findings in the literature based on measures of revenue productivity would hold up reasonably well with measures of productivity based on quantities.[5] The industries for which physical quantity data are available in the US economy are limited, so there is some question as to the wider applicability of these findings across the whole economy. However, evidence for Colombia (where establishment-level price indices are available for all manufacturing establishments) suggests these patterns are robust for a much wider range of industries (Eslava, Haltiwanger, Kugler, and Kugler 2004, 2013).

To be clear, productivity growth in an economy is not only a matter of more-productive entering firms replacing less-productive exiting firms. A common finding in the literature about productivity growth in manufacturing is that about 60 percent of industry-level productivity growth happens within existing establishments and the rest comes from reallocation of productive resources resulting from entry, exit, and the expansion and contraction of existing establishments. For example, Foster, Haltiwanger, and Syverson (2008) find that entrants and young establishments have slightly higher total factor productivity (measured in quantity terms) than more mature incumbents, but the entrants have substantially higher productivity than exiting establishments. In their data, 35 percent of industry-level productivity growth is accounted for by net entry. However, their study looks over a five-year time period, and thus some of the 60 percent of productivity growth happening within existing establishments occurs in young firms. Foster, Haltiwanger, and Krizan (2001, 2006) provide evidence of such selection and learning dynamics and show that within-plant productivity growth is more rapid for surviving young establishments than more mature establishments.[6]

In sectors like retail trade, the evidence suggests that productivity growth within establishments is less important except in the case of young establishments

[5] In section II of the online Appendix available with this paper at http://e-jep.org, we discuss the evidence that shows the relationship between revenue productivity and growth and survival are very similar to the relationship between physical productivity and growth and survival.

[6] Foster, Haltiwanger, and Syverson (2008) also find that the role of net entry is substantially larger using total factor productivity measured in quantity terms as opposed to measuring it in revenue terms. The net entry component accounts for 35 percent of productivity growth with quantity-based productivity and 24 percent of productivity growth with revenue-based productivity. The difference must lie, of course, in price factors that differ between entrants and younger incumbents on the one hand and more mature incumbents on the other. They find that entrants and young establishments face substantially lower prices, which may reflect the challenges of becoming established and building reputation in a market.

that show productivity growth in their early years. Instead, most of the labor productivity growth in this sector has been attributed to net entry (Foster, Haltiwanger, and Krizan 2006). In many cases, existing firms improve productivity in retail trade primarily through adding new, more-productive retail locations rather than expanding existing establishments. Moreover, much of the exit of low-productivity retail establishments in the US economy has been dominated by the exit of "mom and pop" single-establishment firms. Doms, Jarmin, and Klimek (2004) find that more-intensive investment in information technology by large producers is related to both differences in the level of productivity across establishments and differences in the within-establishment growth rates of productivity, thus highlighting the link between technology, productivity, and business dynamics.

An important related question is the role of startups and young firms in innovative activity. Among innovative firms, young and small firms have higher innovation intensities than mature firms as measured by the ratio of research and development spending to sales (Acemoglu, Akcigit, Bloom, and Kerr 2013).

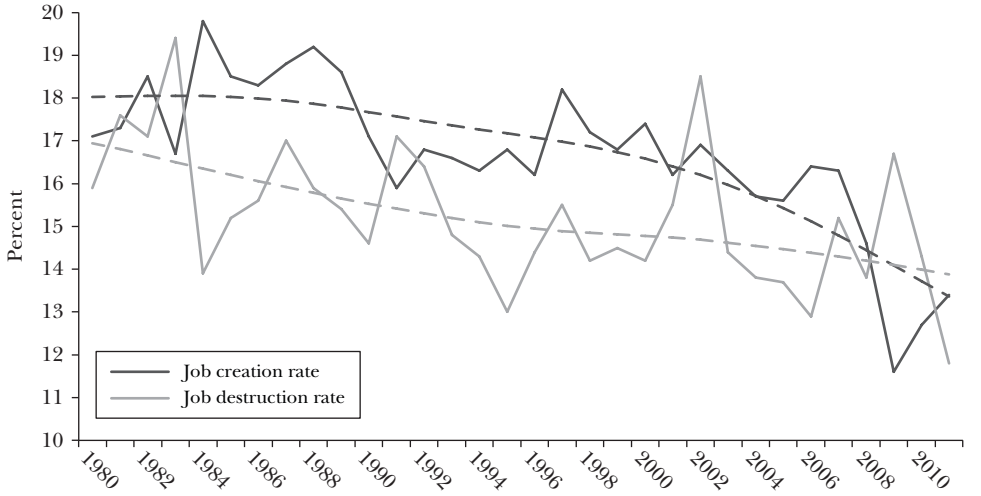
To sum up, while the evidence that startup firms are a powerful driver of job growth is quite clear, the evidence is less definitive for the contribution of startup firms to overall productivity growth. In addition to the limitations in relation to the studies discussed above, micro-based studies of productivity beyond the manufacturing and retail trade sectors are scarce. Measuring productivity in other industries is difficult, given the limitations on data on inputs other than labor in nonmanufacturing sectors. Still, the evidence that does exist for manufacturing and retail trade tends to support the existence of an up-or-out dynamic with high productivity (and high-profitability) young firms growing and low-productivity firms contracting and exiting. Moreover, young firms have a high innovation intensity, which suggests that young firms may be disproportionately important in terms of growth and productivity.

The Decline in the Startup Rate and Business Dynamism

The US economy displays a rapid pace of gross job creation and destruction, as depicted in Figure 3.[7] Even in 2009, when the economy experienced a large net employment decline due to the recession, expanding and new businesses added jobs at a rate equal to 12.4 percent of total employment. In absolute terms, the

[7] The online Appendix available with this paper at http://e-jep.org offers formal definitions of job creation and destruction and other related measures of dynamism in the job market. In this literature, "job reallocation" is the sum of job creation and destruction—it is a summary measure of all the changes in the location of jobs across producers. We also use that measure in the analysis in this section. Another summary measure often used in the literature is the "excess reallocation rate," which is job reallocation less the absolute value of net growth. The excess measure captures the reallocation over and above that needed to accommodate net growth. For the sake of brevity, we do not show patterns by excess reallocation here, but note that our patterns of declining trends carry over to excess reallocation. Davis et al. (2007) show patterns by excess reallocation.

Figure 3
US Annual Job Creation and Destruction Rates, 1980–2011



Source: Author calculations from the US Census Bureau's Business Dynamics Statistics.
Notes: The filter is Hodrick–Prescott with multiplier 400. The vertical axis does not begin at zero.

US private sector created more than 14 million new jobs between March 2008 and March 2009. The nature and pace of reallocation does change over the business cycle: specifically, job creation and job destruction tend to move in opposite directions during expansions and contractions. As Figure 3 shows, the decline in job creation during the 2007–2009 period was especially large.

But abstracting from the business cycle, the rates of job creation and destruction exhibit a general downward trend during the last few decades. Figure 3 illustrates this point by also showing smoothed trends. In particular, the job creation rate averaged 18.9 percent in the late 1980s and decreased in what appears to be a roughly stepwise pattern following recessions to an average of 15.8 percent in the 2004–2006 period just before the Great Recession. Meanwhile, the job destruction rate fell from an average of 16.1 percent in the late 1980s to an average of 13.4 percent in the 2004–2006 period. Moreover, the decline in the pace of job creation and destruction appears to have accelerated since about 2000.[8]

These declining trends appear in a wide range of data and settings. Using multiple measures of business dynamics from the Longitudinal Business Database, Davis, Haltiwanger, Jarmin, and Miranda (2007) found that the trend decline

[8] The annualized decline increases in magnitude from −0.15 to −0.22 of a percentage point for job creation and from −0.14 to −0.19 of a percentage point for job destruction.

is present in both firm-level and establishment-level measures.[9] Similarly, Davis, Faberman, Haltiwanger, Jarmin, and Miranda (2010), using the Bureau of Labor Statistics' Business Employment Dynamics data, show a declining pace of job flows over time. They find that the declining trend in the pace of job destruction is closely linked to the secular decline in the inflow rate to unemployment (at both the national and sectoral level). Davis, Faberman, and Haltiwanger (2012) show that this declining pace of job flows is matched by a declining pace of worker flows in the Job Openings and Labor Turnover (JOLTS) data. Specifically, the decline in the pace of hires exceeds that of the decline in job creation, and similarly the decline in the pace of separations exceeds that of the decline in job destruction. In other words, there is a decline in the pace of excess worker reallocation or worker churn. Lazear and Spletzer (2012) report similar findings using the JOLTS and BED data. Hyatt and Spletzer (2013) find similar patterns using the worker and job flows data from the Quarterly Workforce Indicators.

Job reallocation measures the change in the allocation of jobs across producers. In contrast, excess worker reallocation measures the change in the allocation of workers over a given set of jobs. Total worker reallocation is the sum of job reallocation and excess worker reallocation. It is striking that there have been secular declines in these related but distinct components of overall worker reallocation.

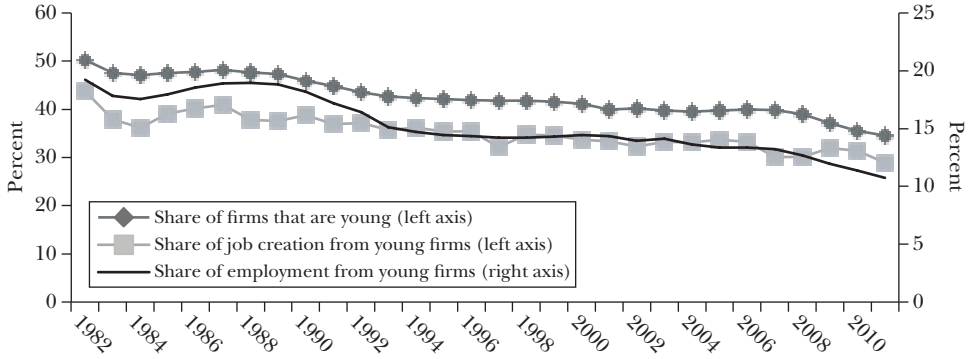
Secular Shifts: The Changing Distribution of Firm Age

A decline in the startup rate is one of the factors affecting the pace of reallocation in labor markets. The firm startup rate is measured by the number of new firms divided by the total number of firms. Our calculations based on the Business Dynamic Statistics data show that the annual startup rate declined from an average of 12.0 percent in the late 1980s to an average of 10.6 percent just before the Great Recession, when it plummeted below 8 percent. We also find that the startup rate has declined in all major sectors. We note, however, that in high-tech sectors (Haltiwanger, Hathaway, and Miranda 2014), the startup rate only began to decline in the post 2000 period.

Meanwhile, the average size of startups, as measured by employment, has either remained approximately the same over this time period as measured by the Census Bureau's Business Dynamics Statistics data (Haltiwanger, Jarmin, and Miranda 2013) or has declined as measured by the Bureau of Labor Statistics' Business Employment

[9] Davis et al. (2007) also show that the trend decline is present in within-firm and within-establishment measures of volatility as well as cross sectional dispersion measures such as the pace of job creation and destruction. The decline in the pace of overall firm volatility does mask an increase in the pace of firm volatility among publicly traded firms through 2000 (Comin and Philippon 2005). Davis et al. (2007) confirm the Comin and Philippon findings using data that have both privately held and publicly traded firms. They show that the decline in the pace of business volatility among privately held firms overwhelms the rise in firm volatility for publicly traded firms. Their findings suggest that the difference in patterns between publicly and privately held firms through 2000 primarily reflects a change in the composition of publicly held firms. In particular, the 1980s and 1990s cohorts of new publicly traded firms are younger when going public and also grow more rapidly after going public than earlier cohorts.

Figure 4
Declining Share of Activity from Young Firms (Firms Age 5 or Less)



Source: Author calculations from the US Census Bureau's Business Dynamics Statistics.
Note: Employment shares in each period based on the average of employment in period $t - 1$ and t (the denominator of the Davis, Haltiwanger, and Schuh (1996) growth rate).

Dynamics (Reedy and Litan 2011; Choi and Spletzer 2012). Either way, the lower startup rate is not being offset by a larger size of startup firms.

A consequence of the declining startup rate and flat or declining startup size is that the share of young firms in the economy, and the share of activity for which they account, is declining. Figure 4 shows that firms aged five years or less made up about 47 percent of all firms in the late 1980s, but this number declined to 39 percent of all firms before the start of the Great Recession, and has declined further since then. Similarly, the share of employment at firms less than five years of age declined from an average of 18.9 percent in the late 1980s to an average of 13.4 percent at the cyclical peak before the Great Recession. Finally, firms that were five years old or less contributed 39 percent of all new jobs in the late 1980s, but about 33 percent of all new jobs before the Great Recession. Because rates of job creation and destruction are much higher at young firms, these changes will contribute to the decline in both job creation and job destruction rates (Fort, Haltiwanger, Jarmin, and Miranda 2013).

To quantify the contribution of the changing age structure of firms on the change in job reallocation, we apply a standard shift-share decomposition for seven unique firm age groups (ages 0 through 5, and then 6 or more). To abstract from business cycle effects, we focus on the long-run change between two cyclical peaks. Specifically, we measure the difference between three-year averages of job flows for the periods of 1987–89 and 2004–06. Using a three-year average allows us to avoid issues that can arise in looking at particular years.

Our shift-share approach decomposes the change in job flows between 1987–89 and 2004–06 into three components: 1) the contribution from within-age group changes in job flows holding constant the employment-weighted firm age distribution at the initial levels; 2) the contribution from changes in the employment

distribution across firm age groups holding constant the job flows by firm age at the initial levels; and 3) the contribution of a cross term relating changes in shares with changes in flows. Doing this, we find that the shifting age composition of firms towards older businesses accounts for 32 percent of the observed decline in job creation, 20 percent of the decline in job destruction, and 26 percent of the decline in job reallocation.

Secular Shifts: The Changing Industrial Structure of Firm Activity

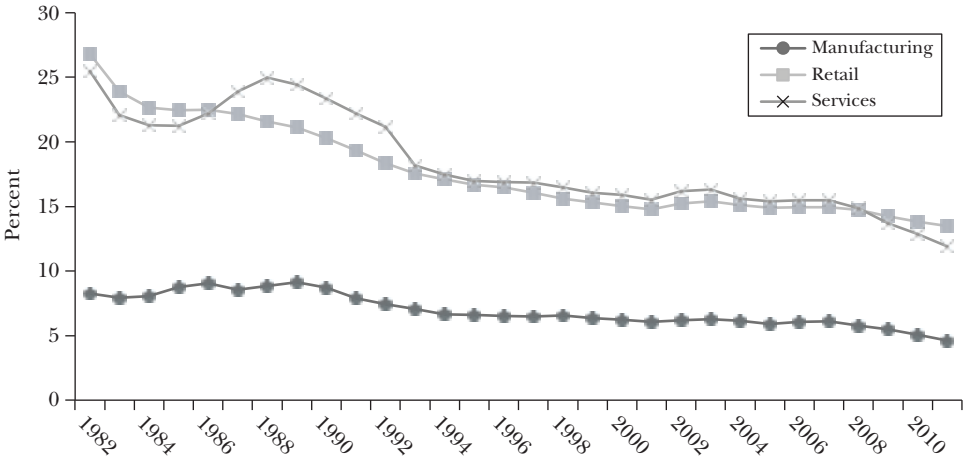
Unlike the changing age distribution of firms, shifts in the industrial composition of the economy would be expected to lead to increases in the pace of job reallocation. For example, the sectoral shift from manufacturing (relatively low pace of reallocation) to services (relatively high pace of reallocation) suggests that we should have expected an increase in the overall pace of business dynamics rather than a decline. In 1980, manufacturing accounted for 28 percent of all US jobs, while retail and services accounted for 24 percent. By 2011, manufacturing accounted for only 11 percent of all jobs, while retail and services accounted for 43 percent (using the broad sector definitions in the Business Dynamic Statistics).

To quantify the contribution shifts in the industry structure of the US economy have on job reallocation, we use the same type of shift-share analysis, but now look at 282 unique four-digit NAICS (2002) industries.[10] We find that the changing industrial share of firm activity indeed dampened the decline in the pace of job reallocation. If the industrial structure had remained the same from 1987–89 to 2004–06, the decline in job creation rates would have been greater by 21 percent, the decline in job destruction would have been greater by 3 percent, and the decline in job reallocation would have been greater by 12 percent.

Given the offsetting effect of the change in industrial composition, we considered a richer shift-share decomposition where we take into account the changes in industrial composition, the age structure, and other firm characteristics that have been associated with systematic differences in the pace of reallocation. In particular, a common finding is that establishments belonging to large firms are less volatile than those that belong to small firms. We consider a shift share decomposition with fully interacted cells defined by 282 unique four-digit NAICS (2002) industries, seven unique firm age groups (ages 0 through 5, and 6+), and eight firm-size groups (1–9, 10–19, 20–49, 50–99, 100–249, 250–499, 500–999, and 1000+ employees). In combination, the changing composition by industry, firm age, and firm size account for 9 percent of the decline in job creation, 18 percent of the decline in job destruction, and 14 percent of the decline in job reallocation. These findings suggest that unexplained within-group variation must account for most of the observed decline in job flows.

[10] We thank Teresa Fort for the development of a methodology that reclassifies all establishments in the Longitudinal Business Database to a consistent NAICS (2002) industry classification system. See Fort (2013) for details. Having a consistent classification system for our entire panel is critical for our analysis. These consistent NAICS codes have not yet been incorporated into the Business Dynamics Statistics.

Figure 5
Share of Employment from Young Firms (Firms Age 5 or Less), Selected Sectors



Source: Author calculations from the US Census Bureau's Business Dynamics Statistics.
Notes: Sector definitions are on an Standard Industrial Classification basis. Employment shares in each period are based on the average of employment in period $t-1$ and t (the denominator of the Davis, Haltiwanger, and Schuh (DHS) growth rate).

Looking for Clues in Sectoral and Geographic Changes in Startups

Insights into the unexplained within-group declines in job flows can be gained from exploring entrepreneurial activity within groups. Figure 5 shows the share of total employment for young firms (five years or less in age) for the retail, services, and manufacturing sectors. The share of employment in young firms has declined in all three sectors, suggesting that factors that are not sector-specific are causing the decline in entrepreneurial activity. Consistent with this view, we find this decline in the share of employment at young firms in every major sector. Figure 5 also shows that the drop is much larger for young retail and services firms than it is for young manufacturing firms. For whatever reason, incentives to start new businesses appear to be declining in all sectors, but disproportionately so in certain sectors such as retail trade, and this has contributed substantially to the declines in the pace of business dynamics.

The especially large decline in entrepreneurial activity in retail trade is consistent with the fundamental transformations ongoing in this sector over many decades. The expansion of "big box" retailers and, more generally, large national firms has dramatically changed the characteristics of the firms and establishments in the industry. Jarmin, Klimek, and Miranda (2005) report that the share of US retail activity accounted for by single-establishment ("mom-and-pop") firms fell from 70 percent in 1948 to 60 percent in 1967, and further still to 39 percent in 1997. If the decline in reallocation were confined to retail trade only, then we might guess that changing retail business models play a dominant role in accounting for

observed decreases in dynamism. But since we observe decreased dynamism in virtually all sectors, additional factors must be at work.

Additional clues might be found by exploring the changing pattern of entrepreneurial activity by US regions and states. We find that the share of employment by young firms has declined in all 50 states. The decline is more pronounced in states in the South and West compared to the declines in the Midwest and the Northeast. Such geographic variation can potentially capture different changes in the business climate, as states differ in regulations across a range of dimensions including occupational licensing requirements, banking regulations, tax burden for businesses and households, employment protection regulations, minimum wages, and others. However, we have found that, for example, states with business climates as different as California and Texas exhibit large and similar declines in entrepreneurial activity. More generally, since all states exhibit a decline in entrepreneurial activity, this suggests that factors other than state-specific business climate effects must be at work.

Understanding the Causes and Consequences of the Decline in Entrepreneurial Activity

We do not yet fully understand the causes of the decline in indicators of business dynamism and entrepreneurship, nor in turn, their consequences. Improving our understanding of the causes and consequences should be a high priority. A straightforward way of considering these issues is to ask whether either the structure of shocks affecting firms or changes in how firms are responding to these shocks can account for the findings. While both possibilities deserve inquiry in future research, we think the latter line of inquiry is likely to prove of greater value. Our prior is that the variance of idiosyncratic shocks affecting firms due to factors like technological change and globalization has, if anything, increased rather than decreased over time.[11] To the extent that such shocks have increased, then the decline in the pace of labor market reallocation is an even greater puzzle.

If the magnitude of economic shocks has not changed in a way that would tend to reduce the amount of labor market reallocation, then any such declines are driven by the way firms respond to shocks. One possibility is that the business climate, broadly defined, has changed in ways that impede job reallocation—that is to say, by impeding entry, exit, expansion, and contraction. Moreover, if the cause of the decline is an increase in the costs of adjustment on one or more of these margins, this can imply adverse consequences for growth, productivity, and welfare. The seminal work of Hopenhayn and Rogerson (1993) offers guidance here. They show that if an economy experiences an increase in adjustment costs for job destruction (for example, due to increased regulation), then not only will there be a decline in job destruction but also a decline in job creation (including a decline in startups)

[11] It is the trend in the pace of second moment shocks impacting firms that matters here and not so much the pace of first moment shocks. The period of the Great Moderation of business cycles is within our sample period, but that should have primarily impacted aggregate volatility and not firm-level volatility.

and ultimately in both productivity and welfare. The loss in welfare and productivity arises because the increase in adjustment frictions reduces the pace at which resources move away from less-productive to more-productive businesses. The same logic applies to changes in regulations or institutions that affect the costs of starting up or expanding a business, including regulations that raise the costs associated with expanding beyond some threshold of size.

A very different possibility is that firms are increasingly able to respond to shocks without as much churning of jobs and firms. After all, the churning of jobs and firms have no social value per se—it only has value to the extent that churning facilitates allocating outputs and inputs to their highest-valued use. Technological changes combined with globalization may have changed how businesses are organized and respond to shocks. For example, information and communications technology has arguably provided greater advantage for large, multinational firms in all sectors since these technologies can facilitate the coordination of production and distribution networks in multiple locations.[12] This change could help to explain the shift away from young firms to large, mature firms. In addition, it might help to explain the especially large declines in the share of startups and the decline in the pace of reallocation in specific sectors and geographic regions where information and communication technology has been especially relevant. Even with this more benign view, there might be a tradeoff between economies of scale induced by information and communications technology and flexibility in terms of how quickly the economy can adjust to changing economic conditions.

Yet another possibility is a hybrid of the first two explanations. Perhaps technological changes have induced changes in costs of hiring and training workers. For example, Cairó (2013) offers evidence that the training requirements of jobs have increased due to both the changing occupational mix of jobs and increases in training requirements within jobs. She develops a model where increased training requirements reduce the pace of job reallocation in the economy. She models this change in training requirements as an increase in adjustment costs, so that it is related to the first explanation above. However, it is an open question as to whether the increased training requirements of jobs is due to changes in technology, as Cairó hypothesizes, or to changing regulations or institutions, such as increased requirements for occupational licensing examined in Kleiner and Krueger (2013).

This discussion illustrates a few of many possible alternative causes, each with different consequences for trends in business dynamism. If the more sluggish pace of adjustment is due to increasingly burdensome regulation and institutions, this has potentially large adverse consequences for intermediate and long-run US job and productivity growth. Alternative explanations, like the notion of information and communications technology favoring large, multi-national firms suggest more benign consequences. Evaluating the productivity and welfare implications

[12] Evidence in support of the hypothesis that information and communications technology has favored large, multi-establishment firms for retail trade can be found in Doms, Jarmin, and Klimek (2004) and Foster, Haltiwanger, and Krizan (2006).

of the change toward lower levels of business dynamics and labor market reallocation will depend critically on the underlying causes of these changes.

Evidence about the role of startups from the rest of the world offers some useful perspective, too. Both the academic literature and international organizations such as the World Bank discuss the importance of entrepreneur-friendly business climates for economic growth; the 2013 *World Development Report* from the World Bank, focused on the theme of jobs, offers an excellent summary of the recent work in this area. But these issues are complicated and sometimes counterintuitive. For example, the evidence suggests that poorly performing emerging economies have plenty of entry—perhaps too much. Most startups in emerging economies are informal micro-enterprises with few if any paid employees. A high startup rate has little value per se. Instead, the problem in many emerging economies is that there is little or no evidence of post-entry growth (Hsieh and Klenow 2012). Using Schoar's (2010) characterization, poorly performing economies seem to have too many subsistence entrepreneurs and too few high-growth transformational entrepreneurs. Translating this issue back to the US context, the open question is whether the observed decline in startups and the associated decline in indicators of business dynamism imply fewer transformational entrepreneurs in the US economy.

Concluding Remarks

Startups and young firms are important contributors to job creation and productivity growth. The patterns for the US economy are roughly consistent with canonical heterogeneous firm models that place an important role for the reallocation of resources away from low-productivity (or low-profitability) businesses to high-productivity businesses. Entry and exit are important components of that ongoing reallocation. However, the contribution of startups and young businesses to jobs and productivity is a noisy and complex process. While startups contribute substantially to jobs immediately, most startups fail or, even if they survive, do not grow, while a small fraction of high-growth young firms contribute disproportionately to job creation in the United States. These findings pose challenges to policymakers seeking to promote job creation by encouraging entrepreneurship, because most young and small businesses are not in fact primary creators of jobs.

The rate of business startups in the US economy has been declining in recent decades, and business dynamism, as measured by the pace of job creation and job destruction, has declined as well. We do not have an explanation for the decline in the pace of entrepreneurship. Some of the structural changes we have observed in the economy such as the shift in the industrial composition actually increase the puzzle rather than provide an explanation since the shift has been towards sectors with higher paces of reallocation historically. The decline in entrepreneurial activity is present in all broad industrial sectors and in all US states. This finding suggests there are some common factors that are not sector- or region-specific accounting for the decline in entrepreneurial activity.

The decline in startups and accompanying decline in the pace of reallocation is a legitimate concern, although the causes of the decline and whether it is having or will have adverse consequences remain open questions for research. In considering these issues, it is worth noting that the declining pace of startups, job creation, and job destruction is mirrored in other measures of the dynamism of American society. In this journal in 2011, Molloy, Smith, and Wozniak noted that internal migration has been declining since a peak in the early 1980s. Like us, they failed to find simple explanations for these trends in changing demographics or broad economic factors. In a related fashion, the churn of workers over and above the job reallocation that we have emphasized here has exhibited a pronounced decline. The building of a career path and finding a good match has traditionally involved a high pace of job switching for workers (especially young workers). The evidence on worker churning implies that the process of building careers through job switching has slowed down. Taken together, there appears to be less scope for the US economy to adjust to changing economic conditions through the migration of workers, the reallocation of jobs across producers, and through the switching of workers across a given allocation of jobs.

■ *We thank the Kauffman Foundation for financial support, Kristin McCue and Lucia Foster for helpful comments, and Sinan Hastorun for excellent research assistance. We also thank the editors of this journal for many useful comments. Haltiwanger notes that the discussion in this paper draws on joint work with Steven Davis, Marcela Eslava, Jason Faberman, Teresa Fort, Lucia Foster, Ian Hathaway, C. J. Krizan, Adriana Kugler, Maurice Kugler, Scott Schuh, and Chad Syverson. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the US Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.*

References

Acemoglu, Daron, Ufuk Akcigit, Nicholas Bloom, and William R. Kerr. 2013. "Innovation, Reallocation, and Growth." NBER Working Paper 18993.

Bartelsman, Eric J., and Mark Doms. 2000. "Understanding Productivity: Lessons from Longitudinal Microdata." *Journal of Economic Literature* 38(3): 569–94.

Cairó, Isabel. 2013. "The Slowdown in Business Employment Dynamics: The Role of Changing Skill Demands." At http://www.econ.upf.edu/eng/graduates/gpem/jm/cairo.html.

Choi, Eleonor J., and James R. Spletzer. 2012. "The Declining Average Size of Establishments: Evidence and Explanations." *Monthly Labor Review*, March, pp. 50–65.

Comin, Diego A., and Thomas Philippon. 2005. "The Rise in Firm-Level Volatility: Causes and

Consequences." Chap. 3 in *NBER Macroeconomics Annual 2005,* edited by Mark Gertler and Kenneth Rogoff. MIT Press.

Davis, Steven J., R. Jason Faberman, and John Haltiwanger. 2012. "Labor Market Flows in the Cross Section and Over Time." *Journal of Monetary Economics* 59(1): 1–18.

Davis, Steven J., R. Jason Faberman, John Haltiwanger, Ron Jarmin, and Javier Miranda. 2010. "Business Volatility, Job Destruction, and Unemployment." *American Economic Journal: Macroeconomics* 2(2): 259–87.

Davis, Steven J., John Haltiwanger, Ron Jarmin, and Javier Miranda. 2007. "Volatility and Dispersion in Business Growth Rates: Publicly Traded versus Privately Held Firms." Chap. 2 in *NBER Macroeconomics Annual 2006* edited by Daron Acemoglu, Kenneth Rogoff, and Michael Woodford. Cambridge, MA: MIT Press.

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

Doms, Mark, Ron S. Jarmin, and Shawn Klimek. 2004. "Information Technology Investment and Firm Performance in U.S. Retail Trade." *Economics of Innovation and New Technology* 13(7): 595–613.

Drozd, Lukasz A., and Jaromir B. Nosal. 2012. "Understanding International Prices: Customers as Capital." *American Economic Review* 102(1): 364–95.

Eslava, Marcela, John Haltiwanger, Adriana Kugler, and Maurice Kugler. 2004. "The Effects of Structural Reforms on Productivity and Profitability Enhancing Reallocation: Evidence from Colombia." *Journal of Development Economics* 75(2): 333–72.

Eslava, Marcela, John Haltiwanger, Adriana Kugler, and Maurice Kugler. 2013. "Trade and Market Selection: Evidence from Manufacturing Plants in Colombia." *Review of Economic Dynamics* 16(1): 135–58.

Fort, Teresa. 2013. "NAICS 2002 Code Assignments to the Longitudinal Business Database." Unpublished paper, Center for Economic Studies.

Fort, Teresa C, John Haltiwanger, Ron S. Jarmin, and Javier Miranda. 2013. "How Firms Respond to Business Cycles: The Role of the Firm Age and Firm Size." *IMF Economic Review* 61(3): 520–59.

Foster, Lucia, Cheryl Grim, and John Haltiwanger. 2013. "Reallocation in the Great Recession: Cleansing or Not?" CES Working Paper 13-42, Center for Economic Studies, US Census Bureau.

Foster, Lucia, John Haltiwanger, and C. J. Krizan. 2001. "Aggregate Productivity Growth: Lessons from Microeconomic Evidence." In *New Developments in Productivity Analysis* (NBER Book Series Studies in Income and Wealth), edited by Charles R. Hulten, Edwin R. Dean, and Michael J. Harper, pp. 303–72. Chicago, IL: National Bureau of Economic Research, University of Chicago Press.

Foster, Lucia, John Haltiwanger, and C. J. Krizan. 2006. "Market Selection, Reallocation, and Restructuring in the US Retail Trade Sector in the 1990s." *Review of Economics and Statistics* 88(4): 748–58.

Foster, Lucia, John Haltiwanger, and Chad Syverson. 2008. "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review* 98(1): 394–425.

Haltiwanger, John. 2012. "Job Creation and Firm Dynamics in the U.S." In *Innovation Policy and the Economy,* edited by Josh Lerner and Scott Stern, pp. 17–38. NBER/Chicago Press.

Haltiwanger, John, Ian Hathaway, and Javier Miranda. 2014. "Declining Business Dynamism in the U.S. High-Technology Sector." The Kauffman Foundation.

Haltiwanger, John, Ron Jarmin, and Javier Miranda. 2009. "Business Dynamics Statistics: An Overview." Kauffman Foundation: Other Research. Available at SSRN: http://ssrn.com/abstract=1456465.

Haltiwanger, John, Ron Jarmin, and Javier Miranda. 2011. "Historically Large Decline in Job Creation from Startups and Existing Firms in the 2008–09 Recession." Business Dynamics Statistics Briefing No. 5, The Kauffman Foundation. Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1793167.

Haltiwanger, John, Ron S. Jarmin, and Javier Miranda. 2013. "Who Creates Jobs? Small vs. Large vs. Young." *Review of Economics and Statistics* 95(2): 347–61.

Hopenhayn, Hugo, and Richard Rogerson. 1993. "Job Turnover and Policy Evaluation: A General Equilibrium Analysis." *Journal of Political Economy* 101(5): 915–38.

Hsieh, Chang-Tai, and Peter J. Klenow. 2012. "The Life Cycle of Plants in India and Mexico." NBER Working Paper 18133.

Hurst, Erik, and Benjamin Wild Pugsley. 2011. "What Do Small Businesses Do?" *Brookings Papers on Economic Activity,* no. 2, pp. 73–142.

Hyatt, Henry R., and James R. Spletzer. 2013. "The Recent Decline in Employment Dynamics." *IZA Journal of Labor Economics* 2(5): 1–21.

Jarmin, Ronald S., Shawn Klimek, and Javier Miranda. 2005. "The Role of Retail Chains: National, Regional and Industry Results." CES Working Paper 05-30, Center for Economic Studies, US Census Bureau.

Kleiner, Morris M., and Alan B. Krueger. 2013. "Analyzing the Extent and Influence of

Occupational Licensing on the Labor Market." *Journal of Labor Economics* 31(2, Part 2): S173–S202.

Lazear, Edward P., and James R. Spletzer. 2012. "Hiring, Churn, and the Business Cycle." *American Economic Review* 102(3): 575–79.

Molloy, Raven, Christopher L. Smith, and Abigail Wozniak. 2011. "Internal Migration in the United States." *Journal of Economic Perspectives* 25(3): 173–96.

Reedy, E. J., and Robert E. Litan. 2011. "Starting Smaller; Staying Smaller: America's Slow Leak in Job Creation." Kauffman Foundation Research Series: Firm Foundation and Economic Growth. At http://www.kauffman.org/research-and-policy/starting-smaller-staying-smaller-americas-slow-leak-in-job-creation.aspx.

Schoar, Antoinette. 2010. "The Divide between Subsistence and Transformational Entrepreneurship." In *Innovation Policy and the Economy*, vol. 10, edited by Joshua Lerner and Scott Stern, pp. 57–81. National Bureau of Economic Research.

Syverson, Chad. 2004. "Product Substitutability and Productivity Dispersion." *Review of Economics and Statistics* 86(2): 534–50.

Syverson, Chad. 2011. "What Determines Productivity?" *Journal of Economic Literature* 49(2): 326–65.

Törnqvist, Leo, Pentti Vartia, and Yrjö O. Vartia. 1985. "How Should Relative Changes Be Measured?" *American Statistician* 39(1): 43–6.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 25–48

Entrepreneurship as Experimentation[†]

William R. Kerr, Ramana Nanda, and Matthew Rhodes-Kropf

Although the fundamental importance of entrepreneurship has been generally recognized by economists since at least Schumpeter (1911 [1934]; 1943), it has not been the subject of substantial research and is more-or-less absent from standard textbooks. Over the last decade, however, entrepreneurship has flourished as a research topic within economics. As entrepreneurship emerges from the shadows, many central and unresolved questions linger: How should we define entrepreneurship? What are its key aspects? Is entrepreneurship about the next Skype or the next self-employed accountant? We assert in this paper that effective entrepreneurship—especially among high-growth ventures and for the economy as a whole—builds upon a process of experimentation in deep and nuanced ways.

Entrepreneurship is fundamentally about experimentation because the knowledge required to be successful cannot be known in advance or deduced from some set of first principles. As Hayek (1948) put it, "the solution of the economic problem of society is . . . always a voyage of exploration into the unknown." For entrepreneurs, it can be virtually impossible to know whether a particular technology or product or business model will be successful, until one has actually invested in it.

Two interesting examples help frame the discussion. In 1999, the venture capital firms Sequoia Capital and Kleiner Perkins Caufield & Byers each invested $12.5 million in Google, a brand new startup that claimed to have a superior search engine. By the time Sequoia sold its stake in 2005, that investment was worth over

■ *William Kerr, Ramana Nanda, and Matthew Rhodes-Kropf are Associate Professors at Harvard Business School, Harvard University, Boston, Massachusetts, and Faculty Research Fellows at the National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are wkerr@hbs.edu, RNanda@hbs.edu, and mrhodeskropf@hbs.edu.*

[†] To access the data Appendix, visit
http://dx.doi.org/10.1257/jep.28.3.25

doi=10.1257/jep.28.3.25

$4 billion, returning 320 times the initial cost (Sahlman 2010). The staggering success should not obscure a key point—Google was by no means a sure-shot investment in 1999. The search algorithm space was already crowded and dominated by other players such as Yahoo! and Altavista. Google, founded by two students, might well have turned out to just be a "me too" investment. In fact, other venture capitalists had turned down the opportunity to invest in Google at the time. One such firm was Bessemer Ventures. Partner David Cowan, on being asked by a friend to meet with the two Google founders who had rented space in her garage, is believed to have quipped, "How can I get out of this house without going anywhere near your garage?" (Bessemer Venture Partners, no date). Over the years, Bessemer has been so successful that they playfully mock at their "anti-portfolio" of deals that they passed on (at http://www.bvp.com/portfolio/antiportfolio).

Not many investors are as candid about great opportunities that they have passed on, but more recently, Fred Wilson of Union Square Ventures has similarly written about his regret at passing on Airbnb, a startup that lets people rent rooms or homes to prospective guests online. Started in 2008, Airbnb currently lists 500,000 properties in more than 34,000 cities across 192 countries, far more than any hotel chain—thereby making it the largest lodging company and brand in the world (Levere 2013). Wilson (2011) explains Union Square Venture's decision to pass on the deal in 2009: "we couldn't wrap our heads around air mattresses on the living room floors as the next hotel room. . . . Others saw the amazing team that we saw, funded them, and the rest is history."

These two examples highlight several challenges associated with commercialization of new ideas, products, and technologies. First, the actual distribution of returns in such ventures has a low median value but very high variance (Scherer and Harhoff 2000; Hall and Woodward 2010). Most new ventures fail badly, but some turn out to be wildly successful. Second, even for professional investors or managers making resource allocation decisions, it is impossible to know in advance which ideas will work. As we describe in greater detail below, venture capital investors make their returns on the one investment out of many that turns out to be a wild success like Google or Airbnb. The vast majority of venture capital investments, however, return less than the face value of the investment.

How entrepreneurs and investors respond to these inherent challenges has important implications for their own success and also for the broader economy in terms of the "best ideas" being commercialized. To flesh out this point, it is important to separate two frames of reference regarding experimentation. The first relates to economic experimentation in a Darwinian sense, which is the natural starting point for most economists. In this conceptual model, new ventures compete with existing products and technologies, and the ensuing competition leads to the survival of the fittest, just as Google surpassed its early rivals due to its superior technology. This competition can be described as experimentation at the level of the economy. In settings where the best approach among several options is unknown, a great benefit of market-based economies is that winners are often chosen by consumers and competition. As Rosenberg (1994) has argued, one of the defining features

Table 1
US Venture Capital Statistics for 2010

Number of active venture capital firms	462
Amount invested by venture capital firms	$22 billion
Number of companies receiving venture capital investment	2,749
Number of companies receiving funding for first time	1,001
Share of dollars invested in Information Technology firms	51%
Share of dollars invested in Life Science firms	27%

Source: Data from National Venture Capital Association, nvac.org.

of capitalism is the freedom it provides entrepreneurs to pursue novel approaches to value creation in the pursuit of economic gain. The promise of large rewards drives entrepreneurs to experiment with new ideas, helping to create a dynamic and growing economy. An institutional environment that facilitates experimentation is thus central to maintaining a vibrant entrepreneurial ecosystem (Dosi and Nelson 2010). This experimentation, combined with a willingness to let losing incumbents fail, is the underlying notion behind Schumpeter's (1943) process of "creative destruction."

While this Darwinian depiction is important and a natural starting point, a second reference point emphasizes that we should be cautious of assuming that market-based mechanisms can always serve as a guiding hand with respect to experimentation. As in the Google example, it is often very difficult to know whether a particular technology or venture will succeed until one has made an investment. Moreover, the investments to obtain this vital information can be nontrivial: in Google's case, a total of $25 million was invested in the first round. Table 1 documents some basic facts about the venture capital industry, including the number of active investors. Given the relatively small number of financiers that invest in this sector, it is easy to imagine scenarios where Google or other highly productive investments fail to receive the required funding.

In an entrepreneurial setting, where the benefit of pursuing different approaches is not clear and the costs of tests are expensive, each individual endeavor is also engaged in a process of experimentation. As these experiments provide information about the likelihood of ultimate success, entrepreneurs and investors gain information about whether to continue the project. However, the investment and continuation decisions for entrepreneurs are often not made in a competitive Darwinian contest, because the decisions to invest further or shut down a firm are often made by only a few investors, well before startups can compete in the product market or have positive cash flow. Moreover, the decisions are made by discrete individuals, often in venture capital firms or other early-stage financing vehicles, whose actions are impacted by a myriad of incentive, agency, and coordination problems. Thus, the extent to which the best idea goes forward may depend on factors such as the organizational structure or incentive system of the firm where the investor is based, available information sets (for example, access to certain networks),

coordination costs, and other such frictions. Taken together, these factors affect how much experimentation is undertaken in the economy and also the trajectory of experimentation, with potentially very deep economic consequences.

The lens of experimentation has implications for what types of innovations will occur, who will pursue them and when. As Stern (2005) argued, "a favorable environment for entrepreneurship and a high level of economic experimentation go hand in hand." Although experiments can be conducted in large companies or in the public sector, new technologies and innovative products are often commercialized by entrepreneurs and often cluster at particular times. We argue that the costs and constraints on the ability to experiment alter the type of organizational form surrounding innovation and influence when innovation is more likely to occur.

This article considers these costs and constraints to experimentation by investors and the implications that these issues have on the type of entrepreneurship across time and across economies. We start with a micro-level examination of experimentation and focus on the processes used by investors to experiment and the issues that individual firms face. We then turn to the macro perspective of the economy and the policies that shape entrepreneurship and its effectiveness.

A Closer Look at the Process of Experimentation

High-impact entrepreneurship requires, almost by definition, going against the grain. Rajan (2012) argues that an entrepreneur "must be willing to strike out, largely on the basis of intuition, on courses of action in direct opposition to the established settled patterns." A consequence of this setting is extreme uncertainty about whether a particular technology, product, or business model will be successful. This uncertainty is fundamentally different from risk, as Knight (1921) stressed. With risk, such as placing money on a spin of a roulette wheel, one can define exact probabilities and expected values. With Knightian uncertainty, these probabilities are not known, and even the form of the potential outcomes may be unclear. For example, looking forward from today, at what rate will electric cars, if at all, replace traditional automobiles and how will the supporting infrastructure for battery recharging be designed? What will be the impact of nascent augmented reality technologies for how humans interact? Which, if any, of the several current ideas to cure cancer will be successfully commercialized?

In this environment of tremendous uncertainty, experimentation allows entrepreneurs and investors to assess and commercialize projects without investing the full amount. Crucially, experimentation offers more than just a possibility of higher returns—it also allows entrepreneurs and investors to pursue projects that are not feasible in an all-or-nothing bet. For example, consider a project that requires $110 to commercialize and will be worth $0 with 99 percent probability or $10,000 with 1 percent probability. This project will not be pursued, because its expected value is negative (−$10). But imagine we can conduct an experiment that will reveal if the project has a 10 percent chance of working. (Suppose further that the probability

of the experiment having a positive outcome is only 10 percent, so that in this example the 1 percent chance of overall success is unchanged.) If the experiment gives a positive signal, the project has a larger expected value of $890. Thus, as long as the experiment costs less than $89 (10 percent × $890), the experiment should be conducted, with the project then being either shut down or commercialized based upon the results.

This simple example provides insights into many features of the entrepreneurial landscape, as we elaborate below. The test resolves uncertainty about the project's potential and creates a real option value for further pursuit. Thus, experimentation is particularly valuable in cases where initial information can be especially informative about the overall quality of the project and this information is cost effective to obtain.[1]

The example also illustrates that the costs of experimenting are important. Our simple example was phrased in terms of the direct costs of conducting the test. More broadly, important indirect costs must be borne by either the investor or entrepreneur when they fail. Even if direct costs are small, significant indirect costs like a stigma of failure (Landier 2005) may still prevent entrepreneurs from pursuing otherwise valuable tests if they only have a 10 percent chance to succeed. Moreover, the ability to document and transmit the results of experiments is important, because the information from the experiment is only valuable if it can be acted upon. Can an investor use the information to decide to provide the follow-on investment required to realize the project? The framework also begins to highlight the distinct roles of the entrepreneur and the investor.

We trace these features out next by first looking more closely at the economics of a venture capital firm. This setting provides a powerful archetype of high-impact entrepreneurship more broadly and the role of experimentation. We then discuss frictions to experimentation that are observed in the economy and the extent to which agents are able to address them.

The Economics of Experimentation in a Venture Capital Firm

The economics and structure of a venture capital firm provide insights into entrepreneurship as experimentation (Sahlman 1990). A crucial starting point is the recognition that venture capital firms do not simply act as a portfolio of risky startups. While the portfolio nature of a venture capital firm is important, it is quite distinct from what is observed in stock market mutual funds, for example, where the portfolio is designed to smooth out the idiosyncrasies of individual stocks. Venture capital firms are better thought of as conducting a portfolio of tests across a number of highly uncertain ideas with skewed economics (similar to the example described above). In most cases, the tests come out negative. In fact, the majority of venture-capital-backed firms fail, despite the investments made into them by highly skilled investors. Once venture capital firms identify the

[1] Weitzman (1979) introduces and analyzes the optimal search behaviors under these uncertainties.

cases that deliver positive test results, however, they pursue them aggressively and invest ever larger amounts on the promising candidates to scale them up. Whereas mutual fund investors do not need to be an early investor in a company in order to be allowed to invest more later, this is precisely the way that venture capital firms aim to make money—by starting small and owning a larger share of the firms that turn out to be successful, while attempting to cut their losses on the unsuccessful ones as early as possible.[2]

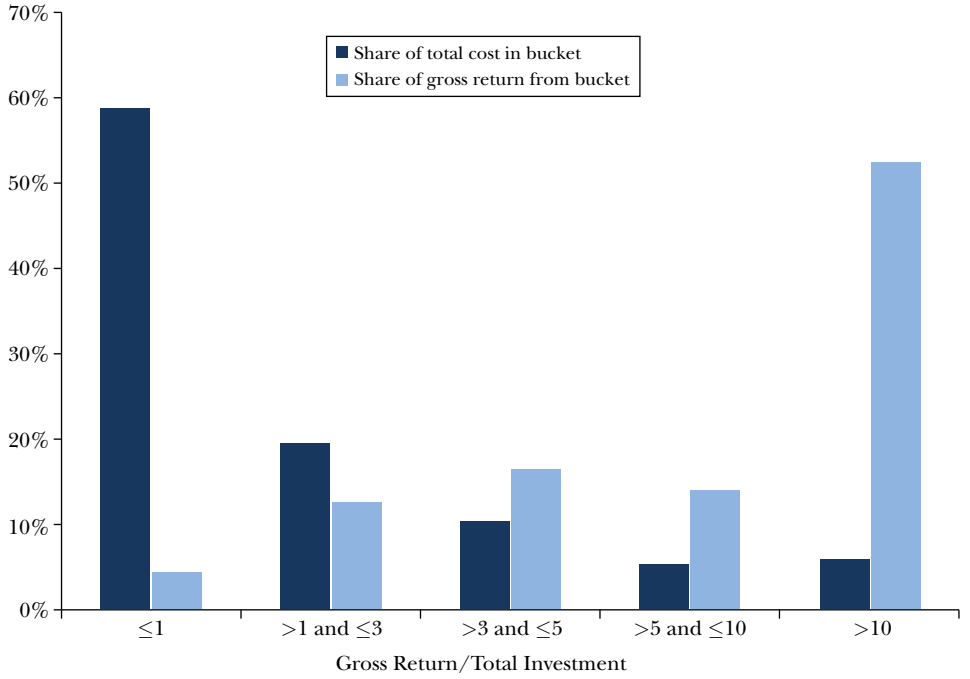
Some simple calculations document these skewed returns using data on all venture capital investments in the United States. Using data from Thompson Venture Economics, we identify all startup ventures that received their first round of early-stage financing for the years between 1985 and 2009. We calculate both the total investment made by venture capital investors in each of these firms, as well as the gross return at the time of exit—either through an acquisition deal, an initial public offering, or bankruptcy. We then group each of these startups based on their economic return to investors at the time of exit—that is, the ratio of the gross return at the point of exit divided by the total amount invested in the startup. About 55 percent of startups that received venture capital over this period were terminated at a loss, and only 6 percent of them returned more than five times their investment. This 6 percent group, however, was extremely successful and together accounted for about 50 percent of the gross return generated over the period.[3] The fact that the returns are so skewed across the portfolio is *prima facie* evidence that the (quite knowledgeable) investors cannot distinguish in advance the next Google from the other cases.

A second example provides more systematic evidence on the difficulty of predicting outcomes, even conditional on a venture capital firm making an investment. For a single large and successful venture capital firm that has invested more than $1 billion over the last decade, we have access to both the outcomes of individual investments and the scores that partners assigned to each venture at the time of their first investment. We place the investments of this firm into buckets according

[2] Gompers and Lerner (2004) and Da Rin, Hellman, and Puri (2013) provide a comprehensive overview of these intermediaries. Chemmanur, Krishnan, and Nandy (2011) and Puri and Zarutskie (2012) provide recent evidence on the performance of venture-capital-backed startups. Parker (2004) provides a broader review of the literature on entrepreneurship.

[3] Total gross returns need to be estimated due to the fact that a venture's value at exit is missing for a large number of acquisitions, for firms that went bankrupt, and for the "living dead"—firms that are coded as still private but have not received follow-on financing for several years. Our research suggests that most of the acquisitions without data are fire sales and that firms that are coded as "alive" but have not received a follow-on round of financing within three years of the last financing are likely to be bankrupt. In the calculation reported above, we have conservatively assumed that unreported acquisitions were undertaken at 1.5 times cost and that firms that did not receive subsequent rounds of financing for three years, but were coded as still private, were liquidated at 0.25 times cost. The results are robust to assuming a wide range of possible values for these outcomes, including an (implausible) average of two times the dollars invested in the firm. These numbers correspond closely with numbers from Sahlman (2010) who uses data from eleven early-stage venture capital firms to show that 64 percent of their investments were terminated at a loss, while 8 percent of the investments, those that returned over five times the investment, generated about 60 percent of the gross return of the portfolio.

Figure 1
Total Cost and Total Return for a Venture Capital Firm



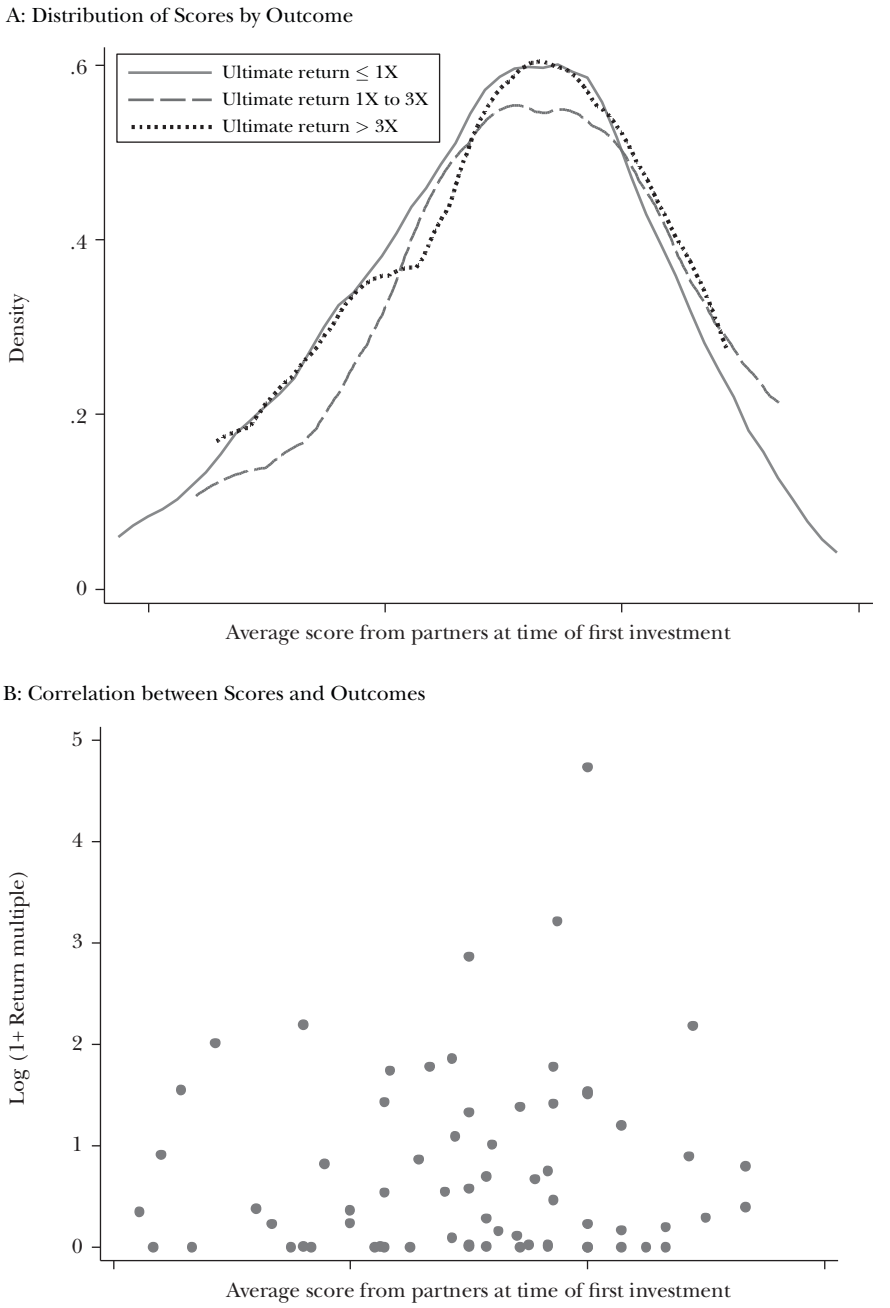
Source: Authors using proprietary data from an anonymous venture capital firm that has invested more than $1 billion over the last decade.
Notes: Investments are placed into buckets according to their return multiple. For each bucket, we show the total cost and total return for that group of investments.

to their return multiples, and in Figure 1, show the total cost and total return for each bucket of investments. The distribution of outcomes for this firm follows the same pattern as the aggregate data on venture capital (with about 60 percent of investments terminated at a loss and 10 percent generating a return more than five times the capital invested).

The firm has an excellent track record in the industry, which implies a well-functioning algorithm for choosing and managing investments. Conditional on making an investment, however, Figures 2A and 2B show the limited ability of the venture capital investors to tell the most successful investments from the less successful ones—even for these experienced investors. Figure 2A highlights that the distribution of scores assigned to investments that ended up performing extremely well is statistically no different from the distribution of scores assigned to investments that ultimately failed or had mediocre returns.[4] (The labels on the horizontal axis

[4] The Kolmogorov–Smirnov test for equality of distribution functions could not reject the hypothesis that these were the same.

Figure 2
Scores Assigned to Investments at Time of First Investment and the Ultimate Returns of Those Investments, for One Venture Capital Firm



Note: The labels on the horizontal axis have been suppressed to maintain the confidentiality of the investor's rating scale, but lower predictions were on the left and higher predictions were on the right.

have been suppressed to maintain the confidentiality of the investor's rating scale, but lower predictions were on the left and higher predictions were on the right.) Figure 2B is a scatterplot showing, for each of the firm's investments, the average score assigned at the time of first investment on the x-axis versus a function of the return multiple on the y-axis. The raw correlation between scores and the return multiple from the investments is 0.1.

This example shows that even conditional on making an investment, it is hard for the investors to predict which firm will be successful. Using a similar framework from a sample of early-stage "angel investors" in Kerr, Lerner, and Schoar (2014), the correlation among the interest levels assigned to funded deals and their ultimate success is less than 0.1. Again, these examples help illustrate the high degree of uncertainty present in these investments.

Thus, venture capital firms need to experiment, and many aspects of their business model facilitate or emphasize this experimentation. One example is their focus on sectors that are capital-efficient for both experimentation and subsequent scaling and that can generate large returns for the successful investments in a short period of time. Under these conditions, most notably associated with information technology investments, venture capital firms can run initial experiments of manageable financial sizes and then fund the winners to completion. The corollary to this, which is further elaborated upon below, is that venture capital activity is concentrated in a narrow range of technological opportunities. Some sectors, like renewable energy production, need to be proven at large scale to demonstrate technical feasibility and unit economics. Commercializing such ventures requires building large manufacturing plants and hence is significantly more capital intensive, and takes much longer. Following a brief period where venture capitalists invested heavily in biofuel and solar technologies only to learn these lessons the hard way, they have largely shied away from funding renewable energy production startups, instead devoting their attention within clean energy to startups commercializing energy efficiency, smart grid, and other software technologies (Nanda, Younge, and Fleming, forthcoming). Fleming (2001) and Fleming and Sorenson (2004) discuss the aggregate implications of a narrower range of search or recombination.

In addition to focusing on specific sectors, venture capital structures and contractual choices also address experimental challenges. For example, in a way that is similar to our earlier numerical example, venture capital investors provide staged financing to startup companies that tie each financial infusion to milestones—points at which information is revealed about the quality of the project. This structured financing builds real options (for example, Gompers 1995; Bergemann and Hege 2005; Bergemann, Hege, and Peng 2008), by matching the amount of money raised in each round to the specific uncertainty that needs to be resolved with that round of funding: for example, proof that the technology works, that consumers will buy the product, and so on. The most successful investors and entrepreneurs are able to identify the most important uncertainties facing a new idea and experiment in a way that resolves the greatest proportion of the uncertainty around them effectively and quickly.

The consequence of this staging approach is that an important role of venture capital firms is to shut down ventures for which they are receiving bad signals. Investors exercise the abandonment options in several ways. Across financing rounds, the investors can choose to not reinvest in the company; within each financing round, venture capital firms also structure contracts to provide them with control and cash flow rights. Control rights allow venture capital firms to replace underperforming chief executive officers, and the investors can also push the company to reposition itself if the chosen path turns sour. These rights appear to be exercised often (for example, Hellman and Puri 2000; Kaplan, Sensoy, and Strömberg 2009; Tian 2011).[5]

The available evidence suggests that this process of venture capital experimentation does lead to successful economic outcomes. Kortum and Lerner (2000) and Samila and Sorenson (2011) document spillover effects from venture capital activity within industries and cities, respectively. Using Census Bureau data, we provide descriptive evidence of how this sort of experimentation is associated with outcomes for ventures. We use name and address matching to identify new entrants that received venture capital investments during the 1986–1997 period. Looking ahead to 2007, 75 percent of these venture-capital-backed ventures have shut down, but the remaining firms have grown to a level where they combine to equal 364 percent of the total employment of the original investments (including those that failed). For comparison, we generated a second sample of non-venture-capital-backed entrants that have the same four-digit industry, year, and starting employment level as the investment sample. In 2007, a greater share of this second group remains alive—only 66 percent have shut down. However, those that are alive only account for two-thirds (67 percent) of the total employment of their original sample. These matched ventures have thus grown, but not nearly to the extent of the venture-capital-backed group. While our simple matching procedure may leave residual differences across the groups and also does not account for either the role of selection or the effect of venture capital on startups' outcomes, it should not obscure the broad overall point: the venture-capital-backed ventures experienced a higher if similar rate of failure (75 versus 66 percent) but also experienced far greater employment gains (364 versus 67 percent of employment in their samples). High-growth ventures typically constitute a small proportion of new startups but are responsible for a disproportionate share of employment and GDP. As a rough benchmark, about 1,000 of the over 500,000 firms founded each year in the United States obtain venture capital finance, but these firms account for about 40 percent of new US publicly-listed companies over the last three decades (Ritter 2014).

[5] Many other aspects of the venture capital model reflect these cash-flow and control rights (for example, Kaplan and Strömberg 2003, 2004). One prominent example is convertible preferred stock that protects venture capital firms in bad outcomes by allowing them to redeem the full value of their investment, thereby treating the investment as debt to the extent that the failed venture has any residual value. On the other hand, in positive outcomes, the venture capital firm can choose to convert the preferred to common stock, allowing them to share in the upside when things go well.

Frictions for the Process of Experimentation—Costs of Experiments

The costs of running experiments play a big role in entrepreneurship. Technological change in the last decade has dramatically lowered these costs, particularly in industries that have benefited from the emergence of the Internet, including trends like open-source software and cloud computing. Industry observers suggest that firms in these sectors that would have cost $5 million to set up a decade ago can be done for under $50,000 today. For example, open-source software lowers the costs associated with hiring programmers. In addition, fixed investments in high-quality infrastructure, servers, and other hardware are no longer necessary at the birth of many software firms because they can be rented in tiny increments from cloud computing providers and efficiently scaled up as demand for their products increases (Blacharski 2013; Palmer 2012). This reduced entry barrier has led to an explosion of experimentation with new entrepreneurial ideas in this area. Funding sources have also proliferated as the ability to finance these smaller and cheaper experiments is possible for individuals and groups beyond the traditional venture capital investors with large pools of capital.

One prominent example of this shift is the Lean Startup methodology advocated by serial entrepreneur Steve Blank and popularized by Eric Ries's (2011) book, "The Lean Startup." Many startup founders have shaped their efforts using the Lean Startup approach, and the ideas have reached the leadership of large companies like Jeff Immelt, the chief executive officer of General Electric. The core of the management style is a focus on identifying and developing "minimal viable products" (MVPs) that reveal how well the overall opportunity will fare. An example of an MVP is the release of a consumer website that has only 10 percent of the functionality that the founders ultimately envision for the product. However, by quickly building a workable version with only the bare essentials, the MVP approach seeks to validate as many assumptions as possible about the viability of the final product before expending enormous effort and financial resources. From the information collected during these experiments, the venture adjusts its course in a way that may involve pivoting from the original agenda. The followers of this methodology frequently discuss how to make their experiments ever more cost-effective, in large part so that they do not need to raise as much money to pursue a range of possible ideas.

This approach to building companies has coincided with the rapid rise of angel investors and crowd-funding platforms,[6] particularly for consumer Internet startups.

[6] Angel investors are defined as individuals investing their own capital in startup ventures. Given the decentralized nature of such investments, it is extremely difficult to accurately quantify the size of the market. The Angel Capital Association (2012) estimates that its membership consists of about 8,000 accredited investors (who are individuals allowed by the Securities and Exchange Commission to make investments in risky, private startups) who invest in approximately 800 companies per year. The size of the broader angel market has been estimated to be as large as $20 billion dollars, but the vast majority of this is tiny investments made by individuals in firms that are already cash-flow positive (Shane 2009). The total does, however, provide a sense of how large the crowd-funding market could be given recent changes in regulation through the Jumpstart Our Business Startups (JOBS) Act of 2012, which allow a larger number of individuals to invest legally in private companies (Nanda and Kind 2013).

For example, AngelList, a platform dedicated to matching startups with angel investors experienced rapid growth since its founding in 2007, and in June 2013 had approximately 100,000 startups and 18,000 accredited investors on its platform, of which 2,000 had been funded (Nanda and Kind 2013). The lower costs associated with initial experiments implies that those with pools of capital smaller than typical venture capitalists can more easily become involved in these financing decisions. Because positive information from experiments leads the valuation of startups in subsequent rounds to increase substantially, this leads to less dilution for early-stage financiers and makes it even more attractive for investors with smaller pools of capital to get involved with financing decisions. Equally important, however, is the fact that as these experiments get better at quickly telling apart projects with high and low potential, they allow startups that would probably not have received funding in the past due to their uncertain outcomes to now receive financing. Ewens, Nanda, and Rhodes-Kropf (2014) show that these are particularly likely to be startups that had a very small chance of being extremely successful, such as those founded by young, inexperienced, first-time entrepreneurs. While these individuals would not have received capital in the past, the financing environment has changed in recent years, creating room for cohort-based accelerator programs with an educational component, such as Y-Combinator, that aim to provide mentoring, networks, and capital to such founders (Graham 2008). Fehder and Hochberg (2014) document that there were no such programs in the United States in 2004, but by 2013 over 40 such accelerators were functioning across the United States.

While a trend toward lower start-up costs is particularly true in software and information technology businesses, where the cost of experimentation has declined dramatically and the frequency with which one learns new information about the product is very high, the same forces exist in very capital-intensive industries as well. An example is Terrapower, a startup trying to commercialize a new way of producing nuclear energy (Sahlman, Nanda, Lassiter, and McQuade 2012). Bill Gates, an early backer of Terrapower, noted that rapid advances in computing power provide this startup with significant advances toward experimentation. In the past, Terrapower would have had to construct an entire nuclear power plant to test whether its new technology would work in practice, costing several billion dollars and taking years to complete, making it impossible to finance in the first place. Now, with the introduction of supercomputers, Terrapower's engineers can simulate the inside of a nuclear reactor, learn whether its technology will work, and make rapid (and much cheaper) iterations to gain more confidence about the potential of the technology before ever constructing the physical nuclear power plant. This in turn makes it viable to finance an initial exploration for an early-stage investor.

These examples reinforce two points about the cost of experimentation. First, differences across industries in the ability and cost to learn about the final outcome have a huge bearing on the degree of experimentation that we see in early stages—the very long time frames and costs for learning about potential technologies in renewable energy or different approaches to curing cancer create a dearth of experimentation, despite intense societal interest (Fernandez, Stein, and Lo 2012). Even

for venture capital firms, the levels of investment required for clean energy efforts or certain healthcare investments tend to be too high per project for them to bear.

Second, the pace of technological progress has a feedback effect on new innovations. As technological advances such as supercomputers or cloud computing diffuse, they affect the cost of experimentation in completely different sectors like nuclear energy and consumer Internet. The advances mean that projects that had a negative expected value in the past become viable. Referring to our earlier numerical example, they may drop the cost of the hypothetical test from several hundred dollars to less than $89. This spillover effect helps sustain and potentially even increase high rates of technological progress.

Frictions for the Process of Experimentation—Organizational

A key part of experimentation is the ability to terminate projects, but organizations differ in their ability and willingness to terminate underperforming ventures. The venture capital industry has long used the expression "throwing good money after bad money" to warn against continual investment in startups for whom the initial experiments reveal poor information. Venture capital firms that shut down more ventures do not necessarily perform worse, and, in fact, some of the best venture capital firms have among the highest rates of abandoning projects—in part because their skill at designing experiments and acting upon the results allows them to pursue more aggressive strategies and enter into more uncertain domains.[7]

This capacity to terminate projects, while certainly not exclusive to the venture capital community, is often difficult to replicate within large corporations or through equity markets. Large corporations, for example, often find it difficult to terminate experiments that aren't working out, due in part to career concerns of the managers in charge of the effort, the "soft budget constraints" that arise when a large corporation can keep providing funding for a time, and similar traits. The initial experiment may also provide ambiguous or slightly negative information that can be interpreted in different ways depending upon the biases of parties. In several ways, the venture capital system is designed to counteract these tendencies, beginning with an industry structure where venture capital firms are separate from the startup itself but still capable of intensive monitoring. Another mechanism often used by venture capital firms is to bring in new investors with each financing round to ensure that the deal is evaluated by outsiders. In academia, the academic tenure process that includes external review after a specific time horizon generates comparable features.

Of course, because venture capital investors consider each investment as contributing to overall portfolio returns, their incentives are not always aligned with entrepreneurs in terms of strategic decisions such as when to shut down or exit

[7] While the higher rate of abandoning projects does not jump out from the statistics we provided above, it is worth noting that venture capital investors also likely select better projects, so that a simple comparison of realized failure rates does not provide a complete picture of how much more they abandon projects compared to non-venture-capital-backed firms.

an investment. For example, Gompers (1996) has documented the incentives of venture capital firms to engage in grandstanding (taking companies public earlier than may be optimal) to develop a reputation that enables them to raise follow-on funds. Venture capital investors may also choose to shut down an underperforming firm if it is unlikely to generate high gross returns, or they may continue to reinvest if they want to build a reputation as entrepreneur-friendly. These decisions may not always be optimal for the firm.

These examples highlight a related tension with organizational design, however. Innovation requires running experiments that will often fail, and such failures will occur even if the idea looked promising at the start and all parties acted in good faith. These features may require an organizational tolerance for failure in the short-term, with compensation rewards over the long-run based upon success (Manso 2011). This tolerance allows opportunities for early experimentation to identify the best course of action, while the long-term compensation tied to performance combats the "moral hazard" concern that decision makers might take excessive risks, and it solicits their best efforts.

How organizations balance the termination of poor experiments with maintaining a tolerance for failure in turn affects the overall set of projects that they choose to start. Nanda and Rhodes-Kropf (2012) show that the inability or the unwillingness to terminate projects when intermediate information is negative leads organizations to start projects that are less experimental in the first place. In other words, organizations can pick a radical range of projects at the start and terminate a high percentage of them, or pick a more limited range of projects at the start and terminate fewer of them. Large bureaucratic organizations often struggle with terminating projects, and these differences provide a rationale for why large companies are comparatively better at pursuing incremental improvements than radical innovations, although this is of course not true in every case. March (1991) and Akcigit and Kerr (2010) provide greater discussion on firm sizes and exploration versus exploitation choices. Lerner (2012) provides a discussion of the relative benefits of the traditional model of research and development within firms versus the model of venture capital investors, and Thomke (2003) discusses how large companies can take insights from the model of experimentation to improve the way in which they develop and advance innovations.

Frictions for the Process of Experimentation—Continuation and Financing Risk

Successful experimentation requires being able to capitalize on experiments that reveal positive outcomes, and these upside scenarios can be as tricky as termination decisions. Financing projects in stages requires that startups return to the capital markets at regular intervals for more capital. Thus far, we have emphasized positive aspects of multiple financing rounds (for example, the advantages of having an outside investor not previously involved with the company value the venture). A negative aspect is that entrepreneurs and early-stage investors feel vulnerable to the state of the financial markets at each round of financing. At times, financial capital is freely available, while in other instances it may be hard to come by, even

for otherwise sound projects. These financing constraints can be marketwide or they can be specific to sectors due to the degree to which investors herd around certain hot sectors.

Venture capital investors routinely refer to "financing risk" (similar to rollover risk) to describe how otherwise sound projects may not obtain capital for the next experiment. One way to protect against financing risk is to go less frequently to the capital markets, by taking larger chunks of money at each stage. This, however, reduces the value of the abandonment options for venture capital firms. Nanda and Rhodes-Kropf (2010, 2013) argue that hot markets—times when financing risk is low—allow projects with the highest real option values to be funded, because the continuation risk is lower for all projects in the economy. Empirically, venture-capital-backed firms that are first funded during hot markets have the highest failure rates, but conditional on being successful, startups funded in more-active periods were valued higher at initial public stock offering or acquisition compared to startups funded in less-active periods exiting at the same time. Financing risk is most salient for projects that need to go back to the capital markets many times, and thus projects that are worst hit by this risk are experimental new technologies that tend to have the highest option value. In fact, a possible implication of this work, which is related to work on the role of stock markets in financing innovation (Brown, Fazzari, and Petersen 2009), is that the most innovative projects in the economy may *need* times of low financing risk ("hot financial markets") to drive their initial commercialization.

Financing risk is representative of a broader class of continuation features that govern the degree to which experimentation can be pursued. To illustrate, consider some of the differences between clean energy and biotech; biotech also requires substantial investment and very long horizons but is the beneficiary of substantially more venture capital investment. Biotech startups and the pharmaceutical industry are part of a vibrant "market for ideas" (Gans, Hsu, and Stern 2002) that allows biotech companies to sell the results of their experiments at milestones. That is, a biotech company can undertake a first experiment to show that a project that originally seemed to have had a 0.1 percent chance of becoming a blockbuster drug has in fact a 5 percent chance. This experiment has generated substantial information value, and pharmaceutical companies are willing to buy the remaining 1-in-20 opportunity for their drug development portfolios. This market works because the science used in the first experiment is observable and verifiable, a patent system protects the startup from intellectual theft, and the long-term opportunity is more protected compared to clean energy innovations (which can face competition from all alternative sources of energy, including the opening up of strategic oil reserves). Such a "market for ideas" is much less developed for clean energy, and as a consequence entrepreneurs and investors who are considering an experiment in this area face additional continuation risks (Nanda, Younge, and Fleming, forthcoming). Relatedly, competition to acquire new technologies among pharmaceutical firms combined with a relatively robust market for initial public offerings have meant that the upside from biotechnology innovation, if successful, is sufficiently high to warrant the financing of upstream

experimentation. We return to the role of the markets for initial public offerings in greater detail in the next section.

Discrete Choices and Individuals—A Process of Experimentation

We close this section by emphasizing its most prominent feature—the role of humans, often just a few humans, in making these decisions. Economic models often focus on the fact that in well-functioning economies, good ideas are rewarded and flourish relative to worse ideas. But they are silent on how these ideas emerge.

A central theme of this section is that entrepreneurs and investors must make discrete choices, often with quite incomplete information and uncertainty about the future, about which ideas to progress and which to shut down. Financiers rather than markets dictate which projects are realized, as they choose which experiments to attempt, how to interpret the results, and whether to continue or abandon the investment. The entrepreneurial landscape at the micro-level is thus characterized by individuals choosing the fate of a venture based on what they have learned about what is likely to work *before* it competes in the product market.

This section has emphasized how constraints on the ability of investors to experiment efficiently can shape which industries, organizations, and time periods see the most radical innovations. It also sets the framework for understanding where barriers to experimentation may lead to market failures. For example, intermediate experiments may reveal very little about the ultimate success of some projects (such as building a particle collider or the nuclear power startup before the advent of simulation), requiring extremely large investments before one can learn about their viability. Alternatively, the horizon for commercialization may be extremely uncertain and distant, such as in the case of research in basic science. Institutional regimes such as academia (and at times the government) may thus be critical to enable experimentation in areas that are of importance to society but where a process of serial experimentation by profit-seeking investors is unlikely to provide a set of stepping-stones to the technologies behind disruptive innovation.

Economic Experiments at the Society Level

While we have focused our attention on the difficulty in predicting outcomes for new technologies, many of these challenges also exist for firms in the broader economy. For example, Shane (2009) documents that of the approximately 500,000 startups that were founded in the United States in 1996, over half had failed by 2002 and only about 3,500 achieved sales of greater than $10 million. Haltiwanger, Jarmin, and Miranda (2013) note the difficulty in predicting upfront which firms will succeed even among producers of homogeneous goods such as cement. The laws and institutions at the regional and national level can therefore play a critical role in driving productivity growth—often through the incentives they create or the costs that they impose on the ability to effectively experiment. For example, Glaeser, Kerr, and Kerr (forthcoming) trace the effects of entrepreneurship on long-run

urban growth by documenting how cities with industrial and mining legacies in 1900 found it harder to create the up-or-out dynamic associated with productivity growth. They argue that the institutions (industrial structure, culture, and skill base) created by the predominantly large mining firms in the 1900s hindered entrepreneurship in the modern era, leading to systematically lower urban growth in these regions post-1970.

Although some institutions are extremely persistent and hard to change, we outline some types of policies that seem to be particularly effective in promoting "economic experiments" in the form of entrepreneurship.

Democratizing Entry and Facilitating Efficient Failure

Bank finance and debt financing more generally is an extremely important source of capital for young companies. The US Small Business Administration has estimated that there were 21 million small business commercial and industrial loans outstanding in 2010 (defined as loans being below $1 million) that were valued at $310 billion (Office of Advocacy, US SBA 2013). While many small businesses are not young, Robb and Robinson (2014) document that debt finance is important even among young firms. They look at the sample of startups in the Kauffman Firm Survey and find that within the first three years of founding, 40 percent of the funding source for these startups is constituted by outside debt, over and above the 4 percent of debt that consists of the owner's credit cards and personal loans.

Changes in the availability and terms of bank finance can therefore have important implications for entrepreneurship. For example, Kerr and Nanda (2009) consider how the banking competition fostered by the state-level US banking deregulations that occurred from the 1970s through the 1990s affected entrepreneurship and the economy. These deregulations facilitated greater competition by allowing out-of-state banks to enter local markets, thereby increasing access to credit and lowering the cost of external finance. Using micro-level data from the US Census Bureau, Kerr and Nanda (2009) find evidence for the standard story of creative destruction—a few strong entrants challenging and later replacing incumbent firms. However, the mechanism through which this was achieved was through widespread entry that largely resulted in failure, so that the most pronounced impact was a massive increase in churning among new entrants.[8] Similarly, Chava, Oettle, Subramanian, and Subramanian (2013) examine innovation outcomes following the banking deregulations and find that banking deregulation facilitated greater risk taking and experimentation by small firms. Because it is hard to know beforehand which projects are going to be successful, "democratizing entry" seems to be an important trait of well-functioning capital markets.

Efficient experimentation implies that institutional environments facilitating exit are as important as those that facilitate entry. Two areas that have received

[8] Prior work by Black and Strahan (2002) and Cetorelli and Strahan (2006) connected the deregulations to higher entry rates, but it was quite puzzling that the elasticity for entrepreneurship was about ten times higher than the response observed on any other dimension like productivity change.

significant attention in recent years are the role of bankruptcy law and employment protection laws and tradeoffs involved as they relate to experimentation.

On the one hand, bankruptcy laws that favor creditors allow them to recoup as much capital as they can from the startup—encouraging them to lend in the first place (Berkowitz and White 2004; Cumming forthcoming; Cerqueiro, Hegde, Penas, and Seamans 2013). Such laws also discourage the moral hazard that might otherwise arise among entrepreneurial agents if they faced little need to repay creditors. On the other hand, entrepreneurial failure allows society to test uncertain projects. This process requires separating the entrepreneur from the firm, as is accomplished to some extent by the limited liability provision that entrepreneurs are not personally liable for the debts of a firm, and thus provides entrepreneurs with the ability to terminate projects and move on. As an example, Eberhart, Eesley, and Eisenhardt (2013) find that a bankruptcy reform in Japan that reduced the consequences of closing a firm encouraged greater levels of entrepreneurship and risk taking following the reform. The best policy appears to involve striking a balance between a good measure of limited liability for entrepreneurs, allowing them to transition across projects without severe and lasting penalties, along with some protections or restrictions for small or unsophisticated investors who may not comprehend fully the low likelihood of the entrepreneur's success. While legal factors play a role in reducing the downside from failure, cultural factors such as a stigma of failure can also play a role in hindering entry. In some regions, such as Silicon Valley, past failure can even be seen as a badge of honor, making it much easier for individuals to take big risks in terms of startup ventures.

A related set of issues can arise from strict employment protection laws that limit the ability of firms to adjust their workforce rapidly and in that way act as an effective tax on employment adjustments (Autor, Kerr, and Kugler 2007). Many ultimately successful startups have had to undertake one or more critical changes in their business model, which in turn require substantial adjustments in their workforce. One of the tradeoffs of legal mandates for employment is that they make it much more difficult for high-growth startups to experiment and for innovative, volatile sectors to form (for example, Saint-Paul 2002; Samaniego 2006). Indeed, venture capital investment is lower in countries with stringent employment protection laws, and the more-volatile sectors are the most affected (Bozkaya and Kerr forthcoming). In this sense, the flexibility of labor markets governs the types of projects that entrepreneurs can undertake (Fallick, Fleischman, and Rebitzer 2006).

Appropriating Value in Successful States of the World

Beyond the policies that shape an entrepreneur's ability to experiment when things are going poorly, policies or conditions that limit the value that entrepreneurs or investors receive in good states of the world are also important. Some conditions like the extent of patent protection were noted earlier in our comparison of biotech and clean energy investments. More generally, the investors and state-contingent financial contracts that we describe require strong property rights

and a sound rule of law (for example, La Porta, Lopez-de-Silanes, Shleifer, and Vishny 1997). Similarly, public equity markets are critical to rewarding startups that are successful, by allowing investors and entrepreneurs to cash out on the expectation of future growth. When equity markets do not work well or there are limited opportunities to exit and retrieve at least a portion of the earlier investment, it becomes much harder to experiment (for example, Black and Gilson 1998; Michelacci and Suarez 2004).

Policy Implications

How should policymakers interested in promoting entrepreneurship think about the role of experimentation? Clearly, difficult entry regulations suppress the experimentation undertaken by startups (Klapper, Laeven, and Rajan 2006) and often are difficult to justify on any grounds. As one example, Fairlee, Kapur, and Gates (2011) describe the extent to which the prevalence of employer-provided health insurance in the US economy serves as an implicit entry barrier for potential founders of new firms, making it more difficult for them to leave current jobs.

In addition, taking the experimental perspective seriously suggests that it is a poor idea for government to seek to pick and promote individual firms. After all, even the most-experienced venture capital firms have substantial success in only one of every ten investments they pick, so we shouldn't expect inexperienced and possibly not-very-objective politicians to do better. Indeed, one of the features that make market-based economies better at commercializing radical innovations is the decentralized and parallel nature with which new ideas are tested (Rosenberg 1994; Qian and Xu 1998; Scherer and Harhoff 2000). Politicians also have greater difficulty terminating projects—that is, telling taxpayers that legislative decisions have spent their money with little or no return. In our experience, most economists buy into this wariness of policymakers acting as a venture capital firm, and many go a step further and caution against picking an individual sector or industry to support. Virtually every state has a biotech initiative, but few will be successful. Lerner (2009) emphasizes the dismal performance of policies attempting to seed specific ventures or industries.

Governments are more likely to facilitate effective entrepreneurship if they work to reduce the costs of experimentation in general. What we have in mind here is not so much government programs that seek to target start-ups with particular benefits, but instead a careful consideration of the broader regulatory framework, including labor laws and requirements with which new entrants need to comply, with a focus on how they affect incentives for entry. These efforts to structure a better playing field are admittedly less glamorous than announcing a new biotech cluster initiative, but they are far more likely to have sustained effects.

As alluded to in the previous section, the experimentation view also suggests that there may be systematic market failures when the costs associated with experimentation are too high or the returns are too uncertain and far into the future. This creates a framework for thinking about which sectors, such as say basic science, may warrant sustained government support.

Finally, government should be cautious about industrial policies that seek to minimize business failures, as such policies may only be propping up firms that need to fail. For example, Acemoglu, Akcigit, Bloom, and Kerr (2013) build a model of firm-level innovation, productivity growth, and reallocation featuring endogenous entry and exit, and calibrate it using firm-level micro data from the Census. Their results highlight a key role for new entrants in policies to promote innovation and growth. Acemoglu et al. (2013) find that while policies like research and development tax credits to entrants can help and may encourage growth, their impact pales in comparison to removing artificial support for inefficient incumbents.

Conclusions

Picking up a quotation originally made by Michael Jordan to describe missing basketball shots, a prominent venture capitalist, Vinod Khosla, emphasizes, "our willingness to fail gives us the ability and opportunity to succeed where others may fear to tread" (Khosla Ventures, no date; Khosla 2013). This paper has outlined our perspective on entrepreneurship that emphasizes this fundamental role for experimentation. Viewing entrepreneurship as experimentation allows individuals and societies to evaluate businesses and technologies in domains with greater uncertainty than otherwise possible, unlocking deep growth opportunities.

Economists should bear in mind two very different forms of experimentation that are associated with entrepreneurship. One form is the economic experimentation that takes place in market-based economies when several new ideas, products and technologies are continually tested and either displace existing technologies or more likely fail themselves. A second less-appreciated form of experimentation happens at the micro-level, before these ideas compete, and relates to the process of bringing new ideas to the market. There are too many opportunities to pursue with scarce resources, and the best-designed experiments can come back with misleading or ambiguous results. Under these conditions, specific choices made by discrete individuals—especially founders of firms and early-stage investors—become very important. Recent research in financial economics, organizational design, and related fields highlight how the costs associated with the ability of investors to experiment can alter the nature of entrepreneurship and help to explain why entrepreneurship is more prevalent in certain industries, regions, or periods of time.

Less than a decade ago, research on entrepreneurship within economics had little internal cohesion; instead, researchers working on entrepreneurship were spread out within broader fields like finance, labor economics, or macroeconomics. Today, while most researchers in this area continue to keep one foot firmly planted in a traditional field, the internal cohesion for entrepreneurial research has formed and is rapidly obtaining scale. Researchers on entrepreneurship are themselves experimenting with new ideas and new directions, and we expect this area of research to develop rapidly in the coming years.

■ *We appreciate comments and editorial direction by David Autor, Chang-Tai Hsieh, and Timothy Taylor. The research in this paper was conducted while the authors were Special Sworn Status researchers of the US Census Bureau at the Boston Census Research Data Center (BRDC). Support for this research from the NSF grant ITR-0427889 [BRDC], the Division of Research and Faculty Development at Harvard Business School, and from the Kaufman Foundation is gratefully acknowledged. Research results and conclusions expressed are the authors and do not necessarily reflect the views of the Census Bureau or the NSF. This paper has used proprietary data from a venture capital firm that requests it remain anonymous, and it has screened the paper to ensure that no confidential data are revealed.*

References

Acemoglu, Daron, Ufuk Akcigit, Nicholas Bloom, and William R. Kerr. 2013. "Innovation, Reallocation and Growth." NBER Working Paper 18993.

Akcigit, Ufuk, and William Kerr. 2010. "Growth through Heterogeneous Innovations." NBER Working Paper 16443.

Angel Capital Association. 2012. "ACA, Angel Groups and Angel-Backed Companies." Slides, September. http://www.angelcapitalassociation.org/data/Documents/Resources/ACAandAngelGroupBackground09-12.pdf.

Autor, David H., William R. Kerr, and Adriana D. Kugler. 2007. "Does Employment Protection Reduce Productivity? Evidence from US States." *Economic Journal* 117(521): 189–217.

Bergemann, Dirk, and Ulrich Hege. 2005. "The Financing of Innovation: Learning and Stopping." *RAND Journal of Economics* 36(4): 719–52.

Bergemann, Dirk, Ulrich Hege, and Liang Peng. 2008. "Venture Capital and Sequential Investments." Working paper.

Berkowitz, Jeremy, and Michelle J. White. 2004. "Bankruptcy and Small Firms' Access to Credit." *Rand Journal of Economics* 35(1): 69–84.

Bessemer Venture Partners. No date. "Anti-portfolio." http://www.bvp.com/portfolio/anti portfolio.

Blacharski, Dan. 2013. "Top Ten Ways the Cloud Has Changed How Startups Launch." Techie.com, November, 26. http://techie.com/top-ten-ways-the-cloud-has-changed-how-startups-launch/.

Black, Bernard S., and Ronald J. Gilson. 1998. "Venture Capital and the Structure of Capital Markets: Banks versus Stock Markets." *Journal of Financial Economics* 47(3): 243–77.

Black, Sandra E., and Philip E. Strahan. 2002. "Entrepreneurship and Bank Credit Availability." *Journal of Finance* 57(6): 2807–33.

Bozkaya, Ant, and William Kerr. Forthcoming. "Labor Regulations and European Venture Capital." *Journal of Economics and Management Strategy.*

Brown, James R., Steven M. Fazzari, and Bruce C. Petersen. 2009. "Financing Innovation and Growth: Cash Flow, External Equity and the 1990s R&D Boom." *Journal of Finance* 64(1): 151–85.

Cerqueiro, Geraldo, Deepak Hegde, María Fabiana Penas, and Robert Seamans. 2013. "Debtor Rights, Credit Supply, and Innovation." Unpublished paper.

Cetorelli, Nicola, and Philip Strahan. 2006. "Finance as a Barrier to Entry: Bank Competition and Industry Structure in Local U.S. Markets." *Journal of Finance* 61(1): 437–61.

Chava, Sudheer, Alexander Oettl, Ajay Subramanian, and Krishnamurthy Subramanian.

2013. "Banking Deregulation and Innovation." *Journal of Financial Economics* 109(3): 759–74.

Chemmanur, Thomas J., Karthik Krishnan, and Debarshi K. Nandy. 2011. "How Does Venture Capital Financing Improve Efficiency in Private Firms? A Look beneath the Surface." *Review of Financial Studies* 24(12): 4037–90.

Cumming, Douglas. Forthcoming. "Measuring the Effect of Bankruptcy Laws on Entrepreneurship across Countries." *Journal of Entrepreneurial Finance*.

Da Rin, Marco, Thomas Hellmann, and Manju Puri. 2013. "A Survey of Venture Capital Research." Chap. 8 in *Handbook of the Economics of Finance*, Vol. 2, Part A, edited by George Constantinides, Milton Harris, and René M. Stulz. Amsterdam: North Holland.

Dosi, Giovanni, and Richard R. Nelson. 2010. "Technical Change and Industrial Dynamics as Evolutionary Processes." In *Handbook of the Economics of Innovation*, Vol. 1 edited by B. H. Hall and N. Rosenberg, 51–127. Amsterdam: North Holland.

Eberhart, Robert, Charles E. Eesley, and Kathleen M. Eisenhardt. 2013. "Failure is an Option: Failure Barriers and New Firm Performance." Stanford University Working Paper 111.

Ewens, Michael, Ramana Nanda, and Matthew Rhodes-Kropf. 2014. "Entrepreneurship and the Cost of Experimentation." Unpublished paper.

Fairlee, Robert W., Kanika Kapur, and Susan Gates. 2011. "Is Employer-based Health Insurance a Barrier to Entrepreneurship?" *Journal of Health Economics* 30(1): 146–62.

Fallick, Bruce, Charles A. Fleischman, and James Rebitzer. 2006. "Job-hopping in Silicon Valley: Some Evidence Concerning the Microfoundations of a High-Technology Cluster." *Review of Economics and Statistics* 88(3): 472–81.

Fehder, Daniel C., and Yael V. Hochberg. 2014. "Accelerators and the Regional Supply of Venture Capital Investment." Unpublished paper.

Fernandez, Jose-Maria, Roger M. Stein, and Andrew W. Lo. 2012. "Commercializing Biomedical Research through Securitization Techniques." *Nature Biotechnology* 30(10): 964–75.

Fleming, Lee. 2001. "Recombinant Uncertainty in Technological Search." *Management Science* 47(1): 117–32.

Fleming, Lee, and Olav Sorenson. 2004. "Science as a Map in Technological Search." *Strategic Management Journal* 25(5–9): 909–28.

Gans, Joshua, David H. Hsu, and Scott Stern. 2002. "When Does Start-up Innovation Spur the Gale of Creative Destruction?" *RAND Journal of Economics* 33(4): 571–86.

Glaeser, Edward L., Sari Pekkala Kerr, and William R. Kerr. Forthcoming. "Entrepreneurship and Urban Growth: An Empirical Assessment with Historical Mines." *Review of Economics and Statistics*.

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

Gompers, Paul A. 1996. "Grandstanding in the Venture Capital Industry." *Journal of Financial Economics* 42(1): 133–56.

Gompers, Paul, and Josh Lerner. 2004. *The Venture Capital Cycle*. Cambridge, MA: MIT Press.

Graham, Paul. 2008. "A New Venture Animal." March, http://paulgraham.com/ycombinator.html.

Grenadier, Steven R., and Andrey Malenko. 2011. "Real Options Signaling Games with Applications to Corporate Finance." *Review of Financial Studies* 24(12): 3993–4036.

Hall, Robert E., and Susan E. Woodward. 2010. "The Burden of the Nondiversifiable Risk of Entrepreneurship." *American Economic Review* 100(3): 1163–94.

Haltiwanger, John, Ron S. Jarmin, and Javier Miranda. 2013. "Who Creates Jobs? Small versus Large versus Young." *Review of Economics and Statistics* 95(2): 347–61.

Hayek, Friedrich A. 1948. *Individualism and Economic Order*. University of Chicago Press.

Hellmann, Thomas, and Manju Puri. 2000. "The Interaction between Product Market and Financing Strategy: The Role of Venture Capital." *Review of Financial Studies* 13(4): 959–84.

Kaplan, Steven N., Berk A. Sensoy, and Per Strömberg. 2009. "Should Investors Bet on the Jockey or the Horse? Evidence from the Evolution of Firms from Early Business Plans to Public Companies." *Journal of Finance* 64(1): 75–115.

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

Kaplan, Steven N., and Per Strömberg. 2004. "Characteristics, Contracts, and Actions: Evidence from Venture Capitalist Analyses." *Journal of Finance* 59(5): 2177–2210.

Kerr, William R., Josh Lerner, and Antoinette Schoar. 2014. "The Consequences of Entrepreneurial Finance: Evidence from Angel Financings." *Review of Financial Studies* 27(1): 20–55.

Kerr, William R., and Ramana Nanda. 2009. "Democratizing Entry: Banking Deregulations, Financing Constraints, and Entrepreneurship." *Journal of Financial Economics* 94(1): 124–49.

Klapper, Leora F., Luc Laeven, and Raghuram G. Rajan. 2006. "Entry Regulation as a Barrier to Entrepreneurship." *Journal of Financial Economics* 82(3): 591–625.

Knight, Frank H. 1921. *Risk, Uncertainty, and Profit.* Boston, MA: Houghton Mifflin.

Kortum, Samuel S., and Josh Lerner. 2000. "Assessing the Contribution of Venture Capital to Innovation." *RAND Journal of Economics* 31(4): 674–92.

Khosla Ventures. No date. Company website mainpage, http://www.khoslaventures.com/ (accessed 5/14/14).

Khosla, Vinod. 2013. "Fail as Often as You Want, And Fail Well." Video of a speech posted March 10, 2013, at Wunderman Reports.com. http://www.wundermanreports.com/fail-as-often-as-you-want-and-fail-well.

Landier, Augustin. 2005. "Entrepreneurship and the Stigma of Failure." Unpublished paper.

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. 1997. "Legal Determinants of External Finance." *Journal of Finance* 52(3): 1131–50.

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 University Press.

Lerner, Josh. 2012. *The Architecture of Innovation: The Economics of Creative Organizations.* Boston, MA: Harvard Business School Press.

Levere, Jane L. 2013. "Airbnb Campaign Uses Birdhouses to Widen Its Reach." *New York Times,* December 15. http://www.nytimes.com/2013/12/16/business/media/airbnb-campaign-uses-birdhouses-to-widen-its-reach.html.

Manso, Gustavo. 2011. "Motivating Innovation." *Journal of Finance* 66(5): 1823–69.

March, James G. 1991. "Exploration and Exploitation in Organizational Learning." *Organizational Science* 2(1): 71–87.

Michelacci, Claudio, and Javier Suarez. 2004. "Business Creation and the Stock Market." *Review of Economic Studies* 71(2): 459–81.

Nanda, Ramana, and Liz Kind. 2013. "AngelList." Harvard Business School Case 814-036.

Nanda, Ramana, and Matthew Rhodes-Kropf. 2010. "Financing Risk and Innovation." Harvard Business School Working Paper 11–013.

Nanda, Ramana, and Matthew Rhodes-Kropf. 2012. "Innovation and the Financial Guillotine." Harvard Business School Working Paper 13–038.

Nanda, Ramana, and Matthew Rhodes-Kropf. 2013. "Investment Cycles and Startup Innovation." *Journal of Financial Economics* 110(2): 403–18.

Nanda, Ramana, Ken Younge, and Lee Fleming. Forthcoming. "Innovation and Entrepreneurship in Renewable Energy." In *The Changing Frontier: Rethinking Science and Innovation Policy* edited by Adam Jaffe and Ben Jones. National Bureau of Economic Research.

Office of Advocacy, US Small Business Association. 2013. *Small Business Lending in the United States 2012.* http://www.sba.gov/sites/default/files/files/sbl_12study.pdf.

Palmer, Maija. 2012. "Cloud Computing Cuts Start-up Costs." *Financial Times,* February 29.

Parker, Simon C. 2004. "The Economics of Self-Employment and Entrepreneurship." Cambridge University Press.

Puri, Manju, and Rebeca Zarutskie. 2012. "On the Lifecycle Dynamics of Venture-Capital- and Non-Venture-Capital-Financed Firms." *Journal of Finance* 67(6): 2247–93.

Qian, Yingyi, and Chenggang Xu. 1998. "Innovation and Bureaucracy under Soft and Hard Budget Constraints." *Review of Economic Studies* 65(1): 151–64.

Rajan, Raghuram G. 2012. "Presidential Address: The Corporation in Finance." *Journal of Finance* 67(4): 1173–1217.

Ries, Eric. 2011. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.* Crown Business Publishing.

Ritter, Jay R. 2014. "Initial Public Offerings: Updated Statistics."

Robb, Alicia M., and David T. Robinson. 2014. "The Capital Structure Decisions of New Firms." *Review of Financial Studies* 27(1): 153–79.

Rosenberg, Nathan. 1994. "Economic Experiments." In Chap. 5 in *Exploring the Black Box: Technology, Economics, and History.* Cambridge University Press.

Sahlman, William A. 1990. "The Structure and Governance of Venture-Capital Organizations." *Journal of Financial Economics* 27(2): 473–521.

Sahlman, William A. 2010. "Risk and Reward in Venture Capital." Harvard Business School Background Note 811-036.

Sahlman, William, Ramana Nanda, Joseph B. Lassiter III, and James McQuade. 2012. "TerraPower." Harvard Business School Case Study 813-108.

Saint-Paul, Gilles. 2002. "Employment Protection, International Specialization, and Innovation." *European Economic Review* 46(2): 375–95.

Samaniego, Roberto M. 2006. "Employment Protection and High-Tech Aversion." *Review of Economic Dynamics* 9(2): 224–41.

Samila, Sampsa, and Olav Sorenson. 2011. "Venture Capital, Entrepreneurship and Economic Growth." *Review of Economics and Statistics* 93(1): 338–49.

Scherer, F. M., and Dietmar Harhoff. 2000. "Technology Policy for a World of Skew-distributed Outcomes." *Research Policy* 29(4–5): 559–66.

Schumpeter, Joseph. 1911 [1934]. *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest and the Business Cycle.* Cambridge, MA: Harvard University Press.

Schumpeter, Joseph. 1943. *Capitalism, Socialism, and Democracy.* Harper Brothers, New York, NY.

Shane, Scott A. 2009. *Fool's Gold? The Truth behind Angel Investing in America.* Oxford University Press.

Stern, Scott. 2005. "Economic Experiments: The Role of Entrepreneurship in Economic Prosperity." In *Understanding Entrepreneurship: A Research and Policy Report*, edited by Carl J. Schramm, 16–20. Ewing Marion Kauffman Foundation.

Thomke, Stefan H. 2003. *Experimentation Matters: Unlocking the Potential of New Technologies for Innovation*. Boston, MA: Harvard Business School Press.

Tian, Xuan. 2011. "The Causes and Consequences of Venture Capital Stage Financing." *Journal of Financial Economics* 101(1): 132–59.

Weitzman, Martin. 1979. "Optimal Search for the Best Alternative." *Econometrica* 47(3): 641–54.

Wilson, Fred. 2011. "Airbnb." March 16. http://www.avc.com/a_vc/2011/03/airbnb.html.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 49–70

Seeking the Roots of Entrepreneurship: Insights from Behavioral Economics[†]

Thomas Åstebro, Holger Herz, Ramana Nanda, and Roberto A. Weber

Frank Knight (1921) proposed that we should not think of entrepreneurship as simply investment under risk, where decisions are made with respect to an objectively known distribution of returns. He argued that entrepreneurship in such a world would not require any particular skill and it would be inconceivable that entrepreneurs could earn rents simply for bearing objective risk as a market should eliminate those rents. Knight therefore put forward the idea that the prerequisites for entrepreneurial activity are a combination of highly uncertain returns that do not have an objectively known distribution, as well as the entrepreneur's skill in perceiving opportunity more clearly than others.

Knight's (1921) work focused attention on the specific individuals pursuing entrepreneurship and what made them distinct. Subsequent research in this vein has aimed to understand the individual traits, motivations, and preferences that make some individuals more likely to pursue entrepreneurship than others. Ironically, while Knight was interested in justifying why entrepreneurs should earn supernormal returns, much of this subsequent research has found the opposite to be true. That is, there is a growing body of evidence that many entrepreneurs seem

■ *Thomas Åstebro is Associate Professor of Strategy and Entrepreneurship, HEC Paris, Paris, France. Holger Herz holds a senior research associate position in Microeconomics and Experimental Economic Research, University of Zurich. Zurich, Switzerland. Ramana Nanda is Associate Professor of Business Administration, Harvard Business School, Boston, Massachusetts, and a Faculty Research Fellow at the National Bureau of Economic Research, Cambridge, Massachusetts. Roberto A. Weber is Professor of Economics, University of Zurich, Zurich, Switzerland. Their email addresses are astebro@hec.fr, holger.herz@econ.uzh.ch, RNanda@hbs.edu, and roberto.weber@econ.uzh.ch.*

[†] To access the data Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.3.49

doi=10.1257/jep.28.3.49

to enter and persist in entrepreneurship despite earning *low* risk-adjusted returns. This finding has led, in turn, to attempts to provide explanations—using both standard economic theory and behavioral economics—for why certain individuals may be attracted to such an apparently unprofitable activity.

In this article, we critically evaluate what the existing research shows regarding the individual determinants of entrepreneurship. We begin by documenting a set of facts that seem to pose a challenge for interpretations of entrepreneurship based on the standard expected utility framework. The expected returns to entrepreneurship tend to be low on average but exhibit a high variance due to the fact that most startups fail completely and only a few are extremely successful. Hall and Woodward (2010) calculate that, for normal degrees of risk aversion, the very low probability of success and high probability of zero exit value make the expected utility of entrepreneurial ventures negative—meaning that people should prefer not to engage in entrepreneurship. Yet each year, over 500,000 individuals in the United States start firms with at least one employee, and approximately 40 percent of American workers experience at least one period of self-employment during their careers (Parker 2009). Entrepreneurs also seem to persist in running businesses for long periods of time despite either low absolute returns (Hamilton 2000; Åstebro 2003) or returns that appear low after controlling for the highly concentrated illiquid stakes they hold in businesses compared to public equity markets (Moskowitz and Vissing-Jørgensen 2002).

The fact that individuals enter and persist in entrepreneurship despite low risk-adjusted returns suggests that standard theories of risk and return provide an incomplete basis for entrepreneurship and may need to be complemented with richer foundations. That is, while it certainly seems plausible that entrepreneurs have different preferences about risk in a broad sense, there is also the possibility that the standard expected utility model based on objectively known distributions of risk may not capture such differences well. Indeed, widely held popular interpretations of entrepreneurial entry often appeal to behavioral explanations, such as those involving high degrees of risk loving among entrepreneurs who "don't need to be rewarded for risk, because they actually get utility out of risk itself" (Harrington 2010); overconfidence and "endemic optimism" in the startup world (Surowiecki 2014); or entrepreneurs who forgo pecuniary rewards because of the genuine pleasure they obtain from creating and controlling a business (Wasserman 2008).

Drawing on research in behavioral economics, in the sections that follow, we review three sets of possible interpretations for understanding the empirical facts related to the entry into, and persistence in, entrepreneurship. Differences in risk aversion provide a plausible and intuitive interpretation of entrepreneurial activity. In addition, a growing literature has begun to highlight the potential importance of overconfidence in driving entrepreneurial outcomes. Such a mechanism may appear at face value to work like a lower level of risk aversion, but there are clear conceptual differences—in particular, overconfidence likely arises from behavioral biases and misperceptions of probability distributions. Finally, nonpecuniary,

taste-based factors may be important in motivating both the decisions to enter into and to persist in entrepreneurship.

While all these candidate explanations have merit and can account for some aspects of the facts above, there is little evidence of a "smoking gun" that can completely account for all the puzzling patterns we observe. In fact, our reading of the literature suggests that even papers that find evidence consistent with one interpretation are often unable to rule out other mechanisms that are also consistent with their results. Hence, while strong statements on what drives entrepreneurs are widespread in the popular literature, the evidence thus far fails to provide compelling evidence for such a unifying interpretation. Indeed, it is unclear whether a single interpretation that can account for the entire puzzle of entrepreneurial behavior even exists.

A deeper understanding of the roots of entrepreneurship is not only important from a theoretical and academic standpoint, but is also critical for policies addressing entrepreneurship, given the central role that entrepreneurs play in driving productivity growth. In particular, distinguishing the extent to which these patterns are driven by behavioral biases versus preferences is important when thinking about policies that might promote entrepreneurship. Therefore, after reviewing the evidence regarding interpretations for the empirical puzzle, we outline promising avenues for further research.

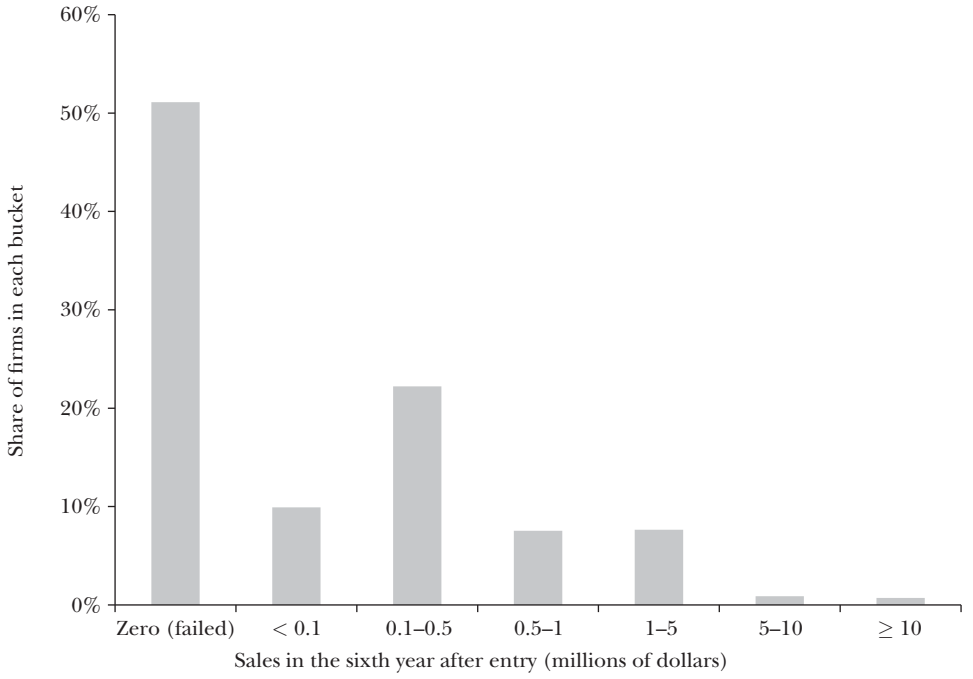
The Entrepreneur's Risk and Return: An Empirical Puzzle

We begin with a set of empirical patterns related to entrepreneurship that create a puzzle when seen through the lens of standard economic models of expected utility and risk.

First, the empirical evidence on returns to entrepreneurship suggests that it tends to be an activity with low median returns but with very high variance—that is, a few entrepreneurs are extremely successful, but the vast majority of entrepreneurs either fail or face meager returns. Figure 1 documents these patterns, both for the broader economy and for the subset of new ventures that are backed by venture capital. Figure 1A is based on data from the Business Information Tracking Series at the US Census Bureau, as documented by Shane (2009). For the 510,654 businesses founded in 1996, Shane (2009) calculates the share of businesses that had either failed or achieved a certain level of annual sales six years later in 2002. Over 50 percent of the businesses had failed within the six years, less than 10 percent achieved more than $1 million in sales, and less than 1 percent had achieved more than $10 million in sales. However, 175 firms or 0.03 percent achieved more than $100 million in sales, making them extremely valuable (and rare) business endeavors. These failure rates are broadly consistent with Kerr and Nanda (2010), who document that 50 percent of all startups founded in the United States between 1976 and 2001 exited within the first four years following entry and 70 percent failed by their tenth year, suggesting there was nothing particularly different about startups founded in 1996.

Figure 1
The Pattern of Failure and Success in Entrepreneurship

A: Sales in 2002 for 510,654 Firms Founded in the United States in 1996

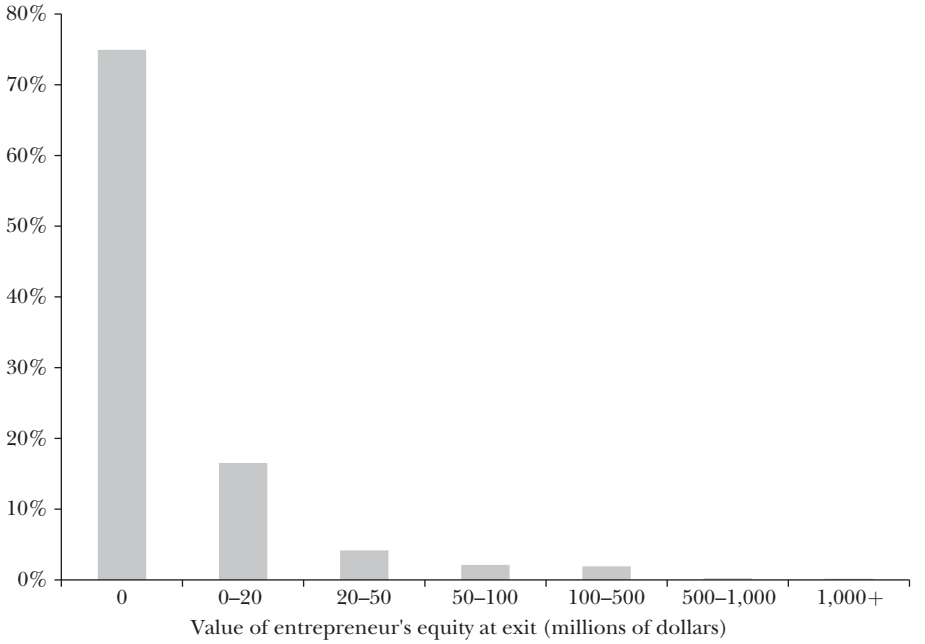


Source: Figure 1A is based on data from the Business Information Tracking Series at the US Census Bureau, as documented by Shane (2009).
Note: Figure 1A shows for the 510,654 US businesses founded in 1996, the share that had either failed or achieved a certain level of sales by 2002.

Figure 1B is based on data from Sand Hill Econometrics, as reported by Hall and Woodward (2010). They analyze the outcomes at exit for the subset of startups founded between 1987 and 2008 in the United States that were financed by venture capital. Venture-capital-backed startups account for under 1 percent of the startups founded each year and typically focus on higher-growth ventures commercializing new technologies or products. The typical contract between the venture capital investor and the entrepreneur involves the entrepreneur taking a below-market salary and a share of the equity. Hall and Woodward find that even for these high-growth ventures, the equity value is zero in almost three-quarters of the startups in their sample. However, a few "billion dollar exits" raise the average value of the entrepreneurs' equity to $5.8 million. Hall and Woodward calculate that, for normal degrees of risk aversion, the very low probability of success and high probability of zero exit value combined with the below market salary makes the expected utility of entrepreneurial ventures presumptively negative—meaning that people should prefer not to engage in entrepreneurship. Because the same

Figure 1 (continued)

B: Value of Entrepreneur's Equity for 22,000 Venture-Capital-Backed Startups Founded between 1987 and 2008



Source: Figure 1B is based on data from Sand Hill Econometrics, as reported by Hall and Woodward (2010).
Note: Figure 1B analyzes outcomes at exit for the subset of startups in the US founded between 1987 and 2008 that were financed by venture capital.

skewed distributions of returns are present for all startups in the United States, and these startups are likely to yield much lower returns than the venture-capital-backed startups, this suggests that Hall and Woodward's (2010) conclusion that people with a normal degree of risk aversion should not become entrepreneurs applies not only to venture-capital-backed entrepreneurial activity, but to entrepreneurial activity more generally.

The expected utility framework does, of course, allow for heterogeneous risk preferences. However, these patterns, when combined with the high frequency of participation in entrepreneurship, suggest that either a sizeable proportion of the population is risk-loving (making it hard to reconcile with other facts about decision making in the general population), or the expected utility framework does not provide a complete characterization of how individuals decide whether to pursue entrepreneurship.

A second dimension of the entrepreneurial puzzle is that not only do many individuals enter despite such low odds of success, but entrepreneurs also seem to persist in running businesses for long periods of time despite either low absolute

returns (Hamilton 2000) or returns that appear low after controlling for the highly concentrated illiquid stakes they hold in businesses compared to public equity markets. For example, Moskowitz and Vissing-Jørgensen (2002) find that entrepreneurial households persistently hold large undiversified stakes in their (mostly private) firms, whose returns are no greater than that of public equity. They find that these "private equity'" investments are at least as volatile and far less liquid than public equity markets, but the returns to these highly undiversified entrepreneurial private equity portfolios are no higher than the returns to public equity. They conclude that the private equity should require a premium of at least 10 percent per year to justify such investment. Furthermore, Åstebro, Jeffrey, and Adomdza (2007) found, using a sample of 820 Canadian inventor entrepreneurs who had sought and paid for assistance from a center originating from the University of Waterloo, that almost one-third continued to spend money and half continued to spend time on projects even after the diagnostic advice from the center advised them to cease; and follow-up data showed little to no value from their further efforts.

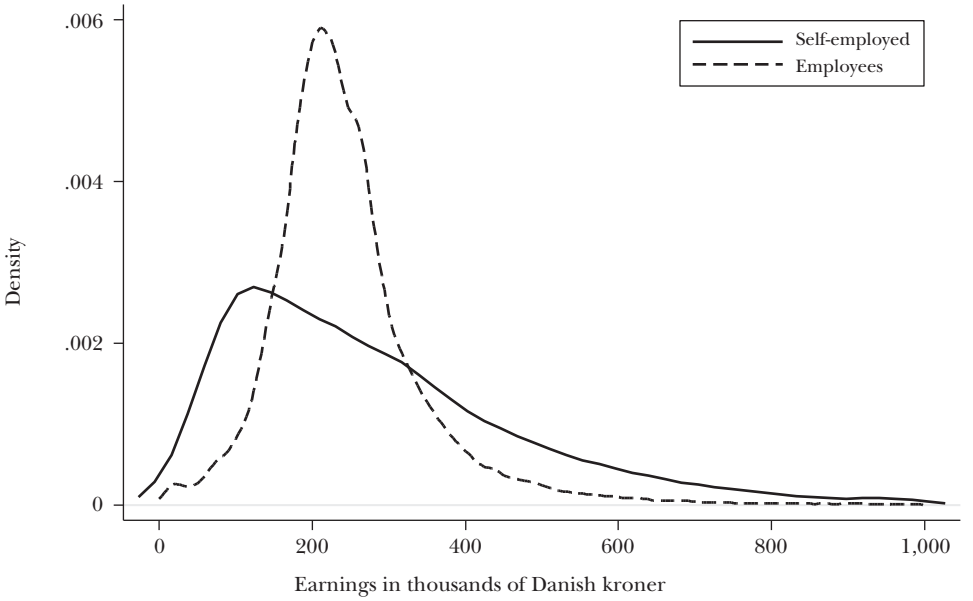
Corroborating evidence of this puzzle is provided in Figure 2, which compares the total earnings of wage employees to those of self-employed individuals, using comprehensive microdata from Denmark. The analysis is based on a 10-percent random sample of all employees and entrepreneurs in 1995, but is then conditioned on individuals whose tenure at their job is at least ten years—in order to compare individuals who would be presumed to have a good match to their job. Similar to Figure 1, Figure 2 documents very high dispersion of earnings among the self-employed, including a large number of individuals whose earnings are lower than that of the typical wage employee. Figure 2 is based on individuals who have been in their job for at least ten years, meaning that this pattern cannot be accounted for purely by lack of time for some entrepreneurs to learn they have low ability and exit (as in Jovanovic 1982). It's true that the comparison in Figure 2 does not control for observable covariates across these groups, and it does not account for sorting based on comparative advantage: thus, it is possible that some of the self-employed who earn less than wage employees are earning the most that they could in either sector. However, Hamilton (2000) finds that, in his sample based on US data, the patterns hold true even when accounting for covariates and for sorting based on sector-specific abilities. The pattern illustrated in Figure 2 seems to suggest the presence of compensating differentials, where some entrepreneurs seem willing to persistently take lower earnings in return for the nonpecuniary benefits associated with self-employment.

Understanding Entrepreneurial Decision Making

Risk Preferences

In the standard expected utility framework, the expected returns to entrepreneurship are determined by the probability distribution over various possible outcomes and the utility obtained from the monetary returns in each of these

Figure 2
Comparison of Wage versus Self-Employment Earnings (Denmark)



Source: Authors using data from the Integrated Database for Labor Market Research (IDA). The database is maintained by the Danish government and consists of an annual panel of all individuals, and firms in Denmark.
Notes: Figure 2 compares the total earnings of wage employees to those of self-employed individuals, using comprehensive microdata from Denmark. The distributions are truncated at 1 million Danish kroner. See text for details.

outcomes. An individual will enter entrepreneurship if this utility assessment is preferable to some alternative occupation, and choose employment otherwise.

Risk preferences are defined by the utility function over wealth in the standard expected utility framework. Most people have utility functions that imply risk aversion, and such people are more willing to take work with regular and less-variable pay. However, a smaller proportion of people—who exhibit less curvature in their utility functions over wealth, and thus less risk aversion—are more likely to be attracted to the possibility of large gains from highly risky ventures such as entrepreneurial activity. Thus, holding constant other factors such as entrepreneurial ability and financing constraints, the individual's preferences over risk can play a critical role in determining the entry decision.

Early models of entrepreneurship attempted to account for entrepreneurial entry within the standard expected-utility framework of economic decision making under risk. For instance, Kihlstrom and Laffont (1979) proposed a theory of entrepreneurship based on differences in risk attitudes, in which optimal risk-sharing between individuals implies that those who are more risk-tolerant become entrepreneurs, while those who are more risk-averse become employees.

A number of empirical studies have attempted to document a difference in risk preferences between entrepreneurs and similar workers who do not start businesses, but the results have been mixed. One approach involves measuring risk taking in other domains of life and using these observations as a proxy for an individual's risk tolerance. Hvide and Panos (2014) look at detailed data on Norwegians who started firms from 2000–2007. They rely on extensive Norwegian government data from tax records that include investment behavior and wealth, as well as on detailed records of all new incorporations in the relevant time period. They show that individuals who participate in the stock market, who invest a higher fraction of their wealth into the stock market, or who have more volatile stock portfolios—presumably, those who possess greater risk tolerance—are more likely to become entrepreneurs. Moreover, Hvide and Panos also find that more risk-tolerant entrepreneurs yield lower-performing firms, measured by number of employees, sales, and profitability. This is consistent with the theoretical prediction that individuals with higher tolerance for risk are willing to enter entrepreneurship in expectation of lower returns, keeping the risk constant.

Another approach to documenting a connection between risk preferences and entrepreneurial entry is to attempt to measure individuals' risk preferences directly. Parker (2009) provides a review of studies comparing such measures between entrepreneur and non-entrepreneur samples. However, no clear picture arises, with some studies pointing toward differences in risk attitudes between the two samples, while others find no such relationship between risk attitudes and entrepreneurship. Many of these studies suffer from small samples and non-incentivized methods of eliciting risk preferences. There also exists the possibility that the samples, often the product of convenience and access, are nonrepresentative of the broader populations of entrepreneurs and non-entrepreneurs.

Other studies use longitudinal data to compare risk attitudes earlier in life with later career paths. Ahn (2010) looks at responses to hypothetical questions about risk that were included in 1993 and in 2002 in the 1979 National Longitudinal Survey of Youth and finds that those who indicate less risk aversion are more likely to become entrepreneurs in the subsequent two years. Cramer, Hartog, Jonker, and Van Praag (2002) use the "Brabant survey" that involved 5,800 Dutch schoolchildren who were originally interviewed and tested in 1952 at the age of 12. In a re-interview in 1993, 1,800 of the original participants answered a hypothetical risk question, and those who had been self-employed at some point in time in the observation period indicated lower degrees of risk aversion. While these studies add value because of their longitudinal nature, they still suffer from the hypothetical nature of the risk elicitation method, and in some instances, the risk measures taken later in life may not match risk preferences at the earlier stage of life.

Further promise for identifying risk attitudes as a driver of entrepreneurial entry comes from experimental economics, which offers tools for incentivized elicitation of individuals' risk preferences. Individuals are confronted with choices between lotteries and certain payments, with real financial consequences, and their profile of choices provides a direct measure of their risk preferences (Holt and

Laury 2002). If risk preferences constitute a stable characteristic of an individual, and these preferences drive entrepreneurship, the application of such methods to eliciting risk preferences for samples of entrepreneurs and non-entrepreneurs could provide evidence that these groups differ in their risk preferences. One study attempting to create such a connection was conducted by Holm, Opper, and Nee (2013). They randomly sampled 700 entrepreneurs heading firms with at least ten employees from local firm registers and 200 control subjects from the Yangzi delta region of China. Both were offered the same incentivized choice menu between various risky and safe outcomes. However, the answers show no difference in preferences towards risk between the entrepreneurs and the control group. When choice menus were offered that involved ambiguity rather than objective risk, the result was the same—again, no difference was found between entrepreneurs and the control group.

In short, the evidence that entrepreneurial entry can be explained by a group of people with very different general risk attitudes than the general population is quite mixed and inconclusive. Some studies suggest that those who start firms are more risk seeking, but others find no association. Indeed, perhaps the most compelling tests from the viewpoint of a critical economist—those in which incentivized elicitation of risk preferences is employed—do not find strong evidence of such entrepreneur versus non-entrepreneur heterogeneity.

Hence, while an interpretation of entrepreneurship as reflecting lower degrees of risk aversion than those in the population remains a potential parsimonious interpretation for some aspects of the puzzle, more evidence is needed before one can conclude that lower risk aversion is, indeed, a primary driver. The evidence on whether entrepreneurs are less risk-averse and whether this preference drives entrepreneurial entry decisions remains, at best, suggestive. Moreover, even if entrepreneurs are less risk-averse than the general population, this finding would not directly imply that entrepreneurs are willing to take the relatively high degrees of risk associated with entrepreneurial entry.

Overconfidence

An alternative explanation that is often proposed to explain entry into entrepreneurship is overconfidence. This explanation implies that individuals enter into entrepreneurship because they subjectively perceive the return distribution too favorably when evaluating their own entrepreneurial project. For example, Cooper, Woo, and Dunkelberg (1988) report that 33 percent of the 3,000 entrepreneurs they surveyed put their odds of success at 10 out of 10, despite putting much lower odds of success for other businesses that were similar to their own. More recently, Shane (2009) reports findings from a Global Entrepreneurship Monitor survey that finds US entrepreneurs report believing it more than five times as likely that they will have at least $10 million in sales than is empirically the case. It has therefore been suggested that those seeking to become entrepreneurs must be imbued with what Adam Smith (1776 [1904], p. 110; see also de Meza and Southey 1996, p. 375) termed "the contempt of risk and the presumptuous hope of success."

Overconfidence may even account for differential patterns of behavior among those who become entrepreneurs. Landier and Thesmar (2009) used survey data collected by Statistics France on a nationally representative sample of French entrepreneurs to construct a measure comparing expectations with future outcomes, which were measured using linked panel data. Those whose expectations exceeded future outcomes were more likely to use short-term debt finance rather than the less-risky option of long-term debt finance.

People often use the general term "overconfidence" to interpret results like those above. However, multiple measures and definitions across empirical studies have made it hard to pin down the precise bias that may be behind entrepreneurship. Moore and Healy (2008) provide a useful distinction between three forms of the general phenomenon of overconfidence. The first concept is *overestimation* of one's ability or performance. The second concept is *overplacement*: individuals assess their skill relative to others as too high. Finally, *overprecision* is the excessive certainty regarding the accuracy of one's beliefs. In addition, one needs to distinguish overconfidence from optimism (Weinstein 1980), which reflects a general view that "good things will happen." Optimism is considered to be a more stable individual trait, not specific to a particular project; to be optimistic is to have generally positively biased expectations. Economists more precisely define an optimist as a person who generally "revises up the probability of favorable events and revises down the probability of unfavorable events" (Hey 1984).

Overestimation, overplacement, and optimism are often observationally equivalent—for example, the above survey evidence from 3,000 entrepreneurs by Cooper, Woo, and Dunkelberg (1988) cannot distinguish between them. However, the underlying psychology is quite different, and the decision environment determines which factors can actually be at work. For example, overplacement requires direct comparisons to a reference group, a feature mainly present in established and contested markets; overestimation applies more broadly to a larger set of situations in which individuals judge their own ability; and optimism indicates a general belief propensity that applies even to situations over which a decision maker has no control. Consequently, the implications of these mechanisms for understanding entrepreneurship and for policy may not be equivalent. Therefore, an understanding of the precise form of overconfidence that might account for the puzzle of excessive entrepreneurial entry requires an ability to distinguish which precise bias drives entrepreneurship, and under what circumstances.

Optimism and overestimation. Researchers have tried to establish a relationship between general optimism and entrepreneurship by measuring optimism in domains of life unrelated to an individual's entrepreneurial skills. Puri and Robinson (2007) constructed such a measure based on data from the Survey of Consumer Finance: specifically, they compared people's own estimates of their life expectancy to what is implied by actuarial tables. They found that more-optimistic people were more likely to be entrepreneurs. They also found that extreme optimists were more likely to make high-risk and even imprudent financial choices. Relatedly, Bengtsson and Ekeblom (2014) use survey data on Swedes' beliefs about

future nationwide economic conditions using responses from 153 monthly surveys conducted between January 1996 and October 2009, again measuring optimism by how expressed beliefs relate to later outcomes. They find that entrepreneurs hold more optimistic beliefs about the general economy, but also that they have lower forecast errors than non-entrepreneurs.

Dawson, de Meza, Henley, and Arabsheibani (2014) used the British Household Panel Study covering 1991–2008 to examine how optimistic forecasts—comparing earnings expectations with future realized outcomes as an employee—predicted performance in subsequent entrepreneurship spells. Since the authors had multiple years of data for individuals as wage earners (on average 5.1 years) they could construct individual fixed-effects estimates of prior optimism net of any environmental influences. The authors also carefully excluded effects from individual ability, which could otherwise co-determine both prior wage earnings—and thus the authors' measure of optimism—and future earnings as an entrepreneur. Dawson et al. found that optimists, on average, earned less than pessimists in entrepreneurship, and that the earnings difference was largest at the top of the earnings distribution and not significant at the bottom.

While optimism and overestimation are often closely related, some work attempts to explicitly differentiate between the two as drivers of entrepreneurial entry. Åstebro, Jeffrey, and Adomdza (2007) compared the behavior of 820 Canadian inventor-entrepreneurs, measuring both overestimation and optimism, with that of a comparable random sample of 300 Canadian citizens. The authors followed a well-established method for measuring *overestimation* by comparing individuals' predictions of their performance to their actual performances on a general knowledge test (Lichtenstein, Fischoff, and Phillips 1982). In addition, the authors measured optimism as reflecting a person's general view that good things will happen (Weinstein 1980). The survey data showed that inventors tended toward both more overestimation and optimism than the comparison group. However, the overestimation measure was not significantly related to increased expenditures of time and money, while entrepreneurs with greater levels of optimism were more likely to keep pursuing an idea even with little chance of success, and thus to incur higher losses.

The controlled decision environments provided by incentivized experiments make them useful to further assess the relevance of overestimation and optimism. Studying the behavior of students and executives with entrepreneurial experience in laboratory experiments, Åstebro, Mata, and Santos-Pinto (2014) employed an experimental design in which success probabilities are exogenously determined and known by subjects, in order to rule out overestimation of own skill as a driver of behavior. Their findings suggest that general optimism, rather than convex utility, drives what appears to be a preference for the kinds of skewed lotteries that characterize entrepreneurship.

Overplacement. Overplacement is different from overestimation and optimism in that it refers to a direct comparison of own skill to competitors. Consequently, overplacement may be a particularly valid explanation for entrepreneurial entry into contested markets, where one could have a biased belief in the likelihood of

coming out ahead of the competition. Early evidence suggesting such a relationship between overplacement and market entry came from an experiment by Camerer and Lovallo (1999). In their experiment, students who were undergraduates or MBAs at either the University of Chicago or the University of Pennsylvania could earn money by entering a "market" where payoffs depend on their rank among all entrants. In the baseline condition, ranks were assigned randomly, but in a skill condition, subjects were told they would be ranked according to their relative performance in a trivia quiz. Camerer and Lovallo found that significantly more subjects entered the market in the skill condition. This excessive entry took place despite the fact that subjects correctly predicted that there would be excessive entry in the skill condition (but not in the baseline condition). The authors concluded that although subjects expect excessive entry, they are willing to enter the market because they hold a biased belief that they are among the most skilled, which makes entry appear profitable in expectation.

The study by Holm, Opper, and Nee (2013) of Chinese entrepreneurs and control subjects discussed in the previous section also provides evidence suggesting a possible relationship between overplacement and entrepreneurial market entry. Their entrepreneur and control subjects participated in a market entry task similar to the one used by Camerer and Lovallo (1999). Holm et al. find that the entrepreneurs were more willing to enter competitive environments, in which success depended on own skill, than the control group. However, the entrepreneurs did not, on average, overplace themselves in expected performance compared to the control group. Hence, the connection between overplacement and entrepreneurship in this study is imperfect and raises the possibility that the entrepreneurs possess a preference for competition per se, rather than biased beliefs about their relative abilities.

Overprecision. While overplacement, overestimation, and optimism all lead to positively biased perceptions of expected returns and hence should foster entrepreneurial entry, the effects of overprecision are less clear. Herz, Schunk, and Zehnder (2014) look at the effect of overprecision on the trade-off between exploration and exploitation, one of the key features of the innovative process that also underlies entrepreneurial activity (Kerr, Nanda, and Rhodes-Kropf, in this symposium). They argue that overprecision, the tendency to underestimate variance of own information, can reduce the perceived option value of exploration and therefore actually reduce incentives to engage in entrepreneurship. Herz, Schunk, and Zehnder experimentally tested these predictions with students and business managers. Subjects participated in an incentivized individual decision-making task in which they had to manage a virtual ice cream stand and repeatedly make decisions over the offered product mix.[1] Subjects faced an overall choice about tweaking a preexisting strategy or trying brand-new strategies to maximize profits. Overprecision was then measured in an independent task using an established method in which

[1] The experiment is adopted from Ederer and Manso (2013), who use a similar task to study the effect of different incentive schemes on the exploration–exploitation trade-off.

individuals state 90 percent confidence intervals for ten trivia questions—that is, such that they are 90 percent certain that the correct answer is contained in the interval (Lichtenstein et al. 1982). Subjects who are overconfident in precision typically provide overly narrow ranges, so that actual values fall outside the range more than 10 percent of the time. Herz, Schunk, and Zehnder find that overprecision is indeed negatively related to experimentation and realized profits, suggesting the possibility that some forms of overconfidence may lead to a bias away from the type of exploration that is central to entrepreneurship.

Taken together, some evidence suggests overconfidence in the form of optimism, overestimation, and overplacement could help explain entrepreneurial entry. Perhaps the strongest support comes from correlational evidence between broad measures of optimism and entrepreneurship. However, even if one believes these correlations reflect a causal relationship, the precise nature of overconfidence driving the relationship is not well understood. Since different forms of overconfidence may differentially impact entrepreneurial decision making, more work is required to better understand the precise type of overconfidence that affects entrepreneurship and how it does so.

Moreover, many open questions remain. For starters, the studies above measure overconfidence using measures unrelated to the domain of entrepreneurship; that is, measures of overconfidence and optimism mostly stem from independent measurements, assuming that these are personality traits that apply generally. This assumption may have some merit, but it would be valuable to have more detailed measures of different forms of overconfidence and optimism directly relating to entrepreneurial activity. For example, in contexts other than entrepreneurial entry, some headway has been made in this direction by Malmendier and Tate (2005a; 2005b) by measuring chief executive officers' overconfidence as continuing to hold stock options in their own firms after the options are fully vested.

Finally, several researchers note that behavior that appears to result from overconfidence may often also have rational, Bayesian interpretations (Benoît and Dubra 2011; Manso 2013; Van den Steen 2004). For example, assessing yourself to be above average is only a bias for those below the average, which may be a small proportion of the population. In addition, if your knowledge about the performance of your comparison group is low, it may make sense to place yourself above it.

Nonpecuniary Benefits

The above interpretations primarily address the observation of too much entry by entrepreneurs. As we note earlier, there is also mounting evidence of persistence in entrepreneurship despite the low average returns from entrepreneurial effort discussed earlier and the availability of more attractive alternative occupations. Hence, if misperceptions of success probabilities drive entrepreneurship, why aren't such initial misperceptions corrected by experience?

One parsimonious interpretation for both entry and persistence in entrepreneurship is the possibility that entrepreneurs receive nonpecuniary benefits from their self-employment. When authors like Hamilton (2000) and Moskowitz and

Vissing-Jørgensen (2002) point out the low average returns for entrepreneurship, as discussed earlier in this paper, they also advance the possibility that nonstandard preferences for autonomy and control could be potential explanations for individuals' inclination to become entrepreneurs as well as their persistence in entrepreneurship. Job characteristics that standard economic theories typically view as a means to obtaining higher pecuniary rewards—such as decision rights and control—may, for some people, be inherently valuable ends themselves.

People with such preferences may be lured to entrepreneurship by the promise of these job characteristics even though earnings may be lower. For instance, Frey, Benz, and Stutzer (2004) argue that independence and autonomy at work are sources of "procedural utility," which raise happiness. In their discussion of small firms in the US economy, Hurst and Pugsley (2011) point out that most start small and remain small, with no new technology and no intention of growing. Many of them are small service firms: lawyers, skilled craftsmen, real estate agents, restaurateurs, and the like. Based on survey evidence from the Panel Study of Entrepreneurial Dynamics, a nationally representative sample of 34,000 individuals during the fall of 2005 and the early winter of 2006, and the Kauffman Firm Survey, a panel study of 4,928 businesses that were newly founded in 2004, Hurst and Pugsley find that, for these firms, entrepreneurs claim nonpecuniary benefits as a first-order motive for self-employment.

However, the precise nature of these nonpecuniary benefits has remained largely unclear. For example, entrepreneurs work longer hours than the average employee. In only two of the 25 OECD countries (Russia and Chile) do the employed work longer hours than the self-employed, and the self-employed tend to work, on average, between 2 and 14 more hours (that is, 5–35 percent more) per week (Åstebro and Chen 2014). Thus, the nonpecuniary benefits do not simply reflect a preference for leisure. One possible source of nonpecuniary benefits is the autonomy and independence that an entrepreneur enjoys in allocating personal work time. More broadly, an entrepreneur can exercise control over the company and need not worry about interference by other parties. The importance of factors such as "control over one's life" and a "sense of purpose" is documented in studies that are not focused on entrepreneurship but instead study hypothetical choices in relation to predicted subjective levels of happiness over varied contexts (for example, Benjamin, Heffetz, Kimball, and Rees-Jones 2012). Other nonpecuniary benefits from self-employment may arise from the pride in bringing one's own business idea to market success or from a taste for variety. Addressing the latter case, Åstebro and Thompson (2011) surveyed 820 Canadian entrepreneurs who sought assistance at the Canadian Innovation Center at the University of Waterloo and compared their responses to those of a matched sample of 300 Canadian non-entrepreneurs. They find that those who have been entrepreneurs tend to be those whose reported behavior suggests a taste for variety, for instance they have varied labor market experience.

There have been attempts to infer the nonpecuniary benefits from self-employment; typically, this involves looking at the lower average returns

earned by entrepreneurs, adjusting for other factors, and then noting that the nonpecuniary benefits must be large enough to offset this difference. Thus, as we note earlier, Hamilton (2000) estimates a median net present value lifetime earnings differential of 35 percent for individuals in business for ten years. Moskowitz and Vissing-Jørgensen (2002) find that nonpecuniary benefits of self-employment may be as large as 143 percent of total annual income. Åstebro and Thompson (2011) find the size of the nonpecuniary benefits of having a large variety of skills in entrepreneurship is on the order of 16 to 22 percent of annual household income.

Obviously, it is almost impossible to infer the nature and magnitude of these nonpecuniary benefits directly from these data. Usually, these benefits are private, and the measured earnings differentials can only provide a crude approximation. Experimental studies thus provide useful complementary evidence in this regard. While such studies are usually restricted to the study of non-entrepreneurial samples and the situations under consideration are rather artificial, they can advance our understanding of these nonpecuniary motives by demonstrating a preference for keeping control over decisions, shedding light on the motivations underlying this behavior and highlighting their potential importance for understanding entrepreneurship.

For example, Cooper and Saral (2013) ran experiments with 184 subjects—a mixture of undergraduates, business school students, entrepreneurs, and other business people—who performed the task of answering questions from the Graduate Management Aptitude Test (GMAT). In the study, subjects decided whether to work alone or in groups. Entrepreneurs showed a greater willingness to pay for working independently, despite monetary incentives to the contrary. In a post-experimental questionnaire, many subjects who revealed a preference for working alone indicated a fear of loss of control or a preference for self-reliance.

Other recent experiments look at the underlying motivations that lead individuals to cede or retain control. Fehr, Herz, and Wilkening (2013) study a structured interaction in which principals must decide whether to delegate decision rights to agents in a situation of incomplete information. Using a subject pool of 504 university students in Zurich, they conducted a laboratory experiment showing that individuals hold on to decision rights in situations in which rendering control would clearly be preferable for all involved parties in terms of expected monetary value. This behavior does not diminish with experience, and appears to be driven by regret aversion: if subjects delegate decision rights but discover later that they would have been better off keeping them, they display strong negative reactions. Owens, Grossman, and Fackler (forthcoming) provide further experimental evidence for inefficient holding on to control. In their study, subjects must choose between an asset that will pay off if they answer a question correctly or an asset that will pay off if their partner answers a question correctly. Results over 108 students show that individuals are willing to sacrifice 8 to 15 percent of expected earnings in exchange for control over their payoff. Similar considerations regarding an inherent value of authority and control may drive an entrepreneur's decision to remain in entrepreneurship even when doing so is unprofitable from a material perspective.

Finally, Bartling, Fehr, and Herz (forthcoming) conducted an experiment that measures individuals' intrinsic valuation of decision rights. Subjects participate in a game in which a principal or an agent can make decisions that have monetary consequences for both. Principals reveal indifference between their own decision and a specific decision by the agent, and these decisions define two lotteries. If decision rights carry no intrinsic value, the certainty equivalents of both lotteries must be the same. This is tested by again presenting the lotteries to the principals but simply as given lotteries over outcomes that are not the result of anyone's decision. Differences in elicited certainty equivalents would therefore reflect the direct impact of decision rights on utility. In a group of 172 students at the University of Zurich, Bartling, Fehr, and Herz identify an average compensating differential of 16.7 percent for letting the agent decide. Hence, unlike with evidence from the field, where the inferred compensating differentials for entrepreneurship may include many components, in this experimental study, a precise value is given to one specific dimension: personal control over decision making.

Taken together, evidence from the field—specifically, the observed compensating differentials and the complementary survey evidence—strongly suggest that nonpecuniary benefits may play an important role in the decision to become and remain an entrepreneur. Experimental evidence, which allows studying the determinants of human behavior more directly, also highlights mechanisms that may underlie such nonpecuniary benefits of entrepreneurship. Entrepreneurs may be those who like to work independently and not rely on others, and control appears to be inherently valuable.

However, while the evidence above suggests an important role for nonpecuniary utility from entrepreneurial activity, our view of the literature is that a gap remains in identifying the importance of these considerations for explaining the entry and persistence puzzle. Much more needs to be understood about the importance and precise nature of nonpecuniary factors for driving entrepreneurship. For example, the evidence by Hurst and Pugsley (2011) indicates that individuals primarily motivated by nonpecuniary factors do not necessarily sort into high-growth sectors and, instead, are satisfied by consuming desirable job characteristics in low-growth sectors. This suggests that preferences for autonomy and control may not only drive the decision to become an entrepreneur but also the kinds of businesses that entrepreneurs pursue. The relevance of nonpecuniary benefits in explaining entrepreneurship in different sectors of the economy is, therefore, a promising possible interpretation but one for which more research is necessary.

New Frontiers in Behavioral Entrepreneurship

Clearly, behavioral interpretations of the drivers of entrepreneurship are potentially valuable in accounting for the entry and persistence puzzle. However, none of the interpretations stands out as the primary factor, and there is little evidence on how much of the behavior of actual entrepreneurs is accounted for by any of the

mechanisms. There is plenty of circumstantial evidence for each of the possible explanations, but no "smoking gun." Indeed, reviewing the evidence on the roots of entrepreneurship, what surprises us most is how little we really know.

In this regard, it is important that research continue to pursue an understanding of the role the above mechanisms, as well as others, play in entrepreneurship. Indeed, research on the behavioral roots of entrepreneurship is proceeding on many fronts. To conclude this article, we lay out three directions that seem especially promising and important.

First, much of the research on entry into entrepreneurship has tended to focus on single factors—such as risk preferences, overconfidence, or nonpecuniary benefits. The time is ripe to compare and contrast these factors since the evidence is often consistent with multiple candidate explanations. For example, the data show that the relationship between personal wealth and entrepreneurship is flat for most of the wealth distribution but rises sharply above the 80th percentile of wealth and is steepest for the wealthiest 5 percent of the population (Hurst and Lusardi 2004). Such evidence is consistent with risk preferences as a driving factor for entry—risk aversion may be decreasing in wealth—but it is also consistent with entrepreneurship being a luxury good that is consumed more by wealthier individuals. Of course, it is also consistent with the presence of financing constraints in entrepreneurship—again, highlighting the necessary caution that must accompany any interpretation of this fact.

Research into the roots of entrepreneurship could also benefit from richer data that allows disentangling different interpretations. For example, to differentiate preference-based explanations from overconfidence and optimism, valuable insights might be gained from a detailed panel study comparing entrepreneurs' assessments of and motivations for becoming entrepreneurs both before they begin and afterwards. Do they regret their entry decision? Such surveys may suffer from after-the-fact justifications of own choices and from hindsight bias but can nonetheless shed further light on the relevance of the different factors in driving entry into entrepreneurship.

Progress on quantifying the relevance of the candidate explanations also requires reliable and precise measurements so that research can cumulatively build towards a consensus. Currently, the most precise measurements often take place in laboratory settings, implicitly assuming that more general measures of optimism, overconfidence, and preferences are stable and generalizable across contexts. Identification that is directly linked to entrepreneurial activity is likely to yield more insight into the mechanisms driving entry. Furthermore, just as many factors may account for the puzzle we outline, it also seems conceivable that different factors may account for various sub-segments of entrepreneurship. While small business owners may mainly be motivated by nonpecuniary benefits, entrepreneurship at the technological frontier may be better explained by overconfidence or risk preferences. Understanding such heterogeneous motivations, where they apply, and how they may interact seems critically important for understanding entrepreneurial entry and persistence.

Second, it is important to understand how individuals' perceptions of entrepreneurship are formed and shaped. Optimism, overplacement, and overestimation manifest themselves in exaggerated beliefs about the profitability of entrepreneurial activity. Such beliefs about the returns to entrepreneurship may be shaped by social context, including role models or peers. For example, there appears to be intergenerational correlation in entrepreneurship and self-employment (Dunn and Holtz-Eakin 2000). Lindquist, Sol, and van Praag (forthcoming) show, using data from biological children and adoptees, that nonbiological factors as well as biological factors indeed contribute to this association but that post-birth factors play a more important role. Using census data from Denmark, Nanda and Sørensen (2010) provide further evidence on the relevance of social factors, finding that individuals are more likely to become entrepreneurs if their coworkers have previously been entrepreneurs (see also Giannetti and Simonov 2009). Lerner and Malmendier (2013) also find that peers shape entry, but their evidence is that of reduced entry into entrepreneurship among MBA students whose peers had entrepreneurial experience. In addition, Lerner and Malmendier find that having entrepreneurial peers reduces the likelihood of starting firms that fail. This latter work suggests that it may be possible to reduce potential distortions in the expected distribution of the returns to entrepreneurship through exposure to those who have experienced it first-hand. Preference-based explanations may also have their root in an individual's social environment. For instance, persistence in entrepreneurship may be influenced by different cultural perceptions of the stigma of failure.

Finally, the societal implications of "excessive entry" need to be better understood. There is a widely held belief that some entrepreneurs generate substantial positive externalities, and excessive entry may be central to the process of creative destruction. For example, it has been estimated that probably well in excess of 90 percent of the benefits of breakthrough innovation go to society as a whole rather than to the individual inventor, their partners, or their financial backers (Baumol 2002; Nordhaus 2004). In this sense, perhaps excessive entry is a blessing for society. Understanding whether excessive entry is in fact welfare enhancing due to these externalities is therefore important for guiding policy. Some initial steps in understanding the interplay of behavioral biases and welfare have been made by Bernardo and Welch (2001), who use an evolutionary model to show equilibrium persistence of overconfident entrepreneurs. They assume that overconfident entrepreneurs have too much confidence in their private information—that is, they suffer from overprecision—and are therefore less likely to imitate their peers. While such overconfidence is harmful to the overconfident entrepreneurs, Bernardo and Welch show that it can be welfare enhancing for society.

Whether excessive entry is indeed optimal from a societal perspective may critically depend on the actual drivers of entry and the sector in which excess entry takes place. For example, while overestimation and optimism may trigger the pursuit of breakthrough innovations with strong positive externalities, overplacement may primarily lead to entry into already contested markets, and it may be associated more with imitation. Hence, while overplacement may still indirectly foster

innovation through increased competition within a market, the potential positive externalities are much less clear.

Similarly, entrepreneurs with strong nonpecuniary motivations may mainly sort into low-growth, non-inventive small businesses. Consequently, policies subsidizing entrepreneurship must be examined to determine the extent to which they provide positive externalities for the economy as opposed to only providing consumption value to the entrepreneur. Policies that simply favor small businesses, which are common in the US and other countries, may not be optimal from a societal viewpoint (Hurst and Pugsley 2011). In particular, the potential prevalence of nonpecuniary benefits as the main driver into small-scale entrepreneurship may call for stricter targeting of policy to foster high-growth industries at the technological frontier.

A comprehensive account of entrepreneurial decision making is likely to include both behavioral and nonbehavioral elements. In our view, behavioral research has not yet provided definitive explanations for puzzling aspects of entrepreneurship. Rather, the real promise of behavioral research lies in the potential for future insights that integrate and enlighten our understanding of this important dimension of economic activity.

■ *We thank David Autor, Chang Tai Hsieh, Ulrike Malmendier, and Timothy Taylor for editorial guidance. Åstebro acknowledges financial support from the HEC Foundation. Herz acknowledges financial support from the European Research Council grant on the Foundations of Economic Preferences. Nanda thanks Jesper B. Sørensen, Niels Westergård-Nielsen, Arhus University, and Statistics Denmark for providing access to the Danish microdata and the Kauffman Foundation and the Division of Research and Faculty Development at Harvard Business School for financial support. Weber is very grateful to Steven Klepper for several discussions that helped shape some of the arguments in this article. This paper has used microdata from the Integrated Database for Labor Market Research in Denmark. Figure 2 has been screened to ensure no confidential data are revealed.*

References

Ahn, Taehyun. 2010. "Attitudes toward Risk and Self-employment of Young Workers." *Labour Economics* 17(2): 434–442.

Åstebro, Thomas. 2003. "The Return to Independent Invention: Evidence of Risk Seeking, Extreme Optimism or Skewness-Loving?" *Economic Journal* 113(484): 226–39.

Åstebro, Thomas, and Jing Chen. 2014. "The Entrepreneurial Earnings Puzzle: Mismeasurement or Real?" *Journal of Business Venturing* 29(1): 88–105.

Åstebro, Thomas, Scott A. Jeffrey, and Gordon Adomdza. 2007. "Inventor Perseverance after Being Told to Quit: The Role of Cognitive Biases." *Journal of Behavioral Decision Making* 20(3): 253–72.

Åstebro, Thomas, José Mata, and Luís Santos-Pinto. 2014. "Skewness Seeking: Risk Loving, Optimism or Overweighting of Small Probabilities?" *Theory and Decision*. Published online January 29.

Åstebro Thomas, and Peter Thompson. 2011. "Entrepreneurs: Jacks of all Trades or Hobos?" *Research Policy* 40(5): 637–49.

Bartling, Björn, Ernst Fehr, and Holger Herz. Forthcoming. "The Intrinsic Value of Decision Rights." *Econometrica*.

Baumol, William. 2002. *The Free-Market Innovation Machine: Analyzing the Growth Miracle of Capitalism*, p. 133–35. Princeton University Press.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones. 2012. "What Do You Think Would Make You Happier? What Do You Think You Would Choose?" *American Economic Review* 102(5): 2083–2110.

Bengtsson, Ola, and Daniel Ekeblom. 2014. "The Bright but Right View? A New Type of Evidence on Entrepreneurial Optimism." IFN Working Paper 1008, February 19.

Benoît, Jean Pierre P., and Juan Dubra. 2011. "Apparent Overconfidence." *Econometrica* 79(5): 1591–1625.

Bernardo, Antonio E., and Ivo Welch. 2001. "On the Evolution of Overconfidence and Entrepreneurs." *Journal of Economics & Management Strategy* 10(3): 301–30.

Camerer, Colin, and Dan Lovallo. 1999. "Overconfidence and Excess Entry: An Experimental Approach." *American Economic Review* 89(1): 306–18.

Cooper, Arnold C., Carolyn Y. Woo, and William C. Dunkelberg. 1988. "Entrepreneurs' Perceived Chances for Success." *Journal of Business Venturing* 3(2): 97–108.

Cooper, David J., and Krista Jabs Saral. 2013. "Entrepreneurship and Team Participation: An Experimental Study." *European Economic Review* 59(C): 126–40.

Cramer, J. S., J. Hartog, N. Jonker, and C. M. Van Praag. 2002. **"**Low Risk Aversion Encourages the Choice for Entrepreneurship: An Empirical Test of a Truism." *Journal of Economic Behavior & Organization* 48(1): 29–36.

Dawson, Christopher, David de Meza, Andrew Henley, and G. Reza Arabsheibani. 2014. "The Power of Non-Positive Thinking: Self-Employed Optimists Earn Less than Self-Employed Realists." Unpublished paper, University of Bath.

De Meza, David, and Clive Southey. 1996. "The Borrower's Curse: Optimism, Finance and Entrepreneurship." *Economic Journal* 106(435): 375–86.

Dunn, Thomas, and Douglas Holtz-Eakin. 2000. "Financial Capital, Human Capital, and the Transition to Self-Employment: Evidence from Intergenerational Links." *Journal of Labor Economics* 18(2): 282–305.

Ederer, Florian, and Gustavo Manso. 2013. "Is Pay for Performance Detrimental to Innovation?" *Management Science* 59(7): 1496–1513.

Fehr, Ernst, Holger Herz, and Tom Wilkening. 2013. "The Lure of Authority: Motivation and Incentive Effects of Power." *American Economic Review* 103(4): 1325–59.

Frey, Bruno S., Matthias Benz, and Alois Stutzer. 2004. "Introducing Procedural Utility: Not Only What, But Also How Matters." *Journal of Institutional and Theoretical Economics* 160(3): 377–401.

Giannetti, Mariassunta, and Andrei Simonov. 2009. "Social Interactions and Entrepreneurial Activity." *Journal of Economics & Management Strategy* 18(3): 665–709.

Hall, Robert E., and Susan E. Woodward. 2010. "The Burden of the Nondiversifiable Risk of Entrepreneurship." *American Economic Review* 100(3): 1163–94.

Hamilton, Barton H. 2000. "Does Entrepreneurship Pay? An Empirical Analysis of the Returns to Self-Employment." *Journal of Political Economy* 108(3): 604–31.

Harrington, Michael. 2010. "Are You a Pirate." Techcrunch, Oct. 31. http://techcrunch.com/2010/10/31/are-you-a-pirate.

Herz, Holger, Daniel Schunk, and Christian Zehnder. 2014. "How Do Judgmental Overconfidence and Over-Optimism Shape Innovative Behavior?" *Games and Economic Behavior* 83: 1–23.

Hey, John Denis. 1984. "The Economics of Optimism and Pessimism." *Kyklos* 37(2): 181–205.

Holm, Hakan J., Sonja Opper, and Victor Nee. 2013. "Entrepreneurs under Uncertainty: An Economic Experiment in China." *Management Science* 59(7): 1671–87.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92(5): 1644–55.

Hurst, Erik, and Annamaria Lusardi. 2004. "Liquidity Constraints, Household Wealth, and Entrepreneurship." *Journal of Political Economy* 112(2): 319–47.

Hurst, Eric, and Benjamin Pugsley. 2011. "What Do Small Businesses Do?" *Brookings Papers on Economic Activity* 43(2): 73–142.

Hvide, Hans K., and Georgios A. Panos. 2014. "Risk Tolerance and Entrepreneurship." *Journal of Financial Economics* 111(1): 200–23.

Jovanovic, Boyan. 1982. "Selection and the Evolution of Industry." *Econometrica* 50(3): 649–70.

Kerr, William R., and Ramana Nanda. 2010. "Banking Deregulations, Financing Constraints and Firm Entry Size." *Journal of the European Economic Association* 8(2–3): 582–93.

Kihlstrom, Richard E., and Jean-Jacques Laffont. 1979. "A General Equilibrium Entrepreneurial Theory of Firm Formation based on Risk Aversion." *Journal of Political Economy* 87(4): 719–48.

Knight, Frank H. 1921. *Risk, Uncertainty, and Profit.* Houghton Mifflin, Boston.

Landier, Augustin, and David Thesmar. 2009. "Financial Contracting with Optimistic Entrepreneurs." *Review of Financial Studies* 22(1): 117–50.

Lerner, Josh, and Ulrike Malmendier. 2013. "With a Little Help from My (Random) Friends: Success and Failure in Post-Business School Entrepreneurship." *Review of Financial Studies* 26(10): 2411–52.

Lichtenstein, Sarah, Baruch Fischhoff, and Lawrence D. Phillips. 1982. "Calibration of Probabilities: The State of the Art to 1980." In *Judgement under Uncertainty: Heuristics and Biases*, edited by D. Kahneman, P. Slovic, & A. Tversky, pp. 306–34. Cambridge University Press.

Lindquist, Matthew J., Joeri Sol, and Mirjam van Praag. Forthcoming. "Why Do Entrepreneurial Parents Have Entrepreneurial Children?" *Journal of Labor Economics.*

Malmendier, Ulrike, and Geoffrey Tate. 2005a. "CEO Overconfidence and Corporate Investment." *Journal of Finance* 60(6): 2661–2700.

Malmendier, Ulrike, and Geoffrey Tate. 2005b. "Does Overconfidence Affect Corporate Investment? CEO Overconfidence Measures Revisited." *European Financial Management* 11(5): 649–59.

Manso, Gustavo. 2013. "Experimentation and the Returns to Entrepreneurship." Unpublished paper.

Moore, Don A., and Paul J. Healy. 2008. "The Trouble with Overconfidence." *Psychological Review* 115(2): 502–17.

Moskowitz, Tobias J., and Annette Vissing-Jørgensen. 2002. "The Returns to Entrepreneurial Investment: A Private Equity Premium Puzzle?" *American Economic Review* 92(4): 745–78.

Nanda, Ramana, and Jesper B. Sørensen. 2010. "Workplace Peers and Entrepreneurship." *Management Science* 56(7): 1116–26.

Nordhaus, William D. 2004. "Schumpeterian Profits in the American Economy: Theory and Measurement." NBER Working Paper 10433.

Owens, Dave, Zachary Grossman, and Ryan Fackler. Forthcoming. "The Control Premium: A Preference for Payoff Autonomy." *American Economic Journal: Microeconomics.*

Parker, Simon C. 2009. *The Economics of Entrepreneurship.* Cambridge University Press.

Puri, Manju, and David T. Robinson. 2007. "Optimism and Economic Choice." *Journal of Financial Economics* 86(1): 71–99.

Shane, Scott A. 2009. *Fool's Gold? The Truth Behind Angel Investing in America.* Oxford University Press.

Smith, Adam. 1776 [1904]. *An Inquiry into the Nature and Causes of the Wealth of Nations.* Edited by Edwin Cannan. London: Methuen & Co.

Surowiecki, James. 2014. "Epic Fails of the Startup World." *New Yorker,* May 19. http://www.newyorker.com/talk/financial/2014/05/19/140519ta_talk_surowiecki.

Van den Steen, Eric. 2004. "Rational Overoptimism (and Other Biases)." *American Economic Review* 94(4): 1141–51.

Wasserman, Noam. 2008. "The Founder's Dilemma." *Harvard Business Review,* February, 103–109.

Weinstein, Neil D. 1980. "Unrealistic Optimism about Future Life Events." *Journal of Personality and Social Psychology* 39(5): 806–820.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 71–88

The Lewis Model: A 60-Year Retrospective

Douglas Gollin

The Lewis (1954) model of economic development is one of the rare 60 year-old papers still featured on many graduate economics reading lists. However, like many classics, the original paper, "Economic Development with Unlimited Supplies of Labor," is probably read less frequently than it is cited. There are numerous commentaries and glosses available that save readers the trouble of wrestling with the original—even though Lewis's writing was lucid and engaging. In particular, many development textbooks offer verbal or graphic summaries of the Lewis model, but the summaries often lose the richness of the original.

Lewis's (1954) paper bundles together theories of growth, structural transformation, inequality and distribution, wage determination, and population. The proliferation of ideas in the Lewis paper was not an accident. Lewis (p. 139) telegraphed his intention in the first paragraph of the paper, where he wrote: "This essay is written in the classical tradition, making the classical assumption, and asking the classical question. The classics, from Smith to Marx, all assumed, or argued, that an unlimited supply of labour was available at subsistence wages. They then inquired how production grows through time. They found the answer in capital accumulation, which they explained in terms of their analysis of the distribution of income. Classical systems thus determined simultaneously income distribution and income growth, with the relative prices of commodities as a minor by-product." This paper was not a modest undertaking, and Lewis himself clearly viewed it as a major contribution.

■ *Douglas Gollin is Professor of Development Economics, Oxford University, Oxford, United Kingdom. His email address is douglas.gollin@qeh.ox.ac.uk.*

http://dx.doi.org/10.1257/jep.28.3.71 doi=10.1257/jep.28.3.71

For Lewis, the target of his analysis was a set of countries that were not only "underdeveloped" but also "overpopulated," a term that he seems to have taken as effectively equivalent to his notion of "unlimited supplies of labour." One very specific interpretation of overpopulation is the idea of a large population relative to fixed factors or natural resource endowments, such as land; Lewis invokes this idea a number of times in the model. But in Lewis's argument, overpopulation also seems related to ideas of underemployment and low labor force participation. This notion of overpopulation seems somewhat problematic to contemporary economists, in the sense that population, employment, and labor force participation all seem endogenous—a concern to which this paper will return. But Lewis sought to identify a distinctive process of development in overpopulated countries. He understood that this theory of development would not be applicable to other countries—including those that had already transitioned to "capitalist" production or those that had abundant natural resources and thus no real Malthusian pressures on a subsistence sector.

The Lewis model is built on the idea of a dual economy. For the "overpopulated" countries which are the focus of the essay, Lewis argued that the central process of development consists of moving a large mass of underemployed workers, with low productivity (in Lewis's terms, workers whose marginal product is "negligible, zero, or even negative," p. 141), out of a "subsistence" sector, where living standards are necessarily low, into a modern "capitalist" sector, where output per worker can be higher because it is "fructified by capital" (p. 147). In this framework, growth consists, in its simplest form, of expanding the capitalist sector. This expansion requires an increase in savings, which can only come from the capitalist sector or from external sources. As capital flows into the economy, it is used to create jobs in the modern sector, which in turn can always be filled by workers from the subsistence sector. As these workers move, the savings rate of the economy rises, and this in turn leads to a virtuous circle that steadily raises the level of income per worker in the economy.

Lewis was somewhat vague on the theoretical underpinnings of the model. Some ingredients were clear, however. Lewis envisioned a capitalist sector that was, at least in the early stages of development, sufficiently small to be a price taker on the labor market. He assumed that the supply of capital was fixed in the short run and could only be used in the capitalist sector. This assumption implicitly requires a market failure or perhaps a technological barrier, and it is the key ingredient in forcing the economy to include two distinct sectors. The labor-surplus subsistence sector determines the wage rate of the economy, at least in early stages of development. At the prevailing wage (or more precisely at a modest premium above the subsistence sector wage, discussed below), the capitalist sector hires labor up to the point where its marginal value product is equated with the wage. The remainder of the labor force remains in the subsistence sector.

At first glance, the capitalist sector appears nicely neoclassical, with a fixed supply of capital and a variable labor input hired at a given wage—behaving, in fact, like an individual firm in a standard introductory micro framework. But this

seemingly neoclassical equilibrium actually rests on a dualistic framework that is imposed by assumption. The critical point is that the subsistence sector cannot make any productive use of capital. Without this assumption, the capitalists of this economy would surely be inclined to use some small portion of their capital to "fructify" part or all of the subsistence sector. The marginal returns to capital would presumably be very high for the large numbers of workers in this sector, and market incentives should encourage capital to flow from the capitalist sector to the subsistence sector. Lewis's explanation for this assumption seems to invoke an essential lumpiness of capital. He writes (p. 145): "If unlimited labour is available, while capital is scarce, we know . . . that the capital should not be spread thinly over all the labour." Given this assumption, then Lewis's capitalist sector will indeed look like a standard neoclassical firm, at least for some model specifications and parameterizations.[1]

Many authors have sought to formalize the Lewis model and to identify a set of assumptions or rigidities that will deliver a version of the Lewis dualism. This paper does not seek to go over the same ground; interested readers can consult Wang and Piesse (2013) for a thoughtful treatment of alternative microfoundations for the Lewis model. Instead, this paper focuses on some key aspects of Lewis's vision of the development process: the dual economy, subsistence wages, patterns of unemployment and underemployment, labor market imperfections, savings, the mechanisms of development, and turning points in the growth process. In this sense, the paper is closest to Kirkpatrick and Barrientos (2004).

I will argue that many of the specific assumptions and mechanisms of the Lewis model have not been well supported by contemporary theory and evidence. This calls into question efforts to use the Lewis model in a very literal fashion for policy analysis. In spite of that, I will argue that the model remains a powerful and useful tool for thinking about growth because it correctly identifies a key feature of the growth process—namely, the importance of *within*-country gaps in income and productivity, or dualism. Lewis made the incisive observation that poor countries are not uniformly poor and that even the poorest countries have firms, sectors, and locations that operate at high levels of productivity. Lewis was perhaps less convincing in explaining why these islands remain and why the within-country gaps are not eliminated through migration and factor mobility. But his framing of the question remains deeply compelling.

This paper will not provide new interpretations of the Lewis model or seek to summarize it in authoritative fashion. Dozens of papers, chapters, and textbooks already offer reviews and interpretations of the Lewis model, and it would be

[1] Consider, for instance, a simple two-sector model economy in which there are two technologies for producing (identical) output: a subsistence technology, $Y = A(1 - n)$ and a capitalist technology, $Y = Mn^{\alpha} k^{(1-\alpha)}$, where n is the labor used in the capitalist sector such that $0 \leq n \leq 1$. For some parameter values, this model will yield an interior equilibrium in which the wage is fully determined by the linear productivity level in the subsistence technology, and the capitalist sector will take the wage as given. However, for other parameterizations, with sufficiently high levels of capital or high levels of productivity in the capitalist sector, the entire labor force can end up in the capitalist sector.

impossible to review all the literature that owes a debt to Lewis. Those looking for a textbook exposition of the model might begin with Ray (1998, Section 10.2) or for a less faithful but equally interesting version, the two-sector model of Eswaran and Kotwal (1993). Basu (1997, chap. 7) offers a critique of the internal consistency of the Lewis model. Several thoughtful papers on the Lewis model and its intellectual implications were written for a two-day conference held at the University of Manchester on the occasion of the 50th anniversary of the paper; these retrospective evaluations can be found in a special issue of *The Manchester School* (vol. 72, no. 6, 2004).

Dualism and Non-neoclassical Foundations

Perhaps the central idea of the Lewis model is the notion that a modern ("capitalist," in Lewis's classical terminology) sector and a traditional ("subsistence") sector coexist in developing countries. The traditional sector is not precisely defined, but it consists of people earning a subsistence wage—perhaps subject to some Malthusian equilibrium, as formalized later in Galor and Weil (2000) or Hansen and Prescott (2002). Lewis determinedly declines to identify the modern sector with industry or the traditional sector with agriculture, noting that commercial agriculture fits his definitions of "modern." He also recognized that the dualism did not correspond entirely to a rural–urban divide, noting that within rural areas (and within the agricultural sector) there are enterprises that seem fully capitalist, and within urban areas of the developing world, there are large numbers of workers in the nontradable service sector earning little more than subsistence wages. As Lewis (1954, p. 141) described them, these urban subsistence workers in this way included

> . . . the workers on the docks, the young men who rush forward asking to carry your bag as you appear, the jobbing gardener, and the like. These occupations usually have a multiple of the number they need, each of them earning very small sums from occasional employment; frequently their number could be halved without reducing output in this sector. Petty retail trading is also exactly of this type; it is enormously expanded in overpopulated economies; each trader makes only a few sales; markets are crowded with stalls, and if the number of stalls were greatly reduced the consumers would be no whit worse off—they might even be better off, since retail margins might fall.

For Lewis, the key feature of this traditional subsistence sector was that it existed alongside the capitalist sector and was effectively unlimited in size, thus potentially providing a perfectly elastic supply of labor to the capitalist sector at a fixed wage. The sheer size of the subsistence sector meant that the modern sector could grow without facing any labor constraints. In particular, wages in the capitalist sector would be determined by the wage in the subsistence sector, which in turn would correspond to something approximating a subsistence level of consumption.

Lewis postulated that this condition would hold until, at some unspecified future point, the growing capitalist sector would finally pull so many workers out of the subsistence sector that the supply of labor would no longer remain perfectly elastic.

How well does Lewis's dualism stand the test of time? The basic idea of dualism remains ubiquitous in the development and growth literature today. A few researchers use the term explicitly (Temple 2005; Temple and Wößmann 2006; Vollrath 2009a, 2009b). Many more use two-sector models in which the sectoral dichotomies are characterized by terminology that is less redolent of the classical literature: formal–informal; modern–traditional; industrial–agricultural. All of these dualistic models in some sense carry on Lewis's thinking, and the dualisms do seem real in the data—even if the boundaries of the dualistic sectors remain ill-defined and occasionally unsatisfying.

Moreover, Lewis's vision of dualism still seems broadly accurate today, although it is difficult to define precisely or to pin down in data. Large numbers of people in poor countries work in quasi-subsistence agriculture and in very low-productivity informal services. This does seem to comprise a "subsistence" sector that is distinct from the high-productivity formal sector. It is difficult to measure the sizes of the two sectors by objective criteria, but we know that there are important differences between agriculture and nonagriculture in developing countries, coinciding imperfectly with a rural–urban split. There are also disparities in urban areas between informal and formal sectors, in terms of average productivity and wages.

Restuccia, Yang, and Zhu (2008) and Caselli (2005) pointed out that the nonagricultural sector in low-income countries appears to be relatively close in average productivity to the nonagricultural sector in high-income countries, with countries at the 90th percentile in the cross-country income distribution having nonagricultural labor productivity about four times as high as those countries at the 10th percentile. In contrast, the 90th to 10th percentile differences in agricultural labor productivity are much larger—about a factor of 45. In a similar vein, in Gollin, Lagakos, and Waugh (2014a), my coauthors and I document differences in average labor productivity between agriculture and nonagriculture and show that these are particularly pronounced in poor countries. In many of the lowest-income countries, the average productivity of labor in agriculture is less than half that of the level in the nonagricultural sectors of the economy. This gap remains even after extensive corrections for differences in hours worked and human capital; the productivity gaps appear in micro data as well as in aggregate data. The average product of labor is not the same as the marginal product, so this is not necessarily evidence of a gap in wages across sectors, nor is it evidence of misallocation per se. The agricultural productivity gap is, however, evidence of a kind of dualism. This dualism extends from the production side of the economy to realized living standards: Young (2013) documents large disparities between urban and rural areas in a number of different measures of well-being.

Lewis's dualism is difficult to pin down, however. Much (but not all) of the agricultural sector in the poorest countries, along with some fraction of the rural and urban services sector, would seem to fall into Lewis's "subsistence" sector. His

"capitalist" sector corresponds more or less to the formal nonagricultural sector, perhaps leaving aside the government sector. But even within narrowly defined sectors, the lines can be blurred. For instance, many formal sector firms in developing countries—clearly capitalist in Lewis's definition—rely on a fringe of subsistence workers for distribution or sales. These individuals may or may not be formally employed by the firm. For instance, the large mobile phone providers in many African countries distribute air time through networks that extend ultimately to young people selling scratch-off vouchers along the roadsides. In the same way, large breweries and soft drink manufacturers typically rely on distribution chains that include informal street vendors and the owners of very small shops.

Perhaps in the end, Lewis's dualism is too stark. The dichotomy between capitalist and subsistence sectors appears on closer examination to be more of a continuum. Taking the retail food sector as an example, there are large formal retail establishments such as supermarkets in many developing countries; there are also people selling oranges and pineapples by the roadside from atop their heads. In between, there is nearly a full range of shops of different sizes, from roadside stands to market stalls to small shops. For example, Woldu, Abebe, Lamoot, and Minten (2013) offer a detailed description of food retailing in Addis Ababa and a taxonomy of sellers. Weatherspoon and Reardon (2003) discuss the evolution of food retailing in Africa and the emergence of supermarkets and chains.

The same argument about a continuum could be made for the agricultural sector in many low-income countries, which typically includes a few producers who are entirely in subsistence but many more who sell small amounts of surplus and others who are nearly fully commercial. Thus, dualism disappears under the microscope. Yet in some larger sense, Lewis's dualism was a useful abstraction—and it remains so. The basic insight seems correct and important—that there are large differences in productivity *within* countries as well as across countries. These within-country disparities are partly linked to sectors and partly to geographic space; perhaps they also reflect underlying inequality in access to capital and other resources. Development must surely involve both a movement of people (and resources) across the dualistic divide and a reduction of the barriers and obstacles that lead to dualism. This central insight of Lewis seems entirely valid today.

Subsistence Wages

A key ingredient of Lewis's model was the notion that in the subsistence sector, wages were determined not by neoclassical logic but by something approaching a biophysical notion of subsistence. Although Lewis did not formally invoke Malthus in his paper, he repeatedly emphasizes that earnings in this sector are determined by the subsistence level. In fact, he uses the word "subsistence" 92 times in the paper. He writes (p. 142) that "[t]he price of labour, in these economies, is a wage at the subsistence level." The subsistence level is slippery to define. Lewis struggles with it at length before waving his hands and dodging the issue. Influenced again by

classical thinking, he begins with the notion that "[t]he classical economists used to think of the wage as being determined by what is required for subsistence consumption, and this may be the right solution in some cases." But he recognizes that in agrarian economies, smallholders may receive land rents, so that they may ultimately earn significantly more than required for bare subsistence. After wrestling with concepts such as "the average product of the farmer," Lewis suggests in the end that the wage in the subsistence sector may be determined by a "conventional standard of living." In the end, with further a waving of hands, he writes, "It is not, however, of great importance to the argument whether earnings in the subsistence sector are determined objectively by the level of peasant productivity, or subjectively in terms of a conventional standard of living. Whatever the mechanism, the result is an unlimited supply of labour for which this is the minimum level of earnings."

In hindsight, it is not clear that the subsistence wage was a necessary ingredient of Lewis's model. Indeed, Ranis and Fei (1961), in their early formalization and extension of the Lewis model, argued simply for a non-neoclassical wage—meaning a wage higher than the marginal value product of labor—in the subsistence sector. This wage was determined by some social norm or "institutional or nonmarket forces" (p. 536). Ranis and Fei equated Lewis's "capitalist" sector with the nonagricultural sector, and their version of the "subsistence" sector was the agricultural sector. In their version of Lewis's model, the key was that the marginal product of labor must be very low in agriculture, if not literally zero, so that labor could move across sectors without reducing the availability of food (and hence reducing the real wage) in the nonagricultural sector. But if the marginal product of labor was low, then if workers received a neoclassical wage, the marginal worker would receive a near-zero wage, making dualism unsustainable. Ranis and Fei saw a way out by invoking an institutionally determined wage, greater than the subsistence wage, which would be received by agricultural workers. Specifically, they proposed a formula such that each worker in the agricultural sector would receive the average product of labor, so that the agricultural wage could be comfortably above subsistence even when the marginal product was effectively zero. Another advantage of their formulation was that it allowed for dynamics within the agricultural sector—such as population growth or agricultural productivity increases—to matter for the development process. The Fei–Ranis approach has been expanded and updated in more complete and more recent treatments, and other interpretations have been offered; for example, by Wang and Piesse (2013), who propose a more completely developed set of microfoundations for a Lewis-inspired model.

Thus, Lewis's insistence on subsistence wages was largely discarded more than a half-century ago. Over the past 60 years, evidence has grown that in most developing countries, wages and living standards are not constant at an absolute level of subsistence; on the contrary, even in those countries that have remained relatively poor, absolute living standards have on average increased substantially. This is not evidence against the more modest Lewis notion of wages being determined by a "conventional standard of living," or the Ranis and Fei (1961) notion of an "institutional force," but it does seem to diverge from a simplistic version of Lewis's

model—that is, the notion that industrialization could proceed in many countries for extended periods without increases in wages. A more nuanced reading of Lewis might allow for wages to rise in the presence of differential productivity growth across sectors, or for a labor surplus to remain even in the presence of an increasing marginal product of labor.

A more striking finding from the micro evidence is that Lewis may have abstracted too readily from heterogeneity within the "subsistence" sector. In Lewis's framing of the issue, essentially everyone in that sector earned the same effective wage, which in turn set the wage for the modern sector up until the "turning point." In an era when household survey data were rare, this generalization may have been reasonable. However, as we have greatly increased our understanding of the heterogeneity within rural areas, agricultural populations, and the urban informal sector, the data show substantial dispersion even within rural populations. For instance, within the rural population of China, the Gini coefficient for rural expenditure (often used as a proxy for income) was 41.5 in 2009, comparable with the national figures for Qatar or Nicaragua; in Indonesia, the rural Gini was 34.0, the same level as reported for national income statistics in the United Kingdom or Italy.[2] These measures of inequality show that rural households vary substantially in their living standards; they are not all living at some absolute level of subsistence, and they are not enduring "shared poverty" through some kind of pooling of income. Lewis's vision of a subsistence sector in which wages are pegged to some kind of Malthusian level seems on closer examination to be inaccurate. Similarly, the notion that everyone receives a wage that approximates the average product looks to be little more than a romanticized view of a world that in reality displays moderate levels of inequality and heterogeneity. Although Lewis's notion of a "subsistence sector" has some an appeal in a stylized sense, it is not clear whether it corresponds to any operationally meaningful category.

Unemployment and Underemployment

Lewis's notion of "unlimited supplies of labor" implicitly required a kind of "disguised unemployment" or "underemployment." Lewis himself was unafraid of using the term "unemployment" to characterize work that involved low-productivity activities. At times, he seemed to associate this concept with the importance of fixed factors in production, such as land. In this sense, Lewis equated "disguised unemployment" with "surplus population," a term he invokes in the original essay. The connection is made explicit in places; thus, he writes (p. 189) that for many sectors of

[2] The estimated rural Gini indexes for China and Indonesia are taken from the World Bank's PovcalNet data tool, along with the national figures for Nicaragua and Qatar. The PovcalNet data were downloaded from: http://iresearch.worldbank.org/PovcalNet/index.htm (last accessed on June 4, 2014). The national figures for the United Kingdom and Italy are from OECD (2013), available online at http://dx.doi.org/10.1787/factbook-2013-en.

the economy: "if the country is overpopulated relatively to its natural resources, the marginal productivity of labour is negligible, zero, or even negative." For Lewis, both natural resources and capital were effectively fixed factors in the short run, implying that some fraction of the labor force was necessarily unemployed. Lewis seems almost to have envisaged the subsistence sector as facing a fixed-coefficient technology, so that the available resource base could only absorb a certain amount of labor; the rest of the labor force was surplus to requirements and could be pulled out of the subsistence sector without giving up any production. Lewis was not arguing that the entire subsistence sector was unemployed or underemployed; rather, he argued that within that sector there was some fraction of labor that could be withdrawn without a consequential loss of output in that sector.

Lewis's view has been more widely accepted in the policy world than in the academic literature. The empirical literature has struggled to define "unemployment" in the poorest countries, where survey data consistently show that almost all able-bodied individuals work in some fashion, often in self-employment or family business. Micro development economists invariably find positive (though low) returns to labor in almost all surveys of individuals, households, and firms. The micro literature tends to find that individuals and households scrape together livings from broadly diversified portfolios of activities, any or all of which may have very low productivity. (For an example from this journal, Banerjee and Duflo, 2007, offer a vivid depiction of this reality.) Even children and the elderly normally generate positive and non-negligible marginal products, in both market-oriented activities and home production. The micro literature on labor markets in developing countries has tended to view with skepticism the notion of widespread unemployment—disguised or otherwise—other than that caused by seasonality, disability, and other unavoidable barriers.

The macro and policy literatures, however, remain open to the possibility that many individuals are in some sense effectively unemployed. To many governments in the developing world, some of the urban informal workforce appears to be effectively unemployed. For instance, recent reports on youth employment in Africa (for example, World Bank 2009) point out that formal unemployment—as defined by labor statistics—is rare, even for populations that struggle to find good jobs. The development policy literature, taking a fairly macro view, sees many people in poor countries employed in jobs that combine informality, part-time or irregular hours, and little or no return to skill or experience. The literature quibbles over semantics—whether this population should be characterized as unemployed, underemployed, or informally employed, and these terms often embed differing narratives and policy implications. Writing in 1954, Lewis did not feel a need to distinguish among the subtleties; from his point of view, people in all these categories formed the reservoir of surplus labor.

One way to reconcile the micro and macro views of unemployment is to consider the possibility that the marginal *social* value of labor may be very low—very much in the sense that Lewis described. When one additional individual joins the queue of roadside sellers of popcorn or flyswatters in Kampala or Chittagong,

the marginal product of that individual's labor may be positive for that individual and his or her family; but whether there is a positive social value is unclear. Arguably, this individual is simply taking business from other sellers, creating little or no additional social value. Policymakers often take this view when they look at the overall abundance of labor and the low social value of what is being done. Lewis also articulated this view in a spirited (and occasionally testy) defense of the surplus labor proposition (Lewis 1968, pp. 12, 14), in which he argued that it was possible simultaneously for the marginal product of labor to be positive on the intensive margin, for any given worker, but also to be approximately zero on the extensive margin, for an additional person.[3]

> I do not believe that the productivity of a manhour is zero in agriculture, domestic service, petty retailing, handicrafts, or any other part of the non-capitalist reservoir. Nevertheless, I have seen nothing in the now vast literature of under-employment to alter my belief that in India or Egypt one could mobilise a group equal to (say) ten per cent of the unskilled non-capitalist labour force without reducing significantly the output of the non-capitalist sectors from which they were withdrawn. . . . However, this is all an irrelevant digression, since the model in no way depends on the marginal product in agriculture, whether per person or per manhour. . . .

There is relatively little micro evidence on Lewis's claim about marginal labor productivity on the extensive margin. One exception is Foster and Rosenzweig (2010), who calculate that approximately 20 percent of the Indian agricultural labor force could be effectively surplus, based on calculations of the minimum efficient scale of farms. Foster and Rosenzweig calculate that if all farms were operated at an optimal scale, there would be a higher average ratio of land per worker. They calculate that some fraction of workers could thus be released from agriculture without reducing overall output; that is, the losses of output due to the release of labor would be offset by the increased efficiency from expanding farm size. In their analysis, the source of this surplus labor is that farm size in India is inefficiently small, reflecting some unspecified barriers to consolidation—perhaps legal and institutional, perhaps related to failures in other markets. Their findings rest on essentially the same logic that Lewis invokes: even though the marginal product of labor is positive on all farms, labor could be released in the aggregate without reducing output. The Foster and Rosenzweig (2010) findings suggest a misallocation of labor between sectors, but of course this is not the same as finding a surplus of labor for the economy as a whole.

To summarize, the Lewis model seems to have been wrong in assuming that wages in the capitalist sector are determined by a subsistence wage, and perhaps

[3] In the same paper, Lewis (1968) also reiterated the point that his model does not require a zero marginal product of labor; it simply needs the supply of labor to the capitalist sector to be more or less perfectly elastic.

also wrong in assuming that growth could proceed in many countries for extended periods without increases in wages or standards of living. His broader point may have been valid: that firms in the modern sector in developing countries face a very large pool of workers who are willing to work for a wage that would give them a modest increase in living standards relative to the subsistence sector. Whether the supply of labor is *literally* perfectly elastic may be somewhat beside the point.

Labor Market Imperfections

The Lewis model is sometimes portrayed as a model with barriers to movement between sectors. But in fact, Lewis posits a perhaps surprisingly free movement of workers across sectors. Labor is more or less indifferent between working in the capitalist or the subsistence sector. Wages are slightly higher in the capitalist sector, Lewis suggests (pp. 150–51), because of differences in costs of living and some nonmonetary compensation for the "psychological cost of transferring from the easy going way of life of the subsistence sector to the more regimented and urbanised environment of the capitalist sector." But this wedge is broadly consistent with a labor market equilibrium in which workers have no desire, in equilibrium, to move across sectors. Marginal productivity is also effectively equalized across sectors, in a peculiar sense: it is zero (or near-zero) in both sectors. In the subsistence sector, marginal product is near zero because of fixed factors and "overpopulation." But then marginal product is also essentially zero in the capitalist sector.

What does differ sharply across sectors is *average* productivity. In the capitalist sector, this is quite high because of the presence of capital. In the subsistence sector, average product is presumed to be lower. The real labor market imperfection lies in the determination of wages in the subsistence sector. Here, Lewis fails to spell out the reason that the wage lies above the marginal product of zero. This implies some non-neoclassical characteristics of the labor market. Indeed, the labor market imperfections of the Lewis model are not related to barriers to mobility—a subject that has received extensive recent treatment in widely differing contexts.[4] Neither are they necessarily related to labor market imperfections or disequilibrium in the capitalist sector, although that view has attracted considerable attention going back to Harris and Todaro (1970) and more recently has been revisited as a point of contention in Brown (2006) and Fields (2006).

[4] A number of recent literatures have looked at different types of barriers to labor mobility across sectors. Bryan, Chowdhury, and Mobarak (forthcoming) consider information barriers to migration; Dercon, Krishnan, and Krutikova (2013) note the importance of subjective well-being; Caselli and Coleman (2001) focus on the costs of acquiring skills needed for migration; In Gollin, Parente, and Rogerson (2004, 2007), we view subsistence factors as a barrier to migration; In Gollin and Rogerson (2014), we consider transportation costs as an additional source of differences in sectoral productivity; and Vollrath (2009b) sees differential fertility patterns as a potential source of dualism.

Lewis's view of a fixed capital stock for the capitalist sector seems difficult to accept, particularly in today's era of rapid and relatively open investment flows. Perhaps in the 1950s, it made sense to think of developing economies as operating in some kind of financial autarky, but this aspect of the Lewis model seems problematic today. Why does capital not move in to low-income countries to employ "surplus" labor? This is, of course, the central question posed by Lucas (1990), and it remains a fundamental puzzle for the development and growth literature today.

Savings and the Mechanisms of Development

A key theme in the Lewis model—perhaps the most important feature of the model, from Lewis's perspective—is the importance of capital investment as a source of growth. For Lewis, capital represents a fixed factor in the short run for most developing countries. There simply isn't enough capital to absorb all of the economy's labor in the modern sector. Lewis's thinking was heavily informed by the Harrod–Domar model and the other planning-oriented growth theories of his day. (In fact, two of Lewis's major works were books on planning: he published *The Principles of Economic Planning* in 1949 and revisited the topic specifically in the context of developing countries in *Development Planning: The Essentials of Economic Policy*, published in 1966. A major focus of these works is how to mobilize sufficient capital for an economy to grow and how to allocate it across sectors to achieve certain planning goals, assuming different values of the incremental capital output ratio for different sectors and different economies.) The "surplus labour" paper that spells out the Lewis model is also the source of the famous quotation that "[t]he central problem in the theory of economic development is to understand the process by which a community which was previously saving and investing 4 or 5 per cent of its national income or less, converts itself into an economy where voluntary saving is running at about 12 to 15 per cent of national income or more. This is the central problem because the central fact of economic development is rapid capital accumulation (including knowledge and skills with capital)."

An important corollary of Lewis's view of capital as the key source of growth in labor surplus economies was that foreign assistance and other forms of foreign capital inflows could play a central role in driving development. As Easterly (1999) pointed out, the Lewis model and an associated view of capital fundamentalism, based on a Harrod–Domar view of the world, remained for many decades important ingredients in the measurement of "financing gaps" and the business of foreign aid.

Lewis's capital fundamentalism is not, I think, an essential ingredient of his theories of structural transformation, but it is related to a key set of puzzles about the model. Why doesn't the modern sector grow? Why doesn't it attract high rates of investment, given the large pool of unemployed or underemployed labor that could be productively used? Why isn't the capital spread more evenly across the labor force? Lewis repeatedly says in the paper that this does not happen, and he seems to have in mind some kind of lumpiness or indivisibility with respect

to investments. But this argument is never made explicit, and the implied model seems to require some kind of nonconvexity and/or market imperfection in the capital market.

Lewis builds his case on a set of propositions, none of which seems to be particularly well supported by the evidence available today. First, he supposes that all investment comes from the capitalist sector—and primarily from the savings of capitalists. This means that the capitalist sector can grow only from its own rents. In countries where the capitalist sector is small, this means that growth in absolute terms cannot be very rapid. As the capitalist sector grows, there is a reinforcing phenomenon, with the steady expansion of the capitalist sector leading to progressively higher savings rates. In this way, Lewis's story offers an explanation for the observed positive correlation between income per capita and investment rates. Lewis also felt that his framework offered an explanation or prediction concerning factor shares. In poor countries, the logic of his argument suggested that the capital share of income would be low. But as the capitalist sector expands through its own investment, there will initially be little or no increase in wages, leading to an increasing share of capital in national income. At the point where wages begin to rise, this trend might eventually be reversed.

Lewis's observations about the correlation between savings rates and income levels reflected a common view at that time concerning the importance of savings rates. This view has remained deeply embedded in the development policy arena in spite of well-founded concerns that the relationship is not causal (Easterly and Levine 2001). In the growth literature, the view of investment rates as an exogenous determinant of income levels was a widespread view in textbook treatments of the Solow model and in empirical papers such as Mankiw, Romer, and Weil (1992). More recently, however, views have arguably shifted; not only is the savings rate seen as potentially endogenous, but also the correlation between investment rates and income levels has been portrayed as misleading, relating more to measurement approaches than to underlying causal links (for example, Laitner 2000; Restuccia and Urrutia 2001; Hsieh and Klenow 2007). Seen through this lens, Lewis's effort to explain the investment–GDP relationship seems, with the benefit of 60 years of hindsight, to have been misguided. Reflecting back on his original work in 1968, Lewis appeared to recognize that his own capital fundamentalism was not the right story for many countries—and that the Green Revolution appeared to be associated with agriculture-led growth in some Asian countries. "This author is delighted that there are economies where the productivity of peasants increases steadily and that some portion of that increase goes into capital formation," Lewis (1968) wrote, somewhat defensively. "This does not render it useless or dangerous to study models of economies where in the initial stages the dynamism of growth is located in capitalist expansion."

In general, Lewis's emphasis on capital as a source of growth seems in retrospect to have overlooked the importance of productivity growth, and his assumption that only capitalists can invest productively seems inconsistent with current micro and macro evidence on savings behavior and investment.

Turning Points in the Growth Process

Lewis understood that in his framework, economies could not and would not indefinitely have unlimited supplies of labor—or in other words, they would not have perfectly elastic supply of labor to the capitalist sector at a wage determined by equilibrium in a subsistence sector. At some point, assuming that population growth did not outpace the accumulation of capital, sufficient labor would be pulled out of the subsistence sector to drive that part of the economy into a neoclassical mode of operation in which wages would be driven upwards by a rising marginal product of labor. For Lewis and his subsequent expositors, this moment represents a turning point. Until that moment, the capitalist sector can expand with fixed wages; beyond that moment, expansion of the capitalist sector will come in a context of rising wages.

This turning point has attracted enormous attention over the years. Much of this attention has focused on the question of whether, in the years before the turning point is reached, growth actually occurs with no increase in wages. This would arguably be an undesirable kind of growth, taken at face value; but potentially one could imagine this as a stage of growth that would allow an economy to expand and diversify. Presumably the model requires some kind of export outlet for goods produced in the growing capitalist sector, since wages are not rising and domestic consumption is then flat. Recent commentators have remarked on apparent similarities between this pre-turning-point growth and the experience of the East Asian economies, including most recently China. The argument is offered that China is reaching its Lewis turning point—presumably leading to a period of rising wages, declining comparative advantage in manufactures, and declining returns to investment (for example, Zhang, Yang, and Wang 2010; Wang and Weaver 2013; Das and N'Diaye 2013). To a large extent, this view of China reprises an earlier set of discussions of the East Asian growth experience. In that context, too, the question was whether their growth was sustainable and replicable elsewhere. Krugman (1994) argued that the East Asian miracle was in all probability unsustainable, founded on input intensification rather than productivity improvements, and cited work by Young (1995) measuring relatively low productivity growth in four East Asian economies. Although Krugman did not mention Lewis explicitly, his assessment of the East Asian experience had many of the same elements: these economies were able to grow by pulling in large supplies of low-cost labor—primarily drawing workers from rural areas and from the urban informal sector.

This paper does not seek to address the disagreement about the underlying causes of the East Asian economic growth miracle nor about China's economic prospects. One point to note, however, is that the Lewis model is neither a necessary nor a sufficient condition for growth through factor accumulation. A standard Solow-style model with growth in the labor force would allow for economic growth through accumulation of factors. The specific prediction of Lewis-style growth before the turning point is that wages in the capitalist sector will remain approximately flat

during a period—potentially quite protracted—of growth. Indeed, a specific prediction of the Lewis model—and one emphasized by Lewis both in the original 1954 paper (for example, p. 190) and in his 1968 follow-up (p. 20)—is that with flat wages and a growing capitalist sector, the economy-wide share of capital should be rising until the turning point. In modern parlance, this is a claim about factor shares. Lewis argues that poor countries with surplus labor should see the capital share rise and the labor share of income fall. But Young (2003, p. 1255) suggests that for China in the period from 1978–1997, factor shares are approximately flat; he also notes (p. 1255) that his earlier work found the same for four other East Asian countries in the period from 1960–1990. Thus, if we take the Lewis model literally, it is not clear that these economies looked like "surplus labor" economies. Perhaps they were growing through factor accumulation; but this looks more like Solow-style economic convergence in the presence of workforce growth than like Lewis growth.

An Assessment

How then to assess the Lewis model? From a long view, Lewis's contribution is not only seminal, but also profoundly useful. The iconic model has become deeply embedded in contemporary thinking about development and growth because its basic structure seems to capture a key reality of the development process. Lewis's stylized description of a dualistic economy rings true with anyone who has spent time in a poor developing country, where modern glass buildings and shiny downtown areas coexist with huge populations of farmers scratching the soil with hand tools. By focusing on this fundamental dualism, Lewis offered a useful way to think about the development process. His model offers a crude but persuasive depiction of the growth process, in which growth occurs through the reallocation of labor and other resources across sectors. The model puts structural transformation processes at the heart of economic growth—a view that has captured renewed attention over the past few years. There is abundant empirical support for the proposition that structural transformation does, in an accounting sense, explain a large fraction of growth and income levels.

What is less clear is that the underlying mechanisms of the Lewis model were correct. The labor market dynamics that he posits do not seem to apply in the real world, and the capital fundamentalism that drives growth in his model seems overly restrictive. The specific assumptions and implications of the model seem almost uniformly to conflict with micro evidence and macro data. Even for countries like those in East Asia that seem at first glance to have been characterized by surplus labor, the evidence conflicts with a key prediction of the Lewis model—namely that the capital share of income should be rising steadily until the Lewis turning point (after which they should fall steadily).

To what extent does it matter that the model is wrong in its details if it is nevertheless compelling in its outlines? The shortcomings of the Lewis model are important, both for our understanding of the growth process and for policies that

are designed to promote development. In the Lewis model, infusions of capital can play a crucial role in unlocking growth; this does not seem to be true in reality. The Lewis model might lead policymakers to imagine that labor can be moved costlessly out of the agricultural sector or the informal services sector; on the contrary, the evidence suggests that the people occupying these sectors are productively engaged and have positive marginal product. There may be settings or sectors where labor could be freed from low-productivity uses at low social costs, as suggested by the Foster and Rosenzweig (2010) evidence on farms that are operating at an inefficiently small size. But implementing such changes depends on identifying and fixing imperfections in the markets for other factors, such as capital and land.

A recent literature does provide evidence that misallocations across sectors—and even across firms within sectors—may play an important role in explaining aggregate income differences and productivity differences across countries. Both micro and macro literatures in development have explored issues of misallocation in recent decades (for example, Banerjee and Duflo 2005; Restuccia and Rogerson 2008; Hsieh and Klenow 2009; Bartelsman, Haltiwanger, and Scarpetta 2013).

Even where there is no clear evidence of misallocation per se, there is overwhelming evidence of spatial and sectoral disparities within countries. Although Lewis was careful not to associate his subsistence sector with agriculture, the data today seem clearly to point to the agricultural sector as a major source of within-country disparities in income and productivity. Large fractions of the labor force in poor countries work in agriculture, and systematically the average productivity of agricultural labor appears to be low, as do living standards in rural areas (Gollin, Lagakos, and Waugh 2014a, b; Young 2013). Understanding the growth process will require a richer understanding of the forces keeping hundreds of millions of the world's poorest people in rural areas and tying them to low-productivity work in agriculture. Lewis's explanation of these forces was perhaps inadequate, but he was surely looking at the right questions.

The Lewis model does invite a set of research questions that remain important today—and that were perhaps neglected for too long in the development and growth literatures. Both academic economics and the world of development policy were arguably hurt by the relative neglect of dual economy models for several decades, beginning in the mid-1960s. The long dominance of one-sector models in the growth literature meant that questions of importance to developing countries were not really addressed. The issues that stand out today are related to the sources of dualism. We need to learn more about the kinds of frictions and barriers that prevent the movement of labor across sectors. These barriers may reside in labor markets, or they may be related to frictions in markets for land and capital. We also need to understand better the reasons why productivity differs so markedly across sectors. But with new data sources and more ability than ever before to collect and analyze data, it seems reasonable to aim for an updated and improved understanding of dualism—one that is consistent with the data and can guide policy choices in the years ahead.

■ *I gratefully acknowledge comments from the editors: David Autor, Chang-Tai Hsieh, Ulrike Malmendier, and especially Timothy Taylor. I also benefited from detailed comments and suggestions from several colleagues, including particularly Stefan Dercon, Cheryl Doss, Chris Udry, and Xiaobing Wang. I also honor the memory of the two professors who introduced me to the Lewis model: Bob Evenson and Gus Ranis, both of whom passed away in 2013. They are greatly missed.*

References

Banerjee, Abhijit V., and Esther Duflo. 2005. "Growth Theory through the Lens of Development Economics." In *Handbook of Economic Growth*, Vol. 1A, edited by Philippe Aghion and Steven N. Durlauf, 473–552. Amsterdam: Elsevier B.V.

Banerjee, Abhijit V., and Esther Duflo. 2007. "The Economic Lives of the Poor." *Journal of Economic Perspectives* 21(1): 141–68.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta. 2013. "Cross-country Differences in Productivity: The Role of Allocation and Selection." *American Economic Review* 103(1): 305–34.

Basu, Kaushik. 1997. *Analytical Development Economics: The Less Developed Economy Reconsidered*. Cambridge, MA.: MIT Press.

Brown, Richard P. C. 2006. "On Labour Market Dualism in the Lewis Model: A Comment." *The Manchester School* 74(3): 350–54.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak. Forthcoming. "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh." *Econometrica.*

Caselli, Francesco. 2005. "Accounting for Cross-Country Income Differences." In *Handbook of Economic Growth*, edited by Phillipe Aghion and Steven Durlauf, 679–741. New York: North Holland.

Caselli, Francesco, and Wilbur John Coleman II. 2001. "The U.S. Structural Transformation and Regional Convergence: A Reinterpretation." *Journal of Political Economy* 109(3): 584–616.

Das, Mitali, and Papa N'Diaye. 2013. "Chronicle of a Decline Foretold: Has China Reached the Lewis Turning Point?" IMF Working Paper WP/13/26.

Dercon, Stefan, Pramila Krishnan, and Sonya Krutikova. 2013. "Migration, Risk-sharing, and Subjective Well-being." Unpublished manuscript, Oxford University, Department of International Development.

Gollin, Douglas, Stephen L. Parente, and Richard Rogerson. 2004. "Farm Work, Home Work, and International Productivity Differences." *Review of Economic Dynamics* 7(4): 827–50.

Gollin, Douglas, Stephen L. Parente, and Richard Rogerson. 2007. "The Food Problem and the Evolution of International Income Levels." *Journal of Monetary Economics* 54(4): 1230–55.

Gollin, Douglas, and Richard Rogerson. 2014. "Productivity, Transport Costs, and Subsistence Agriculture." *Journal of Development Economics* 107(C): 38–48.

Easterly, William. 1999. "The Ghost of Financing Gap: Testing the Growth Model Used in the International Financial Institutions." *Journal of Development Economics* 60(2): 423–38.

Easterly, William, and Ross Levine. 2001. "What Have We Learned from a Decade of Empirical Research on Growth? It's Not Factor Accumulation: Stylized Facts and Growth Models." *World Bank Economic Review* 15(2): 177–219.

Eswaran, Mukesh, and Ashok Kotwal. 1993. "A Theory of Real Wage Growth in LDCs." *Journal of Development Economics* 42(2): 243–69.

Fields, Gary S. 2006. "Labour Market Dualism in the Lewis Model: Reply." *The Manchester School* 74(3): 355–59.

Foster, Andrew D., and Mark R. Rosenzweig. 2010. "Is There Surplus Labor in Rural India?" Yale Economics Department Working Paper 85.

Galor, Oded, and David N. Weil. 2000. "Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond." *American Economic Review* 90(4): 806–28.

Gollin, Douglas, David Lagakos, and Michael E. Waugh. 2014a. "The Agricultural Productivity Gap." *Quarterly Journal of Economics* 129(2): 939–93.

Gollin, Douglas, David Lagakos, and Michael E. Waugh. 2014b. "Agricultural Productivity

Differences Across Countries." *American Economic Review* 104(5): 165–70.

Hansen, Gary D., and Edward C. Prescott. 2002. "Malthus to Solow." *American Economic Review* 92(4): 1205–17.

Harris, John R., and Michael P. Todaro. 1970. "Migration, Unemployment and Development: A Two-Sector Analysis." *American Economic Review* 60(1): 126–42.

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

Hsieh, Chang-Tai, and Peter J. Klenow. 2007. "Relative Prices and Relative Prosperity." *American Economic Review* 97(3): 562–85.

Kirkpatrick, Colin, and Armando Barrientos. 2004. "The Lewis Model after 50 Years." *The Manchester School* 72(6): 679–90.

Krugman, Paul. 1994. "The Myth of Asia's Miracle." *Foreign Affairs* 73(6): 62–78.

Laitner, John. 2000. "Structural Change and Economic Growth." *Review of Economic Studies* 67(3): 545–61.

Lewis, W. Arthur. 1954. "Economic Development with Unlimited Supplies of Labor." *The Manchester School* 22(2): 139–91.

Lewis, W. Arthur. 1949. *The Principles of Economic Planning.* London: Allen and Unwin.

Lewis, W. Arthur. 1966. *Development Planning: The Essentials of Economic Policy.* London: Allen and Unwin.

Lewis, W. Arthur. 1968. "Reflections on Unlimited Labour." Development Research Project, Woodrow Wilson School Discussion Paper 5 (October).

Lucas, Robert E., Jr. 1990. "Why Doesn't Capital Flow from Rich to Poor Countries?" *American Economic Review* 80(2): 92–96.

Manchester School, The. 2004. An issue on the Lewis Model after 50 years, vol. 72, no. 6.

Mankiw, N. Gregory, David Romer, and David N. Weil. 1992. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics* 107(2): 407–37.

OECD. 2013. "Income Inequality." Section in *OECD Factbook 2013: Economic, Environmental and Social Statistics,* 66-67. Paris: OECD Publishing.

Ranis, Gustav, and John C. H. Fei. 1961. "A Theory of Economic Development." *American Economic Review* 51(4): 533–65.

Ray, Debraj. 1998. *Development Economics.* Princeton, NJ: Princeton University Press.

Restuccia, Diego, and Richard Rogerson. 2008. "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments." *Review of Economic Dynamics* 11(4): 707–20.

Restuccia, Diego, and Carlos Urrutia. 2001. "Relative Prices and Investment Rates." *Journal of Monetary Economics* 47(1): 93–121.

Restuccia, Diego, Dennis Tao Yang, and Xiaodong Zhu. 2008. "Agriculture and Aggregate Productivity: A Quantitative Cross-country Analysis." *Journal of Monetary Economics* 55(2): 234–50.

Temple, Jonathan. 2005. "Dual Economy Models: A Primer for Growth Economists." *The Manchester School* 73(4): 435–78.

Temple, Jonathan, and Ludger Wößmann. 2006. "Dualism and Cross-country Growth Regressions. *Journal of Economic Growth* 11(3): 187–228.

Vollrath, Dietrich. 2009a. "How Important are Dual Economy Effects for Aggregate Productivity?" *Journal of Development Economics* 88(2): 325–34.

Vollrath, Dietrich. 2009b. "The Dual Economy in Long-run Development." *Journal of Economic Growth* 14(4): 287–312.

Weatherspoon, Dave D., and Thomas Reardon. 2003. "The Rise of Supermarkets in Africa: Implications for Agrifood Systems and the Rural Poor." *Development Policy Review* 21(3): 333–55.

Woldu, Thomas, Girum Abebe, Indra Lamoot, and Bart Minten. 2013. "Urban Food Retail in Africa: The Case of Addis Ababa, Ethiopia." ESSP Working Paper 50, International Food Policy Research Institute/Ethiopia Strategy Support Program II, Addis Ababa, Ethiopia.

World Bank. 2009. *African Development Indicators 2008/09: Youth and Employment in Africa: The Potential, the Problem, the Promise.* Washington, DC: The World Bank.

Wang, Xiaobing, and Jenifer Piesse. 2013. "The Micro-foundations of Dual Economy Models." *The Manchester School* 81(1): 80–101.

Wang, Xiaobing, and Nick Weaver. 2013. "Surplus Labour and Lewis Turning Points in China." *Journal of Chinese Economic and Business Studies* 11(1): 1–12.

Young, Alwyn. 1995. "The Tyranny of Numbers: Confronting the Statistical Realities of the East Asian Growth Experience." *Quarterly Journal of Economics* 110(3): 641–80.

Young, Alwyn. 2003. "Gold into Base Metals: Productivity Growth in the People's Republic of China during the Reform Period." *Journal of Political Economy* 111(6): 1220–61.

Young, Alwyn. 2013. "Inequality, the Urban–Rural Gap, and Migration." *Quarterly Journal of Economics* 128 (4): 1727–85.

Zhang, Xiaobo, Jin Yang, and Shenglin Wang. 2010. "China has Reached the Lewis Turning Point." IFPRI Discussion Paper 977, International Food Policy Research Institute.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 89–108

The Missing "Missing Middle"[†]

Chang-Tai Hsieh and Benjamin A. Olken

The notion that the distribution of firm size in poor countries is characterized by a bimodal distribution with a "missing middle" is a widely accepted fact in development economics (for example, see Krueger 2013). The idea of the missing middle is that there are a large number of small firms, some large firms, but very few medium-sized firms.

The purported fact about the missing middle is cited as evidence for two broad stories of why many countries are poor. Perhaps surprisingly, these models look for the cause of the missing middle in two fundamentally different places. One approach suggests that *small firms* are disfavored in low-income countries—for example, by a lack of access to financial capital—and thus face difficulties in growing to become middle-sized firms. The other approach posits a "dual economy" of large high-productivity firms and small low-productivity firms, and then suggests that *larger firms* are disfavored in low-income countries—for example, by having to bear large fixed costs of regulation—which make it difficult for middle-sized firms to become established. We begin by reviewing these theories of development based on the purported fact of a "missing middle." We then explore the evidence on distribution of firm sizes in more detail and challenge the presumption that a "missing middle" occurs at all.

We present three main facts. First, there is in fact no evidence of a missing middle in detailed and comprehensive data on the size of manufacturing firms in India, Indonesia, or Mexico, regardless of how we slice the data. To be sure, there

■ *Chang-Tai Hsieh is the Phyllis and Irwin Winkelried Professor of Economics, University of Chicago, Chicago, Illinois. Benjamin A. Olken is Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Their email addresses are chsieh @chicagobooth.edu and bolken@mit.edu.*

[†] To access the Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.3.89

doi=10.1257/jep.28.3.89

are many more small firms in developing countries—but while medium-sized firms are missing in the data, large firms are missing as well. Put differently, while there are fewer middle-sized firms in developing countries than developed countries, there is no missing middle in the sense of a bimodal distribution.

Second, the average product of labor and capital is significantly lower in small firms when compared to larger firms. This is important because some theories say that small firms do not grow because they face high marginal costs of capital; if so, the marginal product of the capital that they do have should be higher. While we do not directly observe the marginal product of capital, it appears that the average product of labor and capital is significantly *lower* in small firms when compared to larger firms. To the extent that marginal and average costs move together, this fact suggests that large firms rather than small firms are the ones suffering the large fixed costs or shortage of capital that could stifle their growth.

Third, we consider the possibility that regulatory obstacles generate a missing middle, but find no evidence of meaningful discontinuities in the firm size distribution. We focus on regulations that kick in at a certain size threshold and test whether there are an unusually large number of firms right under the threshold and an unusually small number of firms right above the threshold: specifically, we focus on a size threshold of 100 employees in India where various labor regulations kick in; a revenue threshold in Indonesia above which firms are required to pay value-added tax; and a revenue threshold in Mexico above which firms face higher tax rates. However, we find no economically meaningful bunching of firms around these thresholds, which suggests that stories based on thresholds due to formality or regulations are unlikely to be causing major distortions in the economy. This evidence does not rule out the possibility that such forces are present, but it suggests that if fixed costs or thresholds are important, they must vary substantially across firms.

Given these facts, a natural question is how this misconception about the missing middle arose in the first place. We suggest that the misconception about the missing middle comes from the two transformations that have been made to data in generating the main evidence cited for the missing middle. Specifically, the main citation for the existence of the missing middle is a table on the distribution of employment shares in a number of countries in Tybout (2000). Due to data limitations, these tabulations are binned into three groups: firms with less than 10 employees, 10–49 employees, and 50 or more employees. The "missing middle" refers to the fact that there is less employment in the middle category (10–49 employees) than in the other two bins. In addition, these tabulations present the distribution of *employment share* by firm size and not the distribution of the number of *firms* by size. However, the relevant theories for which the missing middle is a key fact are about the firm itself (for example, theories that firms over a size are differentially taxed, or firms have trouble getting credit and so can't grow above a certain size) and not about the employment share, which is instead more relevant for understanding where the typical worker in the economy works.

We show that the widely cited facts about the missing middle come from the product of these two transformations of the data (using three broad bins and the employment share); neither one alone will produce the effect. When we bin the

firm size distribution into these three broad bins, it appears unimodal. Similarly, when plotted flexibly as a histogram, the employment share distribution appears unimodal. Only when one groups the employment share distribution into these three bins does the "missing middle" pattern emerge.

The absence of a missing middle suggests that the theories of development that cite the missing middle "fact" are not correct, at least without substantial modifications. Our evidence suggests that a major problem of economic development in low-income countries may not be how to relax the constraints faced by small firms, but instead how to relieve the differential constraints faced by large firms. In turn, this view of the world suggests that programs such as microcredit or tax regimes that seek to benefit small firms can worsen the development problem by further increasing the incentive for firms to remain small.

The Missing Middle in Theories of Development

The missing middle is an important presumed fact behind two models of development, one that emphasizes that small firms are disfavored relative to large ones and another that emphasizes that medium and large firms are disfavored relative to smaller ones.

The first model is the view that the institutional environment in poor countries discriminates against small firms and favors big firms. Such models come in several versions. The most common version, often put forward by supporters of micro-lending, is based on a claim that small firms are credit constrained—that is, they would like to borrow a larger quantity of funds at the prevailing interest rate but are unable to find a lender willing to lend—and large firms are not credit constrained. Closely related mechanisms are based on the idea that property rights are protected for formal firms but not for informal firms (De Soto 1989), or that large firms have better access to output markets. Other related models are based on the idea that government interventions benefit large firms, perhaps because the large firms are state-owned firms, or because industrial policy targets large firms, or because large firms are the main beneficiaries of protectionism and entry barriers.

A central prediction of many of these models is that the marginal return to resources should be higher in small firms compared to large firms. If small firms are constrained in their ability to obtain capital, they will also have high marginal products for the capital that they do have. Indeed, a number of papers estimate a very high return to capital in small firms in developing countries (de Mel, McKenzie, and Woodruff 2008; Udry and Anagol 2006; Kremer, Lee, Robinson, and Rostapshova 2013).

A second model of development that generates a missing middle harkens back to the "dual economy" view, as expressed in Arthur Lewis's (1954, p. 147) famous characterization of poor countries as islands of capitalist employment, "surrounded by a vast sea of subsistence workers . . . a few industries highly capitalized, such as mining or electric power, side by side with the most primitive techniques; a few high class shops, surrounded by masses of old style traders; a few highly capitalized

plantations, surrounded by a sea of peasants." In more recent manifestations of this view, McKinsey Global Institute (2001) argues that the most productive firms in low-income countries are as productive as the firms in high-income countries, but the vast majority of firms in low-income countries are low-productivity ones. Bloom and Van Reenan (2007) provide similar evidence, focusing on the distribution of the quality of management.

The theories behind this view often emphasize that in many low-income countries, medium and large firms face fixed costs or constraints that small firms do not face. While some large firms are able to spread such fixed costs over a large enough volume of sales, or perhaps to offset regulatory costs with counterbalancing government subsidies, medium-sized firms lack such abilities and face differential barriers. For example, Banerjee and Duflo (2005, 2011) argue that the marginal return from increasing scale is low for firms using "primitive" technologies but is high in firms using "modern" technologies. In their model, this situation arises because the fixed cost of modern technologies is prohibitively high in poor countries. Therefore only a small number of firms adopt such technologies.

The classic paper by Harris and Todaro (1970) was the first to model the dual economy view that large firms are subject to constraints and regulations that small firms are able to evade. Their model posits a "modern" sector that pays above-market wages and a "traditional" sector that pays market wages. Rauch (1991) formally shows how this mechanism can generate a "missing middle" by assuming a fixed threshold due to minimum wage laws or labor unions above which firms have to pay above-market wages. Krueger (2013), McKinsey & Co. (2005), and Levy (2008) are recent versions of the same idea, where large firms pay taxes and are subject to regulations (in India, Brazil, and Mexico, respectively) that smaller firms can evade.

The dual economy model generates specific empirical predictions, too. First, if some firms are capital-constrained while others are not, then when one plots the distribution of returns to capital, there should be a "barbell" shape with one group of firms showing high returns to capital and the other showing lower returns. Second, the question of whether one should expect the returns to inputs (capital or labor) to be higher in larger or in smaller firms depends on the type of production function assumed. But if it is the large firms that are constrained in a dual-economy-type model, then under reasonable assumptions (spelled out below), one would expect to find a higher rate of return to capital in those firms. Third, if the missing middle is due to the fixed threshold above which firms face higher taxes or are subject to onerous regulations but these taxes or regulations are imperfectly enforced (say because of an inefficient bureaucracy), the outcome will be a right-skewed firm size distribution instead of a bimodal distribution. The "constrained large firm" model also predicts that the marginal return to resources will be *lower* in small firms compared to large firms.

With this theoretical framework in mind, the next three sections discuss facts about the distribution of firm size, the patterns of returns to capital and labor inputs, and whether one observes a "kink" in the distribution of firms at points where one might expect to find such a kink based on prominent regulations affecting large firms.

The Size Distribution of Firms: India, Indonesia, Mexico

Much of the work on the size distribution of firms in low-income countries looks at partial datasets—in particular, data that lacks much or any coverage of the informal sector. But we were able to obtain complete, representative microdata on the entire manufacturing sector for India, Indonesia, and Mexico, including both formal and informal enterprises. We use microdata from the manufacturing sector in the Mexican Economic Census, the Indonesian Economic Census, and India's Annual Survey of Industries and National Sample Survey (Schedule 2). For each country, we present the most recent wave of data available.[1] The key variable we use is the number of workers, including unpaid family workers. These countries do differ substantially in terms of GDP per capita in the year our data was collected: specifically, real per capita GDP in purchasing power parity terms was $3,700 in India (2011), $3,600 in Indonesia (2006), and $14,200 (2008) in Mexico (World Bank World Development Indicators 2013).

Figure 1 presents the distribution of firm size in bins of ten workers. The first row presents the distribution for India (2011), the second row for Indonesia (2006), and the third row for Mexico (2008). The first column presents the size distribution of *all* firms (since the tails have so few firms, we truncate the range of the graph at size 200 to make it visible). The next columns focus on different ranges of the data so that the patterns are more easily visible. Specifically, we restrict the range to firms with 10 to 200 workers (column 2), 20 to 200 workers (column 3), 50 to 200 workers (column 4), and 200 to 3,000 workers (column 5). The figure shows that the vast majority of firms in all three countries are small, with no evidence of bimodality in the firm size distribution. In all cases, the distribution of firm size is right skewed and generally smoothly declining in firm size, with no evidence of bimodality or discontinuity. This is the first key fact: there is no evidence of a "missing middle" of firms when one examines the raw distributions of firm size in any of these three countries.

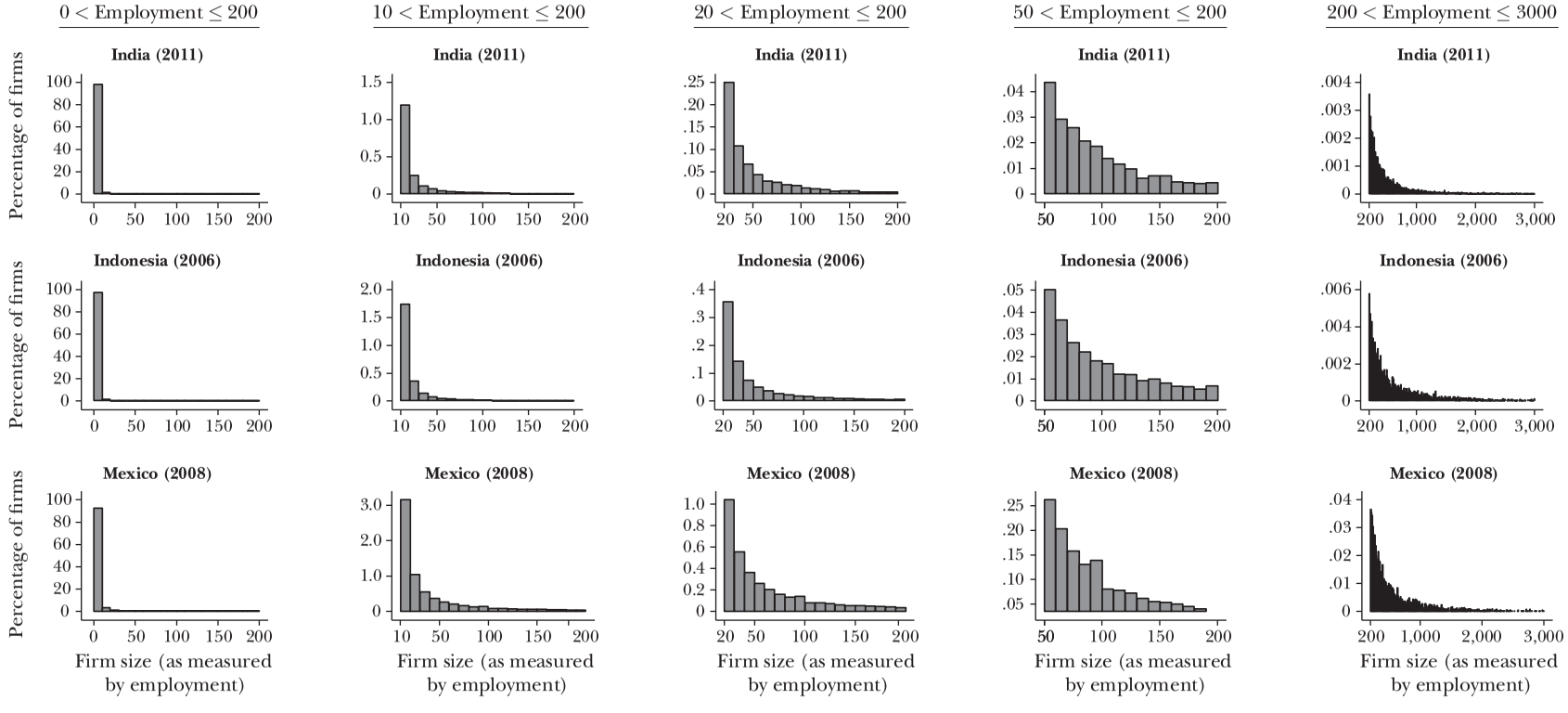
Comparing the three countries, the fraction of small firms is lower in Mexico than in India and Indonesia. About 90 percent of firms in Mexico employ less than 10 workers. In India and Indonesia, the fraction of firms with less than 10 workers is almost visually indistinguishable from 100 percent. Given that the GDP per capita in Mexico is about four times higher than in India and Indonesia, the pattern suggests that development is associated with a decline in the skew of the firm size distribution.

For comparison, the US distribution of manufacturing firms has as its mode mid-sized firms with about 45 employees (Hsieh and Klenow 2014, Figure 14), whereas the mode in each of these countries are firms with one worker. There are

[1] The Mexican Economic Census is a complete enumeration of fixed establishments. The Indonesian Economics Census is a complete enumeration of all establishments with 20 or more employees (medium and large firms) and a random 5 percent sample of establishments with 20 or fewer employees (small firms). We combine these two samples to get a complete picture of the entire Indonesian manufacturing sector, including both formal and informal enterprises. The Indian Annual Survey of Industries is a census of formal establishments with more than 100 employees and a random survey of formal establishments with less than 100 employees. The National Sample Survey is a survey of informal establishments. We combine the data from the two surveys when we present evidence from India.

Figure 1

Distribution of Firm Size as Measured by Number of Workers



Source: We use microdata from the manufacturing sector in the Mexican Economic Census, the Indonesian Economic Census, and India's Annual Survey of Industries and National Sample Survey (Schedule 2). See footnote 1.

Notes: The figure shows distribution of firm size measured by the number of workers. The bin size is 10 workers, and each bin contains the upper bound and not the lower bound. For all graphs, the y-axis indicates the share of all firms in the specified size. The different columns truncate the x-axis in different ways to focus on different parts of the distribution.

fewer mid-sized firms in India, Indonesia, and Mexico than in the United States. But the overwhelming fact is that most firms are small in our three developing countries—large firms are also missing, and there is no missing middle in the sense of a bimodal distribution.

The Distribution of Average Return to Inputs

Even though the evidence shows no evidence of bimodality in the firm size distribution, another approach is to look also for evidence for the supposed forces that would lie behind the purported bimodality in the firm size distribution. For example, models where capital constraints generate a bimodal size distribution also imply that the return to capital is bimodal: that is, small unconstrained firms and large unconstrained firms would have low returns to capital, but firms that are hitting the constraint—the firms that would have grown to be the allegedly "missing" mid-sized firms—would have much higher returns. Other theories, such as those based on the idea that large firms face higher labor costs, those based on the notion that large firms have better access to intermediate inputs, and those based on De Soto's (1989) hypothesis that the property rights of formal firms are better protected, similarly imply that the return to all the resources used by the firm is bimodal, with one set of unconstrained firms with low returns and another set of constrained firms with high returns.

We do not directly measure the marginal return to inputs, but we can measure the average return to capital, labor, and intermediate inputs. However, interpreting the findings of such an exercise requires some caution, because different production functions have different implications for the relationship between average and marginal products. Here are several possibilities.

First, if revenue is generated by a single Cobb–Douglas function of the factor inputs, factor-intensities and markups are constant, and fixed costs are zero, the marginal return of each input is proportional to its average product.

Second, in a dual technology model with high-capital intensity and low-capital intensity technologies and a Cobb–Douglas production function, the average product of capital will be generally lower, and the average product of labor higher, in firms that utilize more capital-intensive technologies. In addition, the average product of the sum of variable and fixed capital will be lower in firms with high fixed-cost technologies. In addition, if some of capital measured in the data includes the fixed cost of the modern technology, this would further lower the average product of capital (the sum of fixed and variable capital) in large firms.

Although a dual technology model with Cobb–Douglas production technologies predicts that the average product of capital is lower for large modern firms, this prediction does not generalize to arbitrary production functions. For example, imagine that the production function for the two technologies is Leontief (so the marginal product of capital and labor for a given technology is zero), and the average product of capital and labor with the modern technology is higher than in firms using traditional technologies. Here, although the average product of capital

and labor in the traditional firm is low, the marginal return from switching to the modern technology is presumably high.

Third, if markups vary across firms—say, because more-productive firms produce higher-quality products that are more price inelastic—then the average product of capital and labor will be higher in larger firms that produce high-quality products.

Figure 2 looks for evidence of bimodality in the average product of factor inputs. Specifically, it plots the distribution of the log ratios of value-added to capital (column 1), value-added to labor (column 2), and gross output to the value of intermediate inputs (column 3). We truncate the top and bottom percentile to make the histograms more easily viewable. Note that we do not have a comparable capital series for large and small Indonesian firms, so we omit Indonesia from the first column. The distributions of the average product of capital and labor are not bimodal, as suggested by theories of capital constraints or labor costs. The distributions of the average product of intermediate inputs are also not bimodal, but are roughly right skewed. This pattern is consistent with theories where a large number of firms use few intermediate inputs.

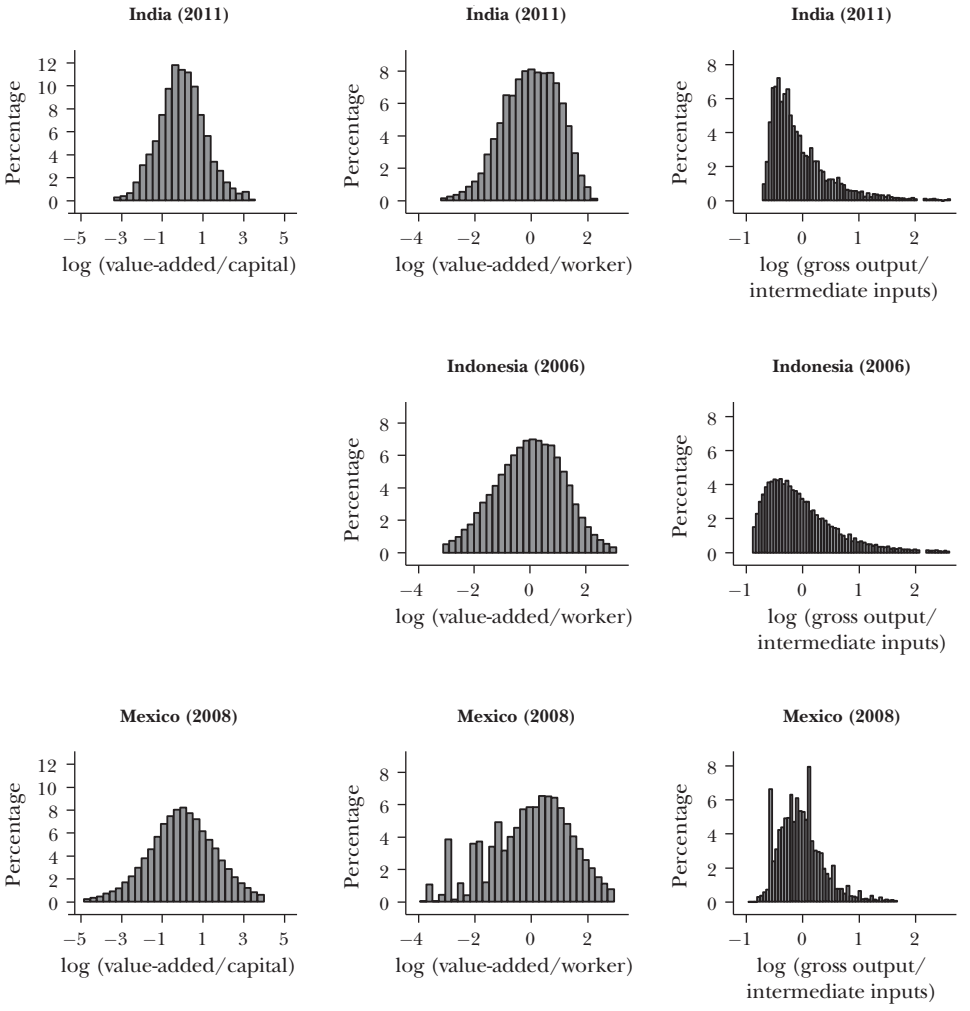
Figure 3 looks directly at the correlation between the average product of inputs and firm size (measured as log employment). The first column presents the nonparametric relationship (from a regression using the approach of Fan 1992) with the average product of capital as the independent variable and firm employment as the explanatory variable. The dashed lines in each figure represent 95 percent confidence intervals.

As can be seen, the average product of capital is increasing with firm employment. If the average product of capital is proportional to the marginal product of capital, this suggests that the marginal cost of capital is *higher* in large firms relative to small firms. This fact is inconsistent with a widely held view that the return to capital is high in small firms in poor countries. Put differently, if the return to capital is high in small firms, the evidence in Figure 3 suggests that the return to capital in large firms is even higher.

This fact would be surprising if one believed in the dual technology view that large firms operate capital-intensive technologies with high fixed costs. For the large firms to have higher average products of capital in this story, it would either need to be that the modern firms have high average products of capital but low marginal products (so they use an L-shaped Leontief production function or close to it), or that modern firms also face higher marginal capital costs and the net effect of the higher marginal cost of capital outweighs the effect of capital-intensive technologies and the higher fixed cost on the average product of capital. Neither of these stories is theoretically impossible, but they are not necessarily what one would have expected from most standard versions of these theories.

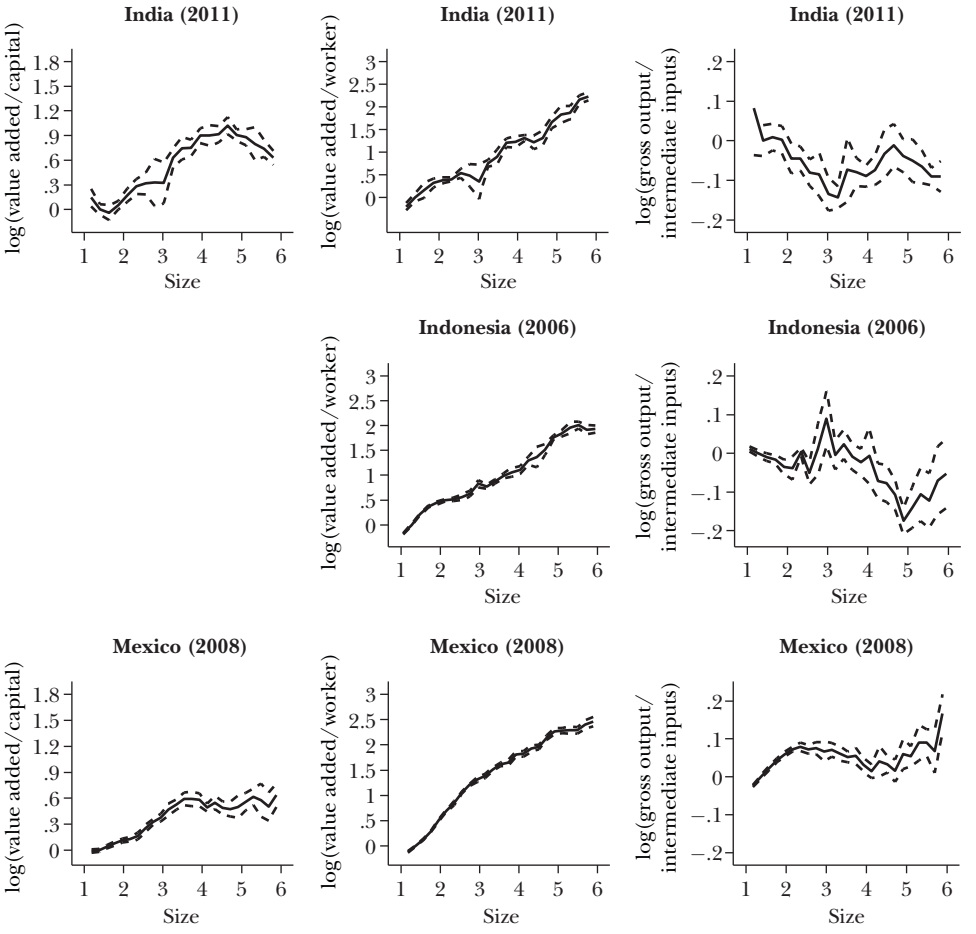
The second column in Figure 3 plots the nonparametric relationship between the average product of labor with firm employment. The relationship is positive, as if the marginal cost of labor inputs is increasing with firm employment. This is a prediction of the Banerjee–Duflo (2005, 2011) dual technology model if modern technologies are more capital-intensive, although we note that this model is not supported by the evidence that the average product of capital is also higher in

Figure 2
Distribution of Average Products



Source: See Figure 1 for sources.
Notes: The figure presents distributions of the demeaned log average product of capital (column 1), log average product of labor (column 2), and the log ratio of revenues to intermediate inputs (column 3). The bin size is the same in each column and chosen such that the histograms for Mexico have 50 bins. We drop the bottom and top 1 percent in each sample. In Indonesia, the questionnaire administered to firms with fewer than 20 employees asked about capital differently than the questionnaire administered to firms with 20 or more employees so we cannot construct a consistent measure of the capital stock across these two samples. In an online Appendix available with this paper at http://e-jep.org, we show qualitatively similar patterns when we separately examine firms with 20 or more employees and firms with less than 20 employees.

Figure 3
Average Product and Firm Size
(size measured as log(employment))



Source: See Figure 1 for sources.
Notes: Figure shows local linear regressions of log average product on log employment. We normalize the y-axis by taking the value of the function at log(employment) = 1.4 to be zero. Dashed lines represent 95 percent confidence bounds. Size is measured as log (employment).

larger firms. The fact that the average product of labor is higher in larger firms also supports the story by Harris and Todaro (1970), McKinsey & Co. (2005), and Levy (2008) that large firms pay above-market labor costs, except that there is no clear discontinuity in this relationship. We note that La Porta and Shleifer (2008, tables X, XI) also find that average labor productivity increases with firm size in the World Bank Enterprise Surveys, except that we interpret the positive relationship as indicating that large firms behave as if they face higher marginal labor costs.

An alternative explanation for why the average product of labor and capital might be higher in large firms is that larger firms charge higher markups. De Loecker

and Warzynski (2012) show that in this case, the markup will be proportional to the ratio of gross output to spending on intermediate inputs. The third column in Figure 3 shows the relationship between revenue per intermediate input and firm size. Here, there is no evidence that the average product of intermediate inputs is higher or lower in large firms relative to small firms. If the marginal cost of intermediate inputs is the same for small versus large firms, Figure 3 indicates that markups are no higher in large firms. In turn, this suggests that the higher average product of capital and labor for large firms do not reflect higher markups, but rather higher marginal costs. Of course, it is possible that large firms charge higher markups but the effect of the higher markup on the average product of intermediate inputs is exactly offset by a lower marginal cost of material inputs.

Together, Figure 2 and Figure 3 produce our second set of stylized facts: the average product of labor and capital is lower in small firms than in large firms, and there is no obvious bimodality in any of these distributions. In this sense, these developing countries look much like the United States, where it also the case that larger firms appear to be more productive (Brown and Medoff 1989; Idson and Oi 1999; Hsieh and Klenow 2014). But the fact that the larger firms are more productive is at odds with the frequent view in the development space that the key constraints are with small firms. If we believed there was a "missing middle" of constrained firms with high returns that could not grow, many models would have predicted either small firms to have higher average products, or potentially an inverted U-shape, with a mass of high average product and constrained firms in the middle of the distribution. However, the data does not support this view.

Discontinuities in Firm Size from Tax and Regulatory Notches

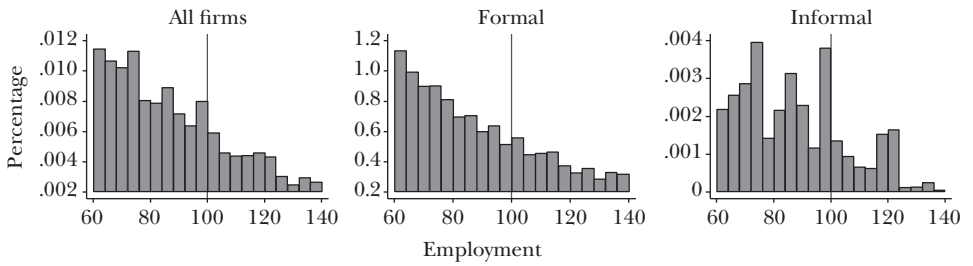
A frequently cited reason for the existence of the purported missing middle is the existence of a tax or regulatory notch that affects firms above a certain size. In this situation, one might expect to find a bunching of firms at the size determined by the regulation or tax, and a missing distribution of firms just above the kink point. There are many possible examples: firms with few employees are frequently exempt from labor regulations (such as benefits and hiring and firing costs), and there is often preferential tax treatment for firms below a certain size threshold.

In our setting, we examine three such notches. In India, the Industrial Disputes Act requires firms employing more than 100 workmen (that is, 100 workers other than managers) to obtain government permission before laying off workers. This suggests a discrete notch in labor regulation at 100 nonmanagerial employees, which some have suggested is an important reason for the small size of firms in India (for example, *The Economist* 2007; Krueger 2013).[2]

In Indonesia, firms below a given revenue threshold are exempt from paying the 10 percent value-added tax. This again creates a discrete notch where we would

[2] Besley and Burgess (2004) offer an empirical assessment of the importance of Indian employment law more generally, and Guner, Ventura, and Xu (2007) present a quantitative model.

Figure 4
Distribution of Indian Firm Size and Labor Regulations
(size as measured by employment)



Source: The data comes from India's Annual Survey of Industries and National Sample Survey.
Notes: Figure shows size distribution of Indian firms around firms with 100 workers. We exclude managerial workers in the sample of formal firms from the Annual Survey of Industries. The bin size is four workers, and each bin contains the upper and not the lower bound.

expect bunching of firms below this cutoff (Kleven and Waseem 2013). The cutoff is not indexed for inflation; instead, it is adjusted discretely by the government periodically. Adjustments were made in 1992 (50 percent nominal increase); 1995 (100 percent nominal increase); 2001 (50 percent nominal increase); and 2004 (66 percent nominal increase). In 2006, the year of our census, the threshold was still where it was in 2004, at 600 million Indonesian rupiah (about $65,000 in US dollars); it was not raised again until 2013.

In Mexico, we focus on the revenue threshold due to the simplified tax regime for small firms. From 1998 until 2013, firms with sales below 2 million pesos (about $125,000 in 2008) pay a flat tax of about 2 percent of their sales and are exempt from payroll taxes, income taxes, and value-added taxes. Firms above the 2 million peso threshold are subject to a 15 percent value-added tax, a 38 percent income tax, and a 35 percent payroll tax.[3]

Although a casual examination of the histograms in Figure 1 does not suggest any discontinuities, it is possible that if we zoom in on these kinds of notches in the regulator environment we will see something. Figure 4 shows the distribution of nonmanagerial employment in India in 2011. (For nonformal firms, we do not have employment separately by managerial and nonmanagerial categories, so we report total employment for these firms.) We zoom in on the range from 60 to 140 nonmanagerial employees so we can focus on the 100 worker cutoff (shown by the vertical line). We focus on the distribution of all firms (left panel) but also show the distribution of formal firms (center panel) and informal firms (right panel). Since the regulation applies only to formal firms, it is possible that even if the regulation doesn't affect the total firm size distribution, it affects the decision to switch from formal to informal.

[3] The tax rate under the simplified tax regime (Repecos) varies across states but averages 2 percent. The simplified tax regime is administered at the state level. See Sánchez-Vela and Valero-Gill (2011).

Visually inspecting the leftmost panel of Figure 4, there is perhaps a slight bit of bunching at 100 employees, but it is small. In the bin just below the cutoff (97–100 workers), there are a total of 1,370 firms. In the next bin (101–104 firms), there are 1,013 firms. Even abstracting from the fact that the overall distribution is downward sloping, so one would expect fewer firms with 101–104 workers than 97–100 workers, the difference amounts to at most two-tenths of 1 percent of all Indian firms—a few hundred firms in all of India out of the 17,177,148 total firms.

Inspecting the central panel, there is no discontinuity whatsoever in formal firms; if anything, there is a slight spike of firms with more than 100 workers. There *is* bunching of informal firms just below 100, but again the economic magnitude is small: the difference between the number of firms with 97–100 workers and 101–104 workers is a little more than two-tenths of 1 percent of all informal firms—at most about 418 firms in total for all of India. Thus, while there may be a small amount of bunching induced by the regulation, the amount we can detect in the data does not suggest that it is an important driver of small firm size in India.

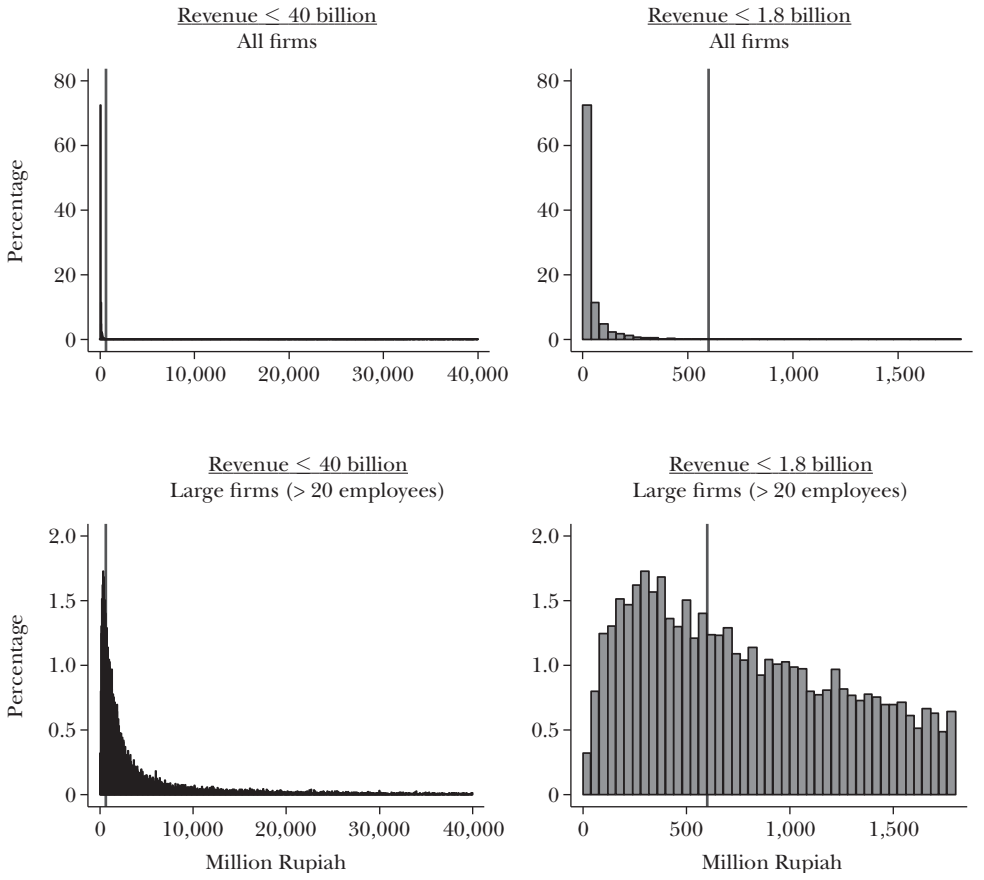
For Indonesia, we focus on the discontinuity in revenue at 600 million Indonesian rupiah. One might expect more heaping in revenue than in employment, since presumably revenue is easier to adjust in order to stay under the threshold (for example, firms may have some flexibility in deciding what year to realize revenue from a given sale). The top left panel of Figure 5 shows the distribution of revenue for all Indonesian firms; the bottom panel zooms in on firms with more than 20 employees. Since virtually no firms with fewer than 20 employees have revenue close to 600 million rupiah, the figures in the bottom panel are easier to read. The two left panels show the distribution for all firms with less than 40 billion Indonesian rupiah in revenue (about $4.3 million in US dollars). To zoom in closer to the discontinuity, the right panels consider only firms with less than 1.8 billion in Indonesian rupiah in revenue (about $200,000 in US dollars).

The figure (particularly the bottom-right), which zooms in on the relevant part of the cutoff shows no bunching at the discontinuity in eligibility to pay the value-added tax. Because virtually all firms in the relevant part of the revenue distribution are from the large firm survey, which is conducted annually, we can re-generate the zoomed-in graph for large firms for each year back to 1990 and check for any changes in firm size associated with the different cutoffs that were in place over the years. In an online Appendix available with this paper at http://e-jep.org, we show this figure, with the relevant cutoff line shown each year. We never find any substantial bunching at the discontinuities.

Next, we focus on a potential discontinuity in Mexico due to a simplified tax regime for firms with less revenue than 2 million pesos. The left panel in Figure 6 shows the distribution of sales for all Mexican firms with less than 6 million pesos in sales; the right panel zooms in on firms with sales between 1 and 4 million. As can be seen, there is no bunching at 2 million pesos (the vertical line) after which firms legally switch from a flat 2 percent sales tax regime to the combination of the value-added tax, income tax, and payroll tax regime.

Combined, the evidence from India, Mexico, and Indonesia suggest a third important fact: at least as we can measure it in our data, we do not see important

Figure 5
Distribution of Indonesian Firm Size and the VAT Threshold

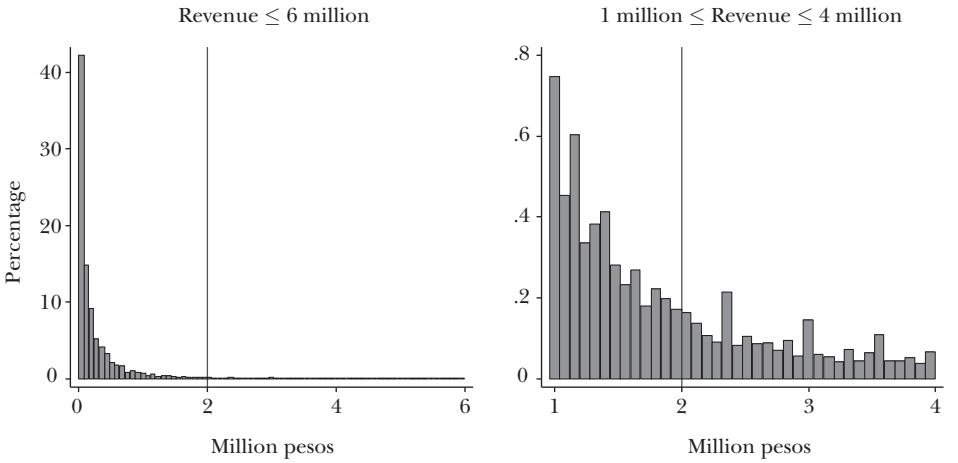


Source: Indonesian Economic Census. See footnote 1.
Notes: Figure shows distribution of the revenue of Indonesian firms. The vertical line (600 million rupiah) denotes the VAT threshold. The bin size is 40 million rupiah, and each bin contains the upper bound but not the lower bound.

discontinuities in firm size, either in general when looking at the distributions or when we zoom in around the places where one would expect them a priori based on regulatory and tax notches.

Of the other papers in the literature that have looked at similar notches, a small number of papers have found some bunching, but in most of these cases the quantitative magnitude of the bunching is small. For example, Onji (2009) examines the introduction of a value-added tax threshold in Japan and looks for bunching around the threshold, much as we do in Indonesia. Although he does find evidence of bunching, the magnitude appears very small: the share of firms below the threshold falls by less than 0.5 percent. Similarly, Schivardi and Torrini (2008) examine a discontinuity in Italian employment regulations that applies to

Figure 6
Distribution of Mexican Firm Size and the Simplified Tax Regime Threshold



Source: We use microdata from the manufacturing sector in the Mexican Economic Census. See footnote 1.
Notes: Figure shows the distribution of revenues of Mexican firms. The vertical line (2 million pesos) denotes the threshold of a simplified tax regime for small firms. The bin size is 80,000 pesos, and each bin contains the upper bound and not the lower bound.

firms greater than 15, much as we do in India. They estimate that after removing the threshold, average firm size would increase by less than 1 percent. Garicano, LeLarge, and Van Reenen (2013) estimate the impact of lifting French regulations that apply to firms with more than 50 workers. Their model implies that about 3 percent of workers are reallocated from firms of size 50 or more to firms of size 49 and below. Under the assumption of flexible wages, their model estimates an output loss of 0.16 percent of GDP associated with this change, although the assumption of fully inflexible wages yields substantially larger estimates. Thus, the evidence we present from India, Indonesia, and Mexico is consistent with the small magnitudes of bunching observed in other contexts.

How Did the "Missing Middle" Misconception Arise?

Given the facts presented in this paper, a natural question is: Where does the misconception about the missing middle—in the sense of the bimodality of the distribution—come from? We suggest it comes from the combination of two transformations that had previously been made to the available data.

In the economics literature, the main evidence typically cited for the missing middle is table 1 of Tybout (2000). In that table, Tybout shows the distribution of employment shares across plant sizes for manufacturing firms for 19 countries. For most countries in the table, he shows the number of workers in firms of size 1–9,

10–49, and 50+; for a few countries, he includes five or six bins of firms instead. The data in the table is, in turn, drawn from other calculations done by a variety of other authors, most notably Leidholm and Mead (1987), who compile similar tabulations from other studies. The "missing middle" refers to the fact that in most developing countries, there is substantially lower employment share in the mid-sized category (that is, firms of 10–49 employees) than in either the small category (fewer than 10 employees) or the large category (50 or more employees). For example, in Indonesia in 1977, the table shows 77 percent of total manufacturing employment is in firms of size 1–9, 7 percent is in firms of size 10–49, and 16 percent is in firms of size 50 or more.

There are two important differences between the facts reported in the Tybout (2000) and Liedholm and Mead (1987) tables and the facts we present here. First, these earlier tables refer to the employment share—that is, what fraction of total manufacturing employment comes from firms of a given size—rather than the distribution of firm size. The employment share distribution reveals in what size firm a typical worker in the economy works, whereas the firm size distribution reveals the distribution of firms. To compute the employment share statistic, one multiplies the number of firms in each bin with the average employment size of firms in the bin. While the employment share statistic is interesting for understanding the aggregate distribution of employment, most theories about the existence of the missing middle discussed above are about firm size itself. For example, theories about tax and regulatory notches and credits constraints are all about whether firms should grow above a certain size, not about the employment share in aggregate.

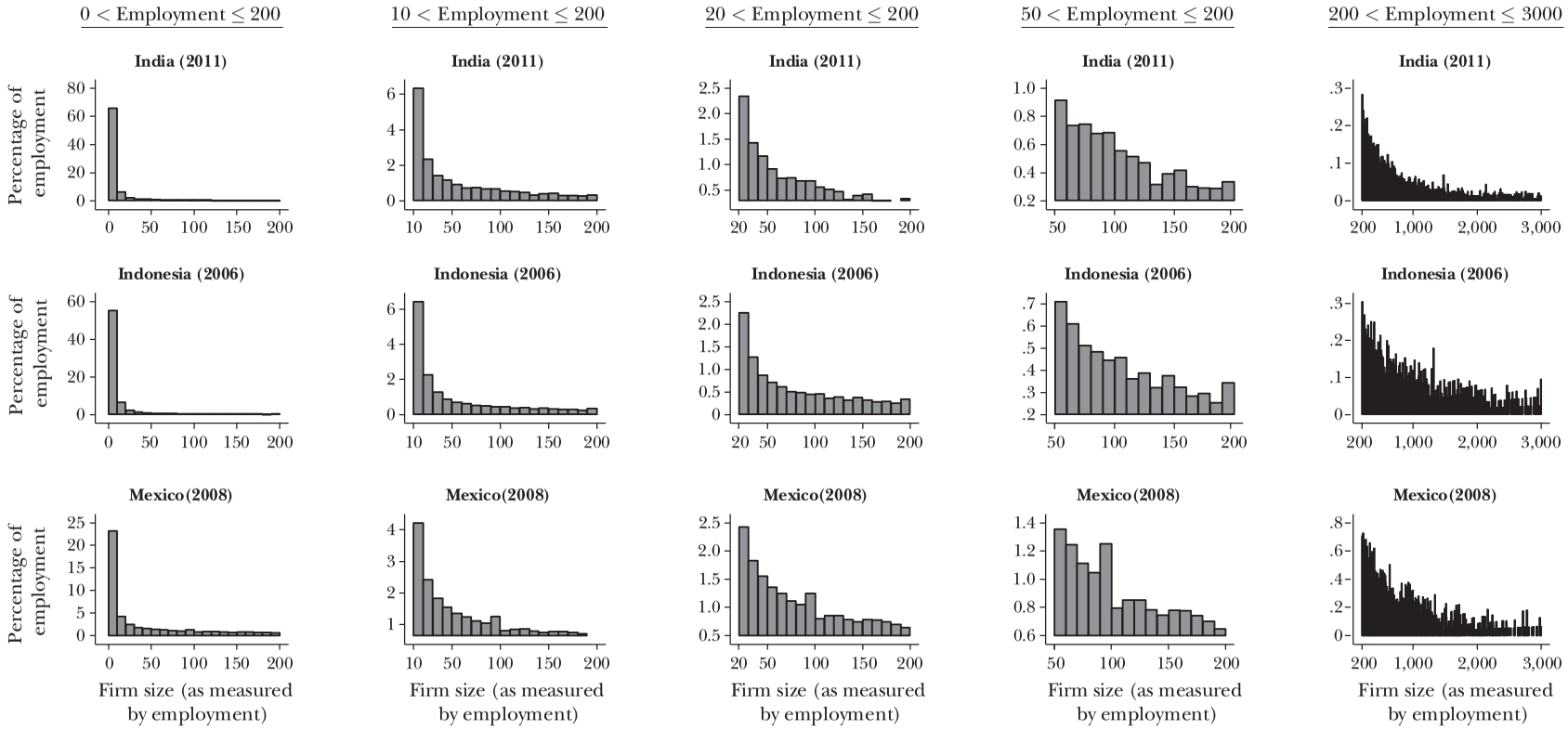
We start with our data from Figure 1 about the distribution of firm size, and transform it into one on the distribution of employment share by firm size, as shown in Figure 7. This transformation, in itself, does not create a missing middle. Figure 7 shows the distribution of employment share, analogous to what is shown in Figure 1 for the distribution of firm size. Although it is shifted to the right (mechanically) from the firm size distribution, it still appears unimodal in all three countries.

Second, in the earlier literature, because of data limitations, the tabulations about employment share are typically binned into a small number of groups: for most countries, the authors report the totals for three bins, firms with less than 10 employees, 10–49 employees, and 50 or more employees.

To see what difference this choice of bins makes, Table 1 reports the distribution of firm size (Panel A) and the distribution of employment shares (Panel B) from our data, grouped into these same three categories. Panel A shows that the firm size distribution, even when binned, shows the same pattern as the histograms—the density of firms is monotonically declining in firm size. But Panel B shows that when we apply the arbitrary binning transformation to the *employment share* distribution, the pattern from Tybout (2000) re-emerges. For example, in Indonesia in 2006, 54 percent of total employment is in firms with 1–9 employees, 12 percent is in firms with 10–49 employees, and 34 percent is in firms with 50 or more employees—that is, the missing middle phenomenon now appears. Thus, the existing facts about the missing middle seem to come from the combination of these two transformations to

Figure 7

Distribution of Employment Share by Firm Size



Source: See Figure 1 for sources.

Notes: The figure shows the distribution of employment share across firms of different size as measured by number of employees. The bin size is 10 workers, and each bin contains the upper bound and not the lower bound. The different columns truncate the x-axis in different ways to focus on different parts of the distribution.

Table 1
Distribution of Firms and Employment Shares in Bins

Firm Size (Employment)	*India 2011*	*Indonesia 2006*	*Mexico 2008*
Panel A: Distribution of Firm Size			
1–9	97.88	96.78	91.74
10–49	1.85	2.83	5.85
50+	0.28	0.39	2.41
Panel B: Distribution of Employment Share by Firm Size			
1–9	64.77	53.95	22.45
10–49	12.10	12.04	10.55
50+	23.13	34.01	66.99

Source: See Figure 1 for sources.

the data: the transformation from the distribution of firms to the aggregate employment share, and the arbitrary binning of the employment share distribution.

Implications for Theories of Development

Ultimately, the main reason that economists and policymakers care about the size distribution of firms in developing countries is what it may reveal about alternative theories of firm development and in turn what that implies for policy. The lack of a "missing middle"—that is, the lack of bimodality in the size distribution of firms—suggests that neither the "small firms are constrained" nor the dual economy theories of development are correct, at least not in their simplest form. In addition, the fact that the average returns to capital and labor are lower in small firms suggests that the view that small firms are constrained—say because they have difficulty accessing capital and thus have a high return to capital—is inconsistent with the simple versions of these models.

What would it take to reconcile the models to the facts? One tempting alternative is to explore the implications of more capital-intensive production technologies for larger firms: after all, it is likely that large firms use more capital-intensive technologies that, all else equal, would tend to lower the average product of capital in large firms. To make a dual-economy model fit the facts without also asserting that large firms are constrained, one would need the high-productivity firms to have high average products of capital but low marginal products of capital, and vice-versa. Moreover, one would need substantial heterogeneity across firms in the employment size of such high capital-intensity firms in order to avoid generating bimodality in the firm size distribution. It is theoretically possible to write down such models, but the facts presented here substantially constrain the types of models one can write down.

An alternative theory that fits all our facts is the view that large firms are constrained, perhaps by taxes or regulations, but that implementation of these

barriers is imperfect. Levy (2008), for example, documents that the vast majority of small and mid-size firms in Mexico evade the 35 percent payroll tax. This view is consistent with the evidence that there is little meaningful discontinuity in the size distribution, even at thresholds at which one would expect a discontinuity if taxes or regulations were perfectly enforced. This view also implies that the problem is unlikely to be the (relatively easy to fix) notch in the tax or regulatory code; rather, it suggests that it a confluence of factors make enforcement of such rules easier in larger firms so that costs from regulation are rising smoothly in firm size. Another key prediction of the "large firms are constrained" view is that the marginal return to resources would be higher in large firms, which is supported by the fact that the average product of capital and labor is consistently higher in large firms when compared to small firms. If so, the fact that the firm size distribution in poor countries is dominated by small firms is explained by firms *choosing* not to exert the effort necessary to grow because their marginal cost would rise if they did grow.

In sum, the evidence we present in this paper suggests that the problem of economic development in low-income and middle-income countries is how to relieve the differential constraints faced by large firms, not how to relax the constraints faced by small firms. Indeed, this view suggests that programs such as microcredit or simplified tax regimes that benefit only small firms may worsen the development problem by further increasing the incentive to stay small.

■ *We thank Arianna Ornaghi, Nick Tsivanidis, Pedro-José Martínez-Alanis, and especially Donghee Jo for outstanding research assistance and Abhijit Banerjee, Esther Duflo, Santiago Levy, Pete Klenow, and John Van Reenan for helpful comments. The results from the Mexican Census have been screened by Mexico's INEGI to ensure no confidential data is released.*

References

Banerjee, Abhijit V., and Esther Duflo. 2005. "Growth Theory through the Lens of Development Economics." Chap. 7 in *Handbook of Economic Growth*, volume 1, edited by Philippe Aghion and Steven Durlauf. Elsevier.

Banerjee, Abhijit V., and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* Public Affairs.

Besley, Timothy, and Robin Burgess. 2004. "Can Labor Regulations Hinder Economic Performance? Evidence from India." *Quarterly Journal of Economics* 119(1): 91–134.

Bloom, Nicholas, and John Van Reenan. 2007. "Measuring and Explaining Management Practices across Firms and Countries." *Quarterly Journal of Economics* 122(4): 1351–1408.

Brown, Charles, and James L. Medoff. 1989. "The Employer Size-Wage Effect." NBER Working Paper 2870.

De Loecker, Jan, and Frederic Warzynski. 2012. "Markups and Firm-level Export Status." *American Economic Review* 102(6): 2437–71.

De Mel, Suresh, David McKenzie, and Christopher Woodruff. 2008. "Returns to Capital in Microenterprises: Evidence from a Field Experiment." *Quarterly Journal of Economics* 123(4): 1329–72.

De Soto, Hernando. 1989. *The Other Path: The*

Invisible Revolution in the Third World. New York, NY: Harper and Row.

Economist, The. 2007. "India's Economy: A Himalayan Challenge." October 11. http://www.economist.com/node/9955756.

Fan, Jianqing. 1992. "Design-adaptive Nonparametric Regression." *Journal of the American Statistical Association* 87(420): 998–1004.

Garicano, Luis, Claire LeLarge, and John Van Reenen. 2013. "Firm Size Distortions and the Productivity Distribution: Evidence from France." NBER Working Paper 18841.

Guner, Nezih, Gustavo Ventura, Xu Yi. 2008. "Macroeconomic Implications of Size-Dependent Policies." *Review of Economic Dynamics* 11(4): 721–44.

Harris, John R., and Michael P. Todaro. 1970. "Migration, Unemployment and Development: A Two-Sector Analysis." *American Economic Review* 60(1): 126–42.

Hsieh, Chang-Tai, and Peter Klenow. 2014. "The Lifecycle of Manufacturing Plants in India and Mexico." http://faculty.chicagobooth.edu/chang-tai.hsieh/research/HsiehKlenow_LifeCycle.pdf. (Forthcoming in *The Quarterly Journal of Economics*).

Idson, Todd L., and Walter Y. Oi. 1999. "Workers Are More Productive in Large Firms." *American Economic Review* 89(2): 104–108.

Kleven, Henrik J., and Mazhar Waseem. 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *Quarterly Journal of Economics* 128(2): 669–723.

Kremer, Michael, Jean Lee, Jonathan Robinson, and Olga Rostapshova. 2013. "Behavioral Biases and Firm Behavior: Evidence from Kenyan Retail Shops." *American Economic Review* 103(3): 362–68.

Krueger, Anne O. 2013. "The Missing Middle." Chap. 9 in *Economic Reform in India: Challenges, Prospects, and Lessons* edited by Nicholas C. Hope, Anjini Kochar, Roger Noll, and T. N. Srinivasan, Cambridge University Press.

La Porta, Rafael, and Andrei Shleifer. 2008. "The Unofficial Economy and Economic Development." *Brookings Papers on Economic Activity*, no. 2 (Fall), pp. 275–352.

Leidholm, Carl, and Donald C. Mead. 1987. "Small Scale Industries in Developing Countries: Empirical Evidence and Policy Implications." Food Security International Development Papers 54062, Michigan State University.

Levy, Santiago. 2008. *Good Intentions, Bad Outcomes: Social Policy, Informality, and Economic Growth in Mexico*. Washington DC: Brookings Institution.

Lewis, Arthur. 1954. "Economic Development with Unlimited Supplies of Labour." *Manchester School* 22(2): 139–91.

McKinsey & Co. 2005. "Eliminando as barreiras ao crescimento econômico e à economia informal no Brasil."

McKinsey Global Institute. 2001. *India: The Growth Imperative*.

Onji, Kazuki. 2009. "The Response of Firms to Eligibility Thresholds: Evidence from the Japanese Value-Added Tax." *Journal of Public Economics* 93(5–6): 766–75.

Rauch, James E. 1991. "Modeling the Informal Sector Formally." *Journal of Development Economics* 35(1): 33–47.

Sánchez-Vela, Claudia, and Jorge Valero-Gil. 2011. "The Effect of Firm-Size Dependent Policies on the Economy: The Case of the Repecos Law in Mexico." IDB Working Paper 216.

Schivardi, Fabiano, and Roberto Torrini. 2008. "Identifying the Effects of Firing Restrictions through Size-Contingent Differences in Regulation." *Labour Economics* 15(3): 482–511.

Tybout, James R. 2000. "Manufacturing Firms in Developing Countries: How Well Do They Do, And Why?" *Journal of Economic Literature* 38(1): 11–44.

Udry Christopher, and Santosh Anagol. 2006. "The Return to Capital in Ghana." *American Economic Review* 96(2): 388–93.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 109–126

Informality and Development[†]

Rafael La Porta and Andrei Shleifer

In developing countries, informal firms account for up to half of economic activity. They provide livelihood for billions of people. Yet their role in economic development remains controversial.

Some, like Hernando De Soto (1989, 2000), see informal firms as an untapped reservoir of entrepreneurial energy, held back by government regulations. In this view, unleashing this energy by reducing entry regulations or improving property rights would fuel growth and development. Others, like Levy (2008), take a more cynical view, stressing the advantages enjoyed by informal firms and workers from avoiding taxes and regulations. A report from the McKinsey Global Institute describes informal firms as parasites competing unfairly with law-abiding formal firms (Farrell 2004). In this view, informality should be suppressed, not unleashed. Still others follow the development tradition of Lewis (1954), Harris and Todaro (1970), and more recently Rauch (1991) and see informality as a byproduct of poverty. From this dual perspective, formal and informal firms are fundamentally different. Productive formal entrepreneurs pay taxes and bear the cost of government regulation to reach new customers, raise capital, and access public goods. These entrepreneurs are often educated and find it more profitable to run bigger formal firms rather than the smaller informal ones. In contrast, informal entrepreneurs are typically uneducated and unproductive, and they run small businesses producing low-quality products for low-income customers using little capital and

■ *Rafael La Porta is Noble Foundation Professor of Finance, Tuck School of Business at Dartmouth, Hanover, New Hampshire. Andrei Shleifer is Professor of Economics, Harvard University, Cambridge, Massachusetts. Their email addresses are rafael.laporta@tuck.dartmouth.edu and ashleifer@harvard.edu.*

[†] To access the Appendix, visit
http://dx.doi.org/10.1257/jep.28.3.109

doi=10.1257/jep.28.3.109

adding little value. Informal firms do not threaten formal firms, and the increase in firm value that the informal entrepreneurs or managers could realize by operating formally is too small to offset the additional costs from taxes and regulations. In this dual view, development comes from formal firms, and their expansion as the economy modernizes eventually dooms the informal economy.

In this paper, we assess these perspectives. To begin, we establish five critical facts about the informal economy. First, it is huge, especially in developing countries. Second, it has extremely low productivity compared to the formal economy: informal firms are typically small, inefficient, and run by poorly educated entrepreneurs. Third, although avoidance of taxes and regulations is an important reason for informality, the productivity of informal firms is too low for them to thrive in the formal sector. Lowering registration costs neither brings many informal firms into the formal sector, nor unleashes economic growth. Fourth, the informal economy is largely disconnected from the formal economy. Informal firms rarely transition to formality, and continue their existence, often for years or even decades, without much growth or improvement. Fifth, as countries grow and develop, the informal economy eventually shrinks, and the formal economy comes to dominate economic life.

We argue that the evidence is most consistent with dual models of informality. We first review these models and stress the supply and demand factors that keep informal and formal economies separate. We then address the critical question of how the informal economy shrinks. In dual models, economic growth comes from the formal sector: that is, from firms run by educated entrepreneurs and exhibiting much higher levels of productivity. The expansion of the formal sector leads to the decline of the informal sector in relative and eventually absolute terms, although informal employment can remain high for a long time, especially when labor force growth is high. A few informal firms convert to formality, but more generally they disappear because they cannot compete with the much more-productive formal firms. Our evidence is not particularly supportive of either De Soto's romantic view of informality as pent-up potential, nor of the McKinsey view that informality is a parasitic organizational form that hinders economic growth. The dual view of informality seems most consistent with the data.

Five Facts about Informality

The Informal Economy is Huge

Measuring the informal economy is inherently difficult. Much of informality is in farming, including both subsistence agriculture and informal sales of marketable crops. A large part, at least in terms of employment, also comes from self-employed sellers and peddlers living at near-subsistence levels (Banerjee and Duflo 2011). Yet even if we look at the more-substantial businesses that employ workers, such as repair shops, furniture or metal-working factories, or transport firms, many forms of informality are apparent. At one extreme, some firms literally do not exist in the

Table 1
Size of the Informal Economy by Alternative Measures

Country		*Measure of informality*				
Income quartile	*GDP/ population*	*% GDP informal (World Economic Forum)*	*% Tax evasion (Enterprise Survey)*	*% Self-employment*	*% GDP informal (electricity consumption)*	*Registered firms/ population (1,000s)*
Bottom	429	35.4	29.0	46.4	38.9	3.2
Second	1,362	33.7	23.3	35.7	42.7	8.2
Third	4,002	27.6	19.7	23.1	31.3	28.7
Top	20,348	17.3	8.2	13.3	17.6	41.8
Sample mean	10,015	27.6	22.5	26.5	29.0	24.7
Difference 1st vs. 4th quartile	−19,919*	−18.1*	−20.8*	−33.1*	−21.4*	38.7*
Observations	185	125	95	133	57	83

Source: La Porta and Shleifer (2008).
Note: Table 1 presents various measures of the size of the informal sector, with 185 countries grouped by the quartile of per capita income.
* , **, and *** indicates significance at the 1, 5, and 10 percent levels, respectively.

eyes of the authorities: they do not register or comply with regulations, they make sales and pay for inputs including labor in cash, they do not have bank accounts, they do not pay taxes. At another extreme, as occurred in transition economies, registered firms hide some of their sales from authorities to reduce profit taxes but still hire formal employees and comply with many regulations (Johnson, Kauffman, and Shleifer 1997). And there is everything in between, such as firms that obtain operating permits but do not pay social security taxes.

With these ambiguities in mind, several methods have been used to assess the size of the informal economy: surveys of experts about their countries, such as those conducted by the Global Competitiveness Report; surveys of entrepreneurs about their own activities, like those conducted by the World Bank Enterprise Surveys; census counts of people reporting that they are self-employed, which is typically a good proxy for informality; and even measures inferred from aggregate electricity consumption (on the plausible assumption that informal firms must also use electricity).

Table 1 presents various measures of the size of the informal sector, with 185 countries grouped by the quartile of per capita income. Fortunately, the very different measures of informality all paint a consistent picture. Depending on the indicator, the informal sector accounts for 30–40 percent of total economic activity in the poorest countries, and a higher share of employment. This falls to something closer to 15 or 20 percent in the richest quartile countries. The last column of

Table 1 offers another perspective: the poorest countries average about three registered firms per 1,000 people; in the richest quartile countries, this number rises to 42 per 1,000 people. Especially in the poor countries, the informal sector is huge, accounting for a giant share of output and employment.

Informal Firms are Small, Unproductive, and Stagnant

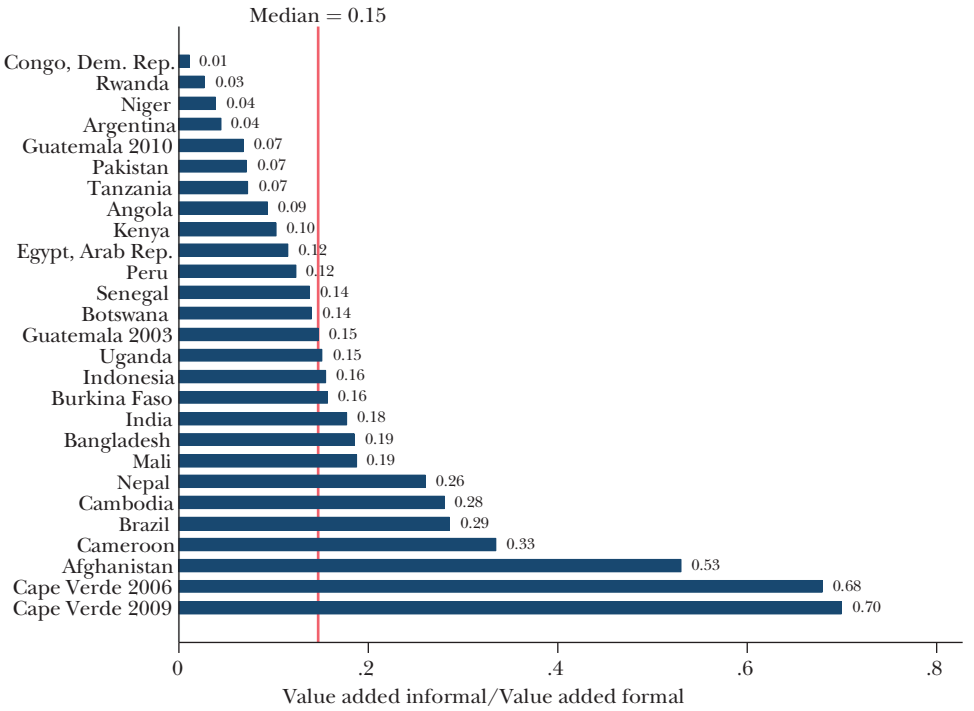
For two decades, the World Bank Enterprise Surveys collected data from entrepreneurs and managers in both formal and informal firms on their sales and inputs, employee and manager education, as well as a variety of assessments of the institutional environment. These data provide considerable evidence on the determinants of productivity of firms in developing countries, including their management. The World Bank surveys deal with actual businesses such as furniture producers or shoe factories; they do not cover the proverbial sellers of flowers and vegetables, who are also informal but have even lower productivity.

In La Porta and Shleifer (2008), we conduct an extensive analysis of size and productivity of formal and informal firms using data from poor countries where the World Bank surveyed both formal and informal businesses. Several findings stand out.[1] First, informal firms—even the real businesses surveyed by the World Bank—are much smaller than formal firms. An average formal firm employs 126 people, while an average informal firm employs only four. Informal firms are also much less productive, with productivity calculated as value added (sales net of expenditures on raw materials and energy) per employee. As Figure 1 shows, in the median sample country, informal firms add only 15 percent of the value per employee of formal firms. The ratio of value added by informal firms to that by formal firms ranges from 1 percent in Congo to 70 percent in Cape Verde. In La Porta and Shleifer (2008), we present some evidence indicating that these productivity differences reflect reality, not just underreporting of sales to interviewers by informal firms.

There are two other ways to see the extreme inefficiency of the informal sector. First, although productivity increases with size within the formal sector (as Hsieh and Olken discuss in their paper in this issue), Figure 2 shows a sharp productivity difference between informal firms and formal firms of the same size (in the median sample country, informal firms add 21 percent of the value per employee of formal firms). Inefficiency of the informal sector is not just a matter of small size. Second, in La Porta and Shleifer (2008), we also find that, averaging across countries, wages in informal firms are roughly one-half of those in small formal firms and less than one-third of those in large formal firms, another indication of low productivity. Many informal entrepreneurs would gladly close their businesses to work as employees in the formal sector if offered the chance, even if wages in the formal sector are taxed while income in the informal sector is not. Few of them have this opportunity.

[1] The online Appendix available with this paper at http://e-jep.org presents a summary table of these findings.

Figure 1
Ratio of the Value Added by Informal Firms to Value Added by Formal Firms

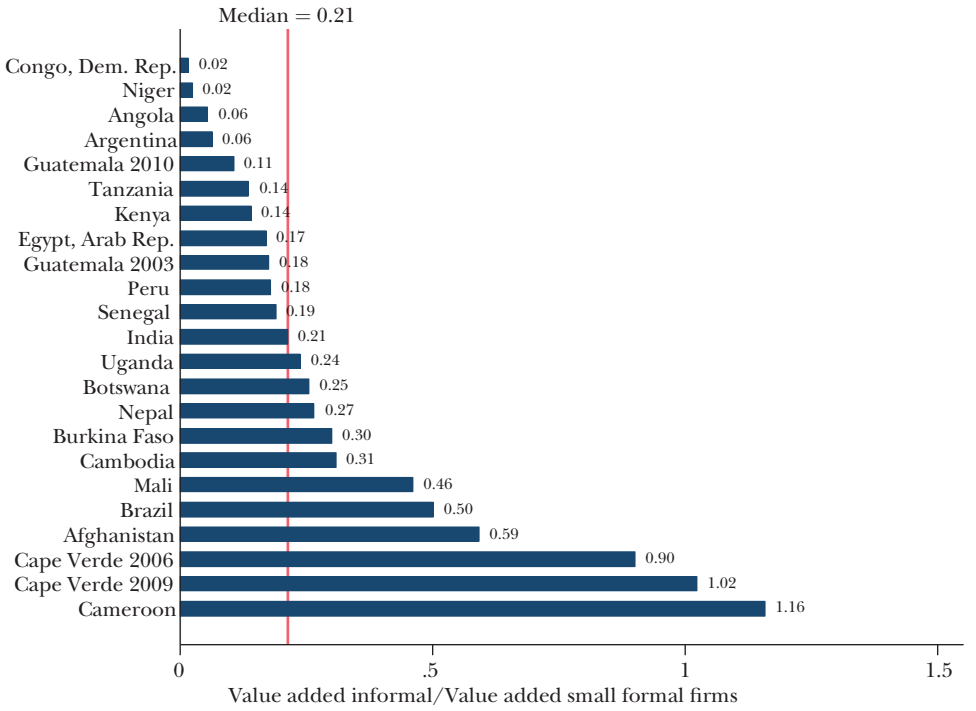


Source: World Bank Enterprise Survey. For details see La Porta and Shleifer (2008), which provides an extensive analysis of the productivity of formal and informal firms using Enterprise Survey data.
Note: In the labels on the left, dates are given for countries that were surveyed twice, and omitted for countries that were surveyed once.

The low value added per employee in the informal sector reflects the extremely low quality of products produced by informal firms. Although quality is difficult to measure, our visits to furniture and metal-working factories in Kenya and Madagascar revealed extreme crudeness of the products being made, usually with fairly basic tools, even when the raw material (as in the case of furniture) was hardwood. Informal factories appear to sell extremely low-quality goods for low prices to low-income customers. Informal entrepreneurs in Africa fear formal competition; they repeatedly expressed their fear of competition from Chinese imports. They are far from threatening to formal firms.

In La Porta and Shleifer (2008), we explore the sources of productivity differences between formal and informal firms. One interesting finding is that differences in the human capital of workers are small, at least as measured by education. The data on formal and informal firms contains no direct measures of capital, although formal firms are much more likely to have their own electricity generators. One of the most striking differences between formal and informal firms is in the human

Figure 2
Ratio of the Value Added by Informal Firms to Value Added by Small Formal Firms

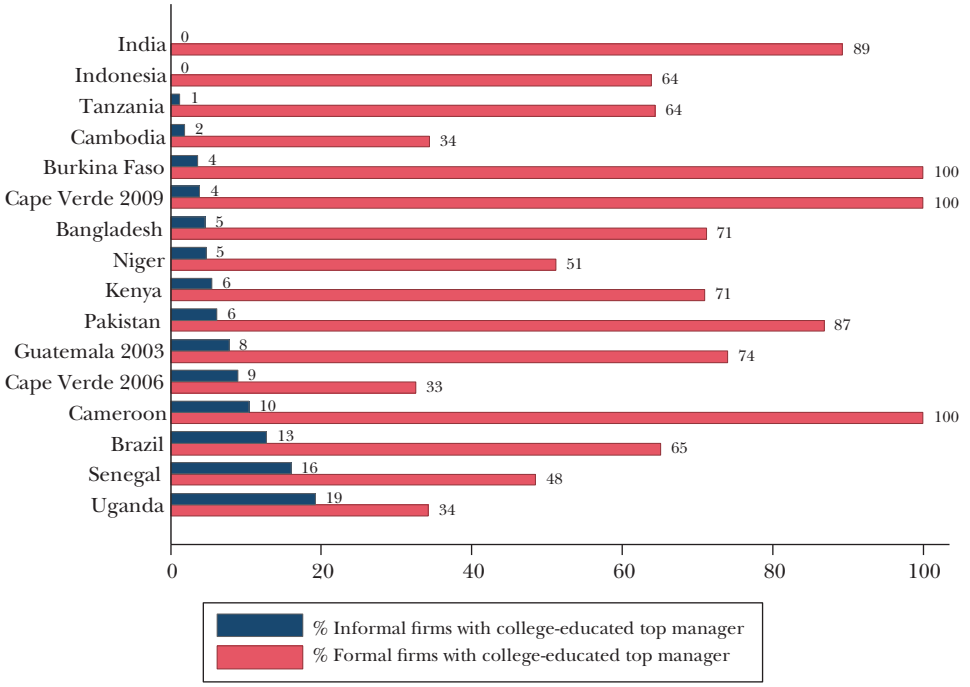


Source: World Bank Enterprise Survey. For details see La Porta and Shleifer (2008).
Notes: Figure 2 shows a sharp productivity difference between informal firms and small formal firms. In the labels on the left, dates are given for countries that were surveyed twice, and omitted for countries that were surveyed once.

capital of their managers. Figure 3 presents World Bank survey data on the fraction of informal and formal firms run by college-educated managers. Consistent with the dual view, only 7 percent of the managers of informal firms have a college degree, while this number is 76 percent for the formal firms. In production function estimates, managerial human capital emerges as a quantitatively large and statistically significant determinant of productivity.

In Gennaioli, La Porta, Lopez-de-Silanes, and Shleifer (2013), we report closely related findings for formal firms around the world. We document enormous productivity gaps between firms run by educated versus uneducated managers and entrepreneurs. Production function estimates imply nearly 30 percent returns per extra year of education of managers, even though estimated returns to an additional year of worker education are in the standard range of 6–7 percent. The message that emerges consistently from this work is that informal firms are hugely unproductive, and a principal reason is the low level of human capital of the people who run them.

Figure 3
Percent of Firms Run by a College-Educated Top Manager in the Informal and Formal Sectors



Source: World Bank Enterprise Survey. For details see La Porta and Shleifer (2008).
Notes: Figure 3 presents Enterprise Survey data on the fraction of informal and formal firms run by college-educated managers. In the labels on the left, dates are given for countries that were surveyed twice, and omitted for countries that were surveyed once.

The low productivity of informal firms is reflected in their growth rates as well. In La Porta and Shleifer (2008), we report sharply lower employment growth rates for informal than for formal firms (5 percent versus 10 percent per year). Indeed, an average informal firm in World Bank Enterprise Surveys had been around for nearly a decade and has continued its existence with only modest growth even during a period of rapid growth of formal firms. In a similar vein, de Mel, McKenzie, and Woodruff (2008) find that roughly 70 percent of own-account workers in Sri Lanka have backgrounds, abilities, and attitudes more similar to those of wage workers than those of owners of firms and that they rarely expand by adding paid employees (see also Ardagna and Lusardi 2010). These findings line up with the evidence from the US economy: most US small businesses have little desire to grow big or to innovate in any observable way (Hurst and Pugsley 2011).

Table 2
Obstacles to Doing Business

	Informal Enterprise Survey	*Formal Enterprise Survey*				*Formal vs. Informal*
		Small	*Medium*	*Big*	*All*	
Obstacles (% of firms identifying an obstacle as the most important)						
Access to financing	43.8%	20.6%	17.8%	13.6%	18.5%	−25.3%*
Political instability	11.4%	9.5%	9.1%	11.7%	9.7%	−1.7%
Access to land	11.2%	5.6%	4.2%	4.1%	5.0%	−6.3%**
Corruption	7.4%	7.3%	8.2%	6.0%	7.4%	0.0%
Electricity	7.3%	10.0%	9.8%	7.4%	9.8%	2.5%
Business licensing and permits	6.3%	2.3%	2.7%	1.7%	2.4%	−3.9%**
Crime	3.4%	5.2%	5.0%	7.2%	5.4%	2.0%
Legal system	3.3%	0.5%	0.5%	1.9%	0.8%	−2.5%*
Customs and trade regulations	2.1%	3.2%	4.4%	5.0%	3.8%	1.8%
Uneducated workforce	1.8%	4.6%	6.0%	10.4%	6.0%	4.2%***
Labor regulations	1.8%	2.6%	3.1%	4.8%	3.3%	1.4%
Tax administration	0.1%	4.3%	6.7%	6.4%	5.3%	5.2%**
Practices of competitors in the informal economy	0.1%	14.4%	13.4%	9.9%	12.9%	12.9%*
Tax rates	0.0%	7.7%	6.2%	6.3%	6.8%	6.8%*
Transportation	0.0%	2.2%	2.9%	3.7%	2.8%	2.8%*

Source: World Bank Enterprise Surveys. See online Appendix for countries and years.
Note: Table 2 compares perceived obstacles to doing business reported by informal and formal entrepreneurs.
* , **, and *** indicates significance at the 1, 5, and 10 percent levels, respectively.

Regulation is Not What Keeps Informal Firms Down

Why don't informal firms become formal? De Soto (1989) has famously argued that informal firms would like to become formal, but are held back by corruption and government regulation. World Bank Enterprise Surveys of informal entrepreneurs allow a direct assessment of this view. Table 2 compares perceived obstacles to doing business reported by informal and formal entrepreneurs. By far the greatest perceived obstacle by both types of firms is lack of access to finance, although informal firms perceive this as a much greater problem. The link between access to finance and registration may not be causal, however. For example, some of the informal firms we visited maintained several months of (extremely slow-selling) inventory without realizing that it is a form of capital. Their owners simultaneously complained that they did not have financing to buy tools. Similarly, banks may only lend to skilled entrepreneurs or want to see some form of control system (like accounting books) that informal entrepreneurs often lack. In these ways, lack of human capital might be at the heart of the perceived inaccessibility of finance.

Next to perceived financing problems, government regulations are distant concerns. Fewer than 10 percent of either formal or informal firms worry about each of the following categories: corruption; business licensing and permits; or the legal system. Lack of access to land is a bigger problem for informal firms, in part

because a large fraction of them occupy their premises illegally and fear eviction. It is difficult to read this evidence as pointing to the institutional environment as the central obstacle to doing business for informal firms.

World Bank surveys for a more limited group of ten countries, mainly in Africa, also offer more direct evidence on how respondents from the informal sector perceive the potential benefits from registering their firms.[2] Three-quarters of respondents from the informal sector in these surveys mention "better access to financing" as a gain, and one-quarter mention "better access to raw materials." In contrast, "better access to markets" and "fewer bribes" are only mentioned by 14 percent of respondents, and potential gains like "better opportunities with formal firms," "more access to government programs," "better legal foundations on the property," and "better access to infrastructure" are only mentioned by fewer than 10 percent of respondents. This evidence does not mean that the institutional environment that informal firms face is good—on the contrary, such firms face terrible problems of corruption, police abuse, and crime. In some countries, informal entrepreneurs report that up to 3 percent of their sales are stolen. Rather, the evidence suggests that informal firms do not see that formality will address these problems: they will face corrupt and abusive policemen, inspectors, and other officials anyhow. Meanwhile, informal firms report that "other firms like theirs" evade about 75 percent of taxes. Evading taxes is too attractive to be offset by the meager benefits of formality that the informal entrepreneurs would realize.

Informal Firms Rarely Become Formal

Informal firms almost never become formal. In La Porta and Shleifer (2008), we report that on average 91 percent of registered firms started out as registered. An average surveyed informal firm has been in business for nearly a decade without attempting to become formal. Also consistent with this observation, only 2 percent of informal firms sell their output to large firms (versus 14 percent of firms in the Enterprise Survey). Informal firms inhabit an economic space of their own, disconnected from the formal space.

In the last decade or so, processes for registering a business have been simplified in many countries around the world, and data on these changes has become available. This push began with De Soto's (1989) emphasis on the costs of registration, which in turn encouraged systematic data collection of entry costs around the world in Djankov, La Porta, Lopez-de-Silanes, and Shleifer (2002). This approach was in turn adopted by the World Bank in its Doing Business report, which since 2003 published a variety of measures of business regulation, including the regulation of entry, and country rankings. The Doing Business website reports 378 policy actions aimed at lowering the cost of registration in 160 countries. These policy reforms

[2] The ten countries are Angola (2010), Burkina Faso (2009), Botswana (2010), Cameroon (2009), Congo Democratic Republic (2010), Capo Verde (2009), Mauritius (2009), Mali (2010), and Nepal (2010). For country-by-country survey responses, see the online Appendix available with this article at http://e-jep.org.

have generated a wealth of data on the effect of registration costs on the decision to register and—to a lesser degree—on the impact of formality on productivity.

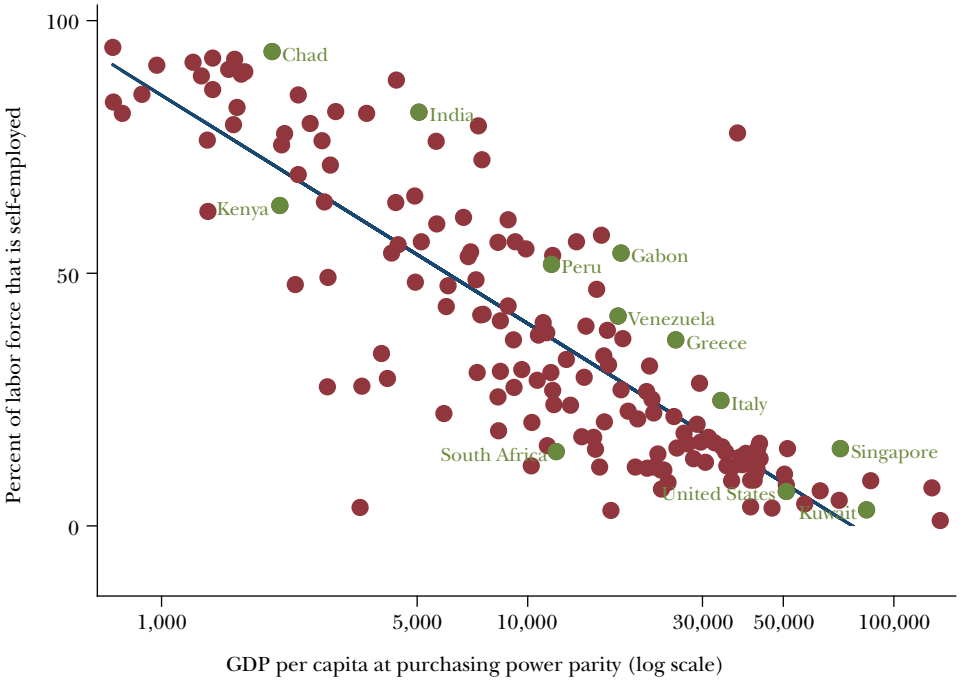
The most compelling evidence comes from two field experiments. The first was carried out in Belo Horizonte, a city in the Brazilian state of Minas Gerais, to test which government actions induce informal firms to register.[3] Firms were randomly assigned to a control group or one of four treatment groups: the first received information about how to formalize; the second received this information and free registration costs along with the use of an accountant for a year; the third group was assigned to receive an enforcement visit from a municipal inspector; while the fourth group was assigned to have a neighboring firm receive an enforcement visit to see if enforcement has spillovers. De Andrade, Henrique, Bruhn, and McKenzie (2013) report that the likelihood of registering increases by 21 to 27 percentage points if the firm receives an actual inspection, but it is unaffected by the other three interventions. Apparently, most informal firms do not formalize unless forced to do so.

The second field experiment was carried out in Sri Lanka. De Mel, McKenzie, and Woodruff (2013) report that information about the registration process and even actual reimbursement of direct costs of registration had no effect on formality. In contrast, around one-fifth of eligible firms registered when offered payments equivalent to one-half to one month of the median firm's profits, and one-half of eligible firms registered when offered payments equivalent to two months of the median firm's profits. Firms were visited 15, 22, and 31 months after the intervention. Firms that formalized had higher profits, but this effect was largely due to a few firms that experienced substantial growth. Jaramillo (2009) reports similarly small effects of easier registration from an experiment in Lima, Peru, the city whose informal sector De Soto has celebrated.

The evidence from changes in registration costs is one lens on informality. Another lens is the evidence on the effects of microcredit, which shows that such credit helps informal entrepreneurs a bit, but almost never jump-starts significant growth or transforms them into formal businesses (Karlan and Zinman 2011). Still another lens comes from the emerging image of slums as domains of permanent informality rather than hubs of transition between agriculture and the formal sector (Marx, Stoker, and Suri 2013). These studies suggest that informal firms start out and live out their lives informal, they avoid taxes and regulations, and they do not trade with the formal sector. It is difficult to lure them into becoming formal, even with subsidies. Far from being reservoirs of entrepreneurial energy, they are swamps of backwardness. They allow their owners and employees to survive, but not much more.

[3] The background of the field experiment is of independent interest. The process of simplifying the process of business registration started in 1996 with the SIMPLES program, which consolidated multiple tax payments and contributions into a single payment, lowering the tax burden on small firms. It was followed by the Minas Facil program in the state of Minas Gerais in 2005 to reduce number of procedures and time to start a business. Despite these efforts, survey data from 2009 reveal that 72 percent of firms in Minas Gerais remained informal.

Figure 4
Self-Employment and GDP per Capita in 2013



Source: World Development Indicators. See online Appendix for countries.

As Countries Develop, Informality Becomes Less Important

Important as the informal economy is in low-income countries, it becomes much less significant in high-income ones. Table 1 shows that the estimates of the size of the informal economy decline with per capita income. Figure 4 illustrates this point more clearly by showing a strong negative correlation between per capita income and the share of economic activity that is informal as measured by the share of self-employment in total employment. Very similar results obtain with the other indicators from Table 1. As an economy develops, informality shrinks.

The Dual View of Informality

We have focused on perhaps the most basic facts about the informal economy. It is extremely large in the poorest countries, but it eventually shrinks as countries develop. It exhibits very low productivity. Informal firms rarely make a transition to formality, even when encouraged or subsidized to do so; rather they carry on without much growth for long periods of time. They are run by uneducated entrepreneurs. Government policies are definitely a hindrance for informal firms, but they are a

hindrance to formal firms as well. Bad government is not the main competitive problem of informal firms: their main problem is that they add so little value.

The evidence we have presented does not support De Soto's (1989, 2000) romantic view of informal firms as reservoirs of productive entrepreneurial energy; most of these firms are too inefficient to survive in the formal sector and do not join it even when barriers to entry are eliminated. Nor is the evidence consistent with McKinsey Global Institute's emphasis on the dangers of informal firms (Farrell 2004); they mostly do not appear to pose much of a competitive threat. The evidence appears most consistent with Lewis's (1954) dual view of informality, which sees the formal and informal economies as largely segregated, producing different products with different labor, capital, and entrepreneurial inputs, and serving different customers.

What we call the dual view encompasses several ideas and formal theories. They all shed light on the separation of the formal and informal economies, and on the slow movement of activity into the formal sector. To put the issue most directly, if regulation is not to blame, why don't we see more informal entrepreneurs restarting their businesses formally? For instance, why don't informal furniture producers buy capital and start modern factories? Why don't street peddlers open modern stores? Alternatively, if formal firms are so efficient, why don't they capture the whole market straight away and drive out the informal firms? In all these cases, the transition to formality should be rapid.

Of course, the most obvious answer to all these questions is that formal firms have to pay taxes and comply with regulations, so they have a huge cost disadvantage relative to the informal firms. Joining the formal sector would raise the costs of informal firms significantly. The wedge between formal and informal labor costs is the major ingredient of all theories of dualism. But taxes and regulations are only part of the story. Other economic forces—on both the demand and the supply side—keep the two sectors separate.

The first force goes back to the original theories of dualism and focuses on demand as a constraint on transition to modern production technologies. Modernizing entrepreneurs need to generate sufficient sales to cover the fixed costs of investment. When the economy is poor, the demand for modern products may not suffice to cover these fixed costs. The problem is particularly severe in economies with significant levels of poverty and inequality, where the vast majority of the population buys almost no modern manufactured goods. In such economies, the informal sector delivers low-quality goods cheaply to people who are themselves informal workers and who cannot afford the output of the higher-quality but more-expensive formal sector. In contrast, the formal sector remains small and offers high-quality goods to a minority of formal workers. These ideas about demand constraints gave rise to the so-called Big Push theory, in which the simultaneous modernization of multiple sectors of the economy generates sufficient demand for the products of the modern sector from its own employees to actually make the transition to formality profitable (for example, Murphy, Shleifer, and Vishny 1989a, b). Formal workers making quality shoes in formal leather factories buy quality chairs made by formal workers employed by formal furniture makers, and vice versa.

These demand-based theories of dualism make an important prediction. Specifically, they suggest that population growth may slow down the decline of the informal sector. To the extent that the formal sector is capital intensive and the informal sector is labor intensive, population growth, particularly if concentrated among the poor, would keep up both the demand for informal goods and the supply of informal workers. Even if labor flows toward the formal sector, with high population growth it would not flow fast enough to kill informality. We return to this prediction in the empirical analysis below.

In our work we have emphasized another input that might explain a slow transition to formality—namely entrepreneurial and management skills. As we have already discussed, the evidence in World Bank surveys, as well as in other data, shows that managerial inputs are extremely important for productivity and that the managers of informal firms are considerably less educated than the managers of formal firms (La Porta and Shleifer 2008; Gennaioli, La Porta, Lopez-de-Silanes, and Shleifer 2013). The evidence demolishes the idea that managers of informal firms can just start larger formal firms and operate them well enough to survive in the formal sector: they would not know how. A shortage of educated entrepreneurs might be the most important constraint on transition to formality, much more important than lack of demand.

Although the dual view does not see government policies as fundamental to shaping the size of the informal sector, such policies may play a contributing role to separating the formal and informal economies. As argued by Johnson, Kauffman, and Shleifer (1997), tax avoidance by informal firms undermines government tax collection and therefore the provision of public goods, which makes joining the formal sector to access the public goods less attractive. Levy's (2008) concern with the fiscal costs of informality and distortions arising from differential tax treatment of formal and informal sectors also point to the high costs of formality relative to benefits, which keep the informal sector large.

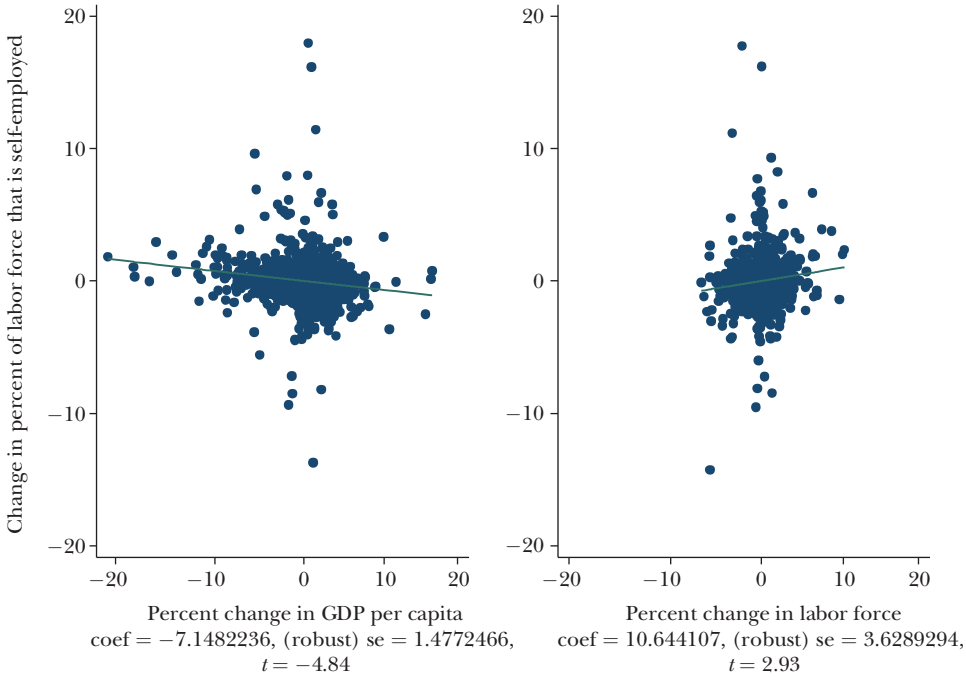
The dual view explains how the informal economy shrinks as the formal economy grows. Indeed, the decline of informality is the result of replacement of inefficient informal firms by efficient formal ones in the process of economic development. The available evidence both from the cross-section and from country examples strongly supports this perspective on the decline of informality and indeed suggests that both demand and supply factors play a role in this process.

Figure 5 presents graphically the results of the following regression run at annual frequency (with country fixed effects "δ") for a panel of 68 countries during the period 1990–2012:

$$\Delta(\textit{self-employment}) = \alpha + \beta \ln(\textit{GDP pc}_t / \textit{GDP pc}_{t-1}) + \gamma \ln(\textit{labor force}_t / \textit{labor force}_{t-1}) + \delta_i + \varepsilon.$$

That is, we regress the change in the percent of labor force in self-employment, a reliable and widely available measure of informality, on change in log GDP per capita and change in log labor force. We run this regression in first differences to alleviate

Figure 5
Partial Correlation of Change in Self-Employment, Growth in GDP per Capita, and Growth in the Labor Force



Source: World Development Indicators.
Notes: Figure 5 presents graphically the results of the following regression run at annual frequency (with country fixed effects "δ") for a panel of 68 countries during the period 1990–2012:

$$\Delta(\textit{self-employment}) = \alpha + \beta \ln(\textit{GDP pc}_t / \textit{GDP pc}_{t-1}) + \gamma \ln(\textit{labor force}_t / \textit{labor force}_{t-1}) + \delta_i + \varepsilon.$$

the concern that the strong negative correlation between self-employment and GDP per capita illustrated by Figure 4 is driven by omitted variables.

The left panel of Figure 5 shows that faster economic growth is associated with a more rapid decline in self-employment. Doubling GDP per capita is associated with a reduction in self-employment of 4.95 percentage points (the mean of self-employment is 26 percent, and its standard deviation is 16 percent). This estimate says that a low-income country that starts with 50 percent self-employment and then grows consistently at 7 percent per year so that per capita income doubles every 10 years will see its self-employment fall to the high-income countries' level of 20 percent after 60 years—a remarkably slow transition to formality. The right panel of Figure 5 shows that faster labor force growth is associated with a slower decline in self-employment, consistent with theoretical prediction. Doubling the labor force is associated with an increase in self-employment of 7.38 percentage points. This means that self-employment in the low-income country of the previous example

would converge to the high-income countries' level in 105 years if its labor force grows consistently at a 2 percent rate while GDP per capita grows at 7 percent. In sum, the transition to formality is driven by economic growth, but it is very slow, and even slower when labor force growth is fast.

A comparison of three rapidly growing countries—Korea, Chile, and Peru—illustrates the significance of labor force growth. For Korea, per capita income rose 2.5-fold between 1990 and 2012. During this period, the share of labor force classified as employees rose from 60 to 72 percent, and the share of self-employed fell from 40 to 28 percent. Korea's labor force during this period grew by only 34 percent in total, so the growth of formal employment share comes largely from an increase in formal employees and an actual decline in the number of self-employed. During the same period, GDP per capita rose 2.3-fold in Chile, but its labor force grew almost 70 percent from 5.0 million to 8.5 million. We do not see the same kind of reallocation of labor between formal and informal sectors. Indeed, the formal share of employment in Chile has increased only slightly even though, during this period, formal employment increased from fewer than 4 million to nearly 6 million persons.

These conflicting forces are illustrated most dramatically by the case of Peru, the birthplace of De Soto's (1989) enthusiasm for the potential of the informal sector. During the period 1990–2012, Peruvian per capita GDP has grown 2.1-fold, nearly as fast as Korea's economy, thanks to aggressive liberalization policies and the defeat of Maoist guerillas. Yet the Peruvian labor force grew almost as fast as its per capita income, and much of the labor absorption was done by the informal sector. As a consequence, despite the tremendous economic growth over this period, the share of formal employment in Peru has declined and that of informal employment has increased even though the level of total formal employment rose from 5 to 8 million people. Figure 6 uses Peruvian household survey data to graph annual GDP per capita, annual income of formal employees, and annual income of the self-employed for the period of 1990–2010. Figure 6 shows rapid economic growth during this period but also rapid increases in formal sector wages. In a striking contrast, incomes of the self-employed did not rise over this period. Economic growth was driven by the formal sector; the informal sector stagnated.

What about the supply side? Does scarcity of human capital indeed slow down the transition to a formal economy? Table 3 presents some evidence on this hypothesis. It uses subnational data for 1,090 regions in 71 countries to examine firm formation and employment composition within countries as a function of each region's level of education. Table 3 shows that, within countries, the more-educated regions have more formal establishments per capita, more formal employees both relative to population and relative to the number of establishments (that is, there are larger firms), and a larger fraction of formal employees employed in large firms. Looking across regions within a country (so at least national institutions are held constant), transition to a formal economy appears to be driven by human capital, consistent with the supply-side theories of dualism we discussed.

The bottom line of this evidence seems straightforward. Economic growth comes from the formal sector, which absorbs labor in part from the informal sector

Figure 6
Income of Employees, Self-Employed, and GDP per Capita in Peru

Source: Encuesta Nacional de Hogares (Peru).
Note: Figure 6 uses Peruvian household survey data to graph annual GDP per capita, annual income of formal employees, and annual income of the self-employed for the period of 1990–2010.

but mostly from the new generations of workers. As economies grow, productivity and income in the informal sector stagnate. Labor force growth slows down the absorption of labor in the formal sector, but eventually this process does take place. The supply of human capital, in contrast, speeds up the creation of formal firms. Some survey evidence on Peruvian slums collected and provided to us by Nathan Nadramija shows that it is mostly the children of the informal sector workers, rather than these workers themselves, who become educated and join the modern economy. As they do, the share of the informal economy declines because the unproductive informal firms cannot survive in the modern economy.

Conclusion

The evidence we have presented is broadly consistent with the dual view of informality: informal firms stay permanently informal, they hire informal workers for cash, buy their inputs for cash, and sell their products for cash, they are extremely unproductive, and they are unlikely to benefit much from becoming formal. This approach generates the strong prediction that the cure for informality is economic growth. The evidence strongly supports this prediction: informality declines, although slowly, with development.

This approach suggests that structural policies designed to promote formality should be introduced with caution. Their wisdom depends, in part on whether they encourage formalization, or discourage informal activity. Thus the simplification of business registration advocated by De Soto (1989) is probably a good idea, even though the evidence suggests that it is unlikely to have large benefits. On the other hand, we are skeptical of all policies that might tax or regulate informal firms. Rather

Table 3

Regional Human Capital, the Size of Establishments, and Participation in the Economy

	Dependent variable			
	ln(Establishments/ Population)	*ln(Employees/ Population)*	*ln(Employees/ Establishments)*	*ln(Employees Big Firms/Employees)*
Years of education in the region	0.2936* (0.0311)	0.3372* (0.0263)	0.1205* (0.0220)	0.2215* (0.0349)
Constant	−5.8407* (0.2554)	−4.3678* (0.2052)	0.8918* (0.2064)	−3.4071* (0.4041)
Observations	970	1,016	1,035	510
Adjusted R^2	92%	93%	82%	95%
Country fixed effects	Yes	Yes	Yes	Yes

Source: See online Appendix.

Notes: Table 3 uses subnational data for 1,090 regions in 71 countries to examine firm formation and employment composition within countries as a function of each region's level of education. The table reports country fixed effect regressions for the following four dependent variables: 1) logarithm of the number of formal establishments per capita; 2) logarithm of the number of formal employees per capita; 3) logarithm of the number of formal employees per formal establishment; and 4) logarithm of the number of formal employees working in firms that employ at least 100 employees as a percent of total employment. All regressions include the number of years of education in the region.

*, **, *** indicate significance at the 1, 5, and 10 percent levels, respectively.

than encourage informal firms to become formal, such policies may have the effect of driving them out of business, leading to poverty and destitution of informal workers and entrepreneurs. The recognition of the fundamental fact that informal firms are extremely inefficient recommends extreme caution with policies that impose on them any kind of additional costs.

There is accumulating evidence that growth that kills the informal sector is driven by the formation and expansion of formal firms managed by educated entrepreneurs. Uneducated entrepreneurs—in both informal and formal sectors—generally run small and inefficient firms; educated entrepreneurs and managers run larger and more-efficient firms. This is the dark side of dualism: informal economies are so large in poor countries because their entrepreneurs are so unproductive.

The evidence suggests that an important bottleneck to economic growth is not the supply of better-educated workers; indeed, at least on many observable characteristics the workers are rather similar in informal and formal firms. Rather, the bottleneck is the supply of educated entrepreneurs—people who can run productive businesses. These entrepreneurs create and expand modern businesses with which informal firms, despite all their benefits of avoiding taxes and regulations, simply cannot compete. This is how the informal economy dies out in the process of

development. From this perspective, the policy message for how to grow the formal economy and shrink the informal one is to increase—whether through immigration or education and training—the supply of educated entrepreneurs.

References

Ardagna, Silvia, and Annamaria Lusardi. 2010. "Explaining International Differences in Entrepreneurship: The Role of Individual Characteristics and Regulatory Constraints." In *International Differences in Entrepreneurship*, edited by Josh Lerner and Antoinette Schoar. University of Chicago Press.

Banerjee, Abhijit V., and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* New York: Public Affairs, Perseus Books.

de Andrade, Gustavo Henrique, Miriam Bruhn, and David McKenzie. 2013. "A Helping Hand or the Long Arm of the Law? Experimental Evidence on What Governments Can Do to Formalize Firms." World Bank Policy Research Working Paper 6435.

de Mel, Suresh, David McKenzie, Christopher Woodruff. 2008. "Who Are the Microenterprise Owners? Evidence from Sri Lanka on Tokman v. De Soto." World Bank Policy Research Working Paper 4635.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2013. "The Demand for, and Consequences of Formalization among Informal Firms in Sri Lanka." *American Economic Journal: Applied Economics* 5(2): 122–50.

De Soto, Hernando. 1989. *The Other Path: The Invisible Revolution in the Third World.* New York: Harper and Row.

De Soto, Hernando. 2000. *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else.* New York: Basic Books.

Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2002. "The Regulation of Entry." *Quarterly Journal of Economics* 117(1): 1–37.

Farrell, Diana. 2004. "The Hidden Dangers of the Informal Economy." *McKinsey Quarterly* 3: 27–37.

Gennaioli, Nicola, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2013. "Human Capital and Regional Development." *Quarterly Journal of Economics* 128(1): 105–164.

Harris, John R., and Michael P. Todaro. 1970. "Migration, Unemployment, and Development: A Two-Sector Analysis." *American Economic Review* 60(1): 126–42.

Hurst, Erik, and Benjamin Wild Pugsley. 2011. "What Do Small Businesses Do?" *Brookings Papers on Economic Activity*, Fall: 73–118.

Jaramillo, Miguel. 2009. "Is There Demand for Formality among Informal Firms? Evidence from Microfirms in Downtown Lima." German Development Institute Discussion Paper 12/2009.

Johnson, Simon, Daniel Kaufmann, and Andrei Shleifer. 1997. "The Unofficial Economy in Transition." *Brookings Papers on Economic Activity*, Fall: 159–240.

Karlan, Dean, and Jonathan Zinman. 2011. "Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation." *Science* 332(6035): 1278–84.

La Porta, Rafael, and Andrei Shleifer. 2008. "The Unofficial Economy and Economic Development." *Brookings Papers on Economic Activity*, Fall: 275–352.

Levy, Santiago. 2008. *Good Intentions, Bad Outcomes: Social Policy, Informality, and Economic Growth in Mexico.* Brookings Institution Press.

Lewis, W. Arthur. 1954. "Economic Development with Unlimited Supplies of Labor." *Manchester School of Economic and Social Studies* 22(2): 139–91.

Marx, Benjamin, Thomas Stoker, and Tavneet Suri. 2013. "The Economics of Slums in the Developing World." *Journal of Economic Perspectives* 27(4): 187–210.

Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny. 1989a. "Industrialization and the Big Push." *Journal of Political Economy* 97(5): 1003–1026.

Murphy, Kevin M., Andrei Shleifer, and Robert Vishny. 1989b. "Income Distribution, Market Size, and Industrialization." *Quarterly Journal of Economics* 104(3): 537–64.

Rauch, James E. 1991. "Modeling the Informal Sector Formally." *Journal of Development Economics* 35(1): 33–47.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 127–148

Do Poverty Traps Exist? Assessing the Evidence[†]

Aart Kraay and David McKenzie

In 1960, per capita incomes in Burundi, Haiti, and Nicaragua were $347, $1,512, and $2,491, respectively. Despite the sevenfold difference in incomes between Nicaragua and Burundi, all three countries were poor by developed-country standards. Nicaragua's per capita income was just 16 percent of the level enjoyed by citizens of the United States at the time, while Burundi's income was a paltry 2 percent of US levels. Fifty years later in 2010, per capita incomes in these three countries were basically unchanged, at $396, $1,411, and $2,289 respectively (all measured in 2005 purchasing power parity-adjusted US dollars as reported in the Penn World Tables), reflecting negligible real growth in per capita income over this half-century.

Why did per capita incomes not increase in these countries? One possible explanation is a "poverty trap," which can be understood as a set of self-reinforcing mechanisms whereby countries start poor and remain poor: poverty begets poverty, so that current poverty is itself a direct cause of poverty in the future (Azariadis and Stachurski 2005).[1] The idea of a poverty trap has a striking implication for policy. It implies that much poverty is needless, in the sense that a different equilibrium is possible and also that one-time policy efforts that break the poverty trap may have lasting effects. For example, poverty trap analysis is often used as the basis for advocating a massive increase in foreign aid to low-income counties that would act as

[1] Others also refer to a single, poor, dynamic equilibrium as a structural poverty trap (for example, Barrett and Carter 2013, Naschold 2013), but we do not use this definition in our paper.

■ *Aart Kraay and David McKenzie are economists in the Development Research Group, World Bank, Washington, DC. Their email addresses are akraay@worldbank.org and dmckenzie@worldbank.org.*

[†] To access the data Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.3.127

doi=10.1257/jep.28.3.127

"a 'big push' to instigate a virtuous circle of higher rates of savings, investment and economic growth" (UNCTAD 2006, p. 3), as well as for industrial policies designed to overcome coordination failures.

The concept of a poverty trap at the level of national economies is related to, and sometimes based on, microeconomic foundations that argue for the existence of poverty traps at the household level. Again, the general idea is that current poverty might be what is causing future poverty. This kind of analysis has led to recommendations like that of the Millennium Villages project, which plans to spend $6,000 per household to implement a complex package of interventions designed to lift selected African households out of extreme poverty (Clemens 2012). It also helps provide justification for support for microfinance loans designed to allow households to lift themselves out of poverty by buying some fixed-cost asset necessary to operate a business.

Economic theory has long offered a number of different models that can give rise to such traps at both the macro and micro levels. An early example is given by Nelson (1956), who develops a growth model with low saving and investment rates at low income levels. Even almost 60 years ago, the idea of poverty traps was so well-established in the thinking of the profession that Nelson felt obliged to acknowledge in his introduction: "Although the notion of low-level stagnation is scarcely new or different, it is hoped that this paper does more than express the common knowledge of economists in a complicated manner."

But what does the modern evidence suggest on the extent to which poverty traps exist in practice and what underlying mechanisms they might involve? To address this question, we begin by outlining a simple model of a poverty trap at the country level that was a staple of many macro views of development in the 1950s and 1960s, and use it to clarify what is meant by a poverty trap. We next examine evidence on the persistence of poverty at the country and household level to show that the stagnant incomes of countries like Burundi, Haiti, and Nicaragua over long periods is rare in practice, with the typical poor country growing at least as fast as the global average over the last 60 years.

We then turn to examining the main macro and micro mechanisms that are hypothesized to give rise to poverty traps. We deliberately restrict the scope of our paper to focus on the potential for poverty traps to account for the post–World War II growth experience of developing countries and for the persistence of poverty within households over years rather than generations.[2] Our focus is on multiple-equilibria traps that have often been used to motivate foreign aid efforts, with the idea being that a technical fix exists to enable countries or households to move to a different equilibrium. A related strand of literature considers political economy reasons for why poor countries remain poor, stressing the possibility of self-reinforcing

[2] As a result, we do not take a view on the importance of poverty traps for understanding very long-run development over hundreds or even thousands of years as for example in the work of Galor and Weil (2000) emphasizing demographic transitions, or that of Acemoglu and Zilibotti (1997) who emphasize how fixed costs of production limited the ability of countries to adopt diversified portfolios of risky but high-return technologies in the pre–Industrial Revolution period.

low-quality institutions (for example, Acemoglu and Robinson 2012). We acknowledge the potential importance of these explanations, but note that they have rarely been used as a justification for aid policy, given the difficulties in aid influencing the dynamics of political institutions.

The main mechanisms we examine include S-shaped savings functions at the country level; "big-push" theories of development based on coordination failures; hunger-based traps which rely on physical work capacity rising nonlinearly with food intake at low levels; and occupational poverty traps whereby the combination of borrowing constraints and lumpy production technologies means that poor individuals who start businesses that are too small will be trapped earning subsistence returns. We conclude that these types of poverty traps are rare and largely limited to remote or otherwise disadvantaged areas. We discuss behavioral poverty traps as a recent area of research for which the evidence is just starting to accumulate, and geographic poverty traps as the most likely form of a trap. The policy prescriptions that result are then quite different from the calls for a big push in aid or expansion of microfinance to allow people to overcome credit constraints. Rather, they call for action in less-traditional policy areas such as promoting more migration.

A Simple Model of a Poverty Trap at a Country Level

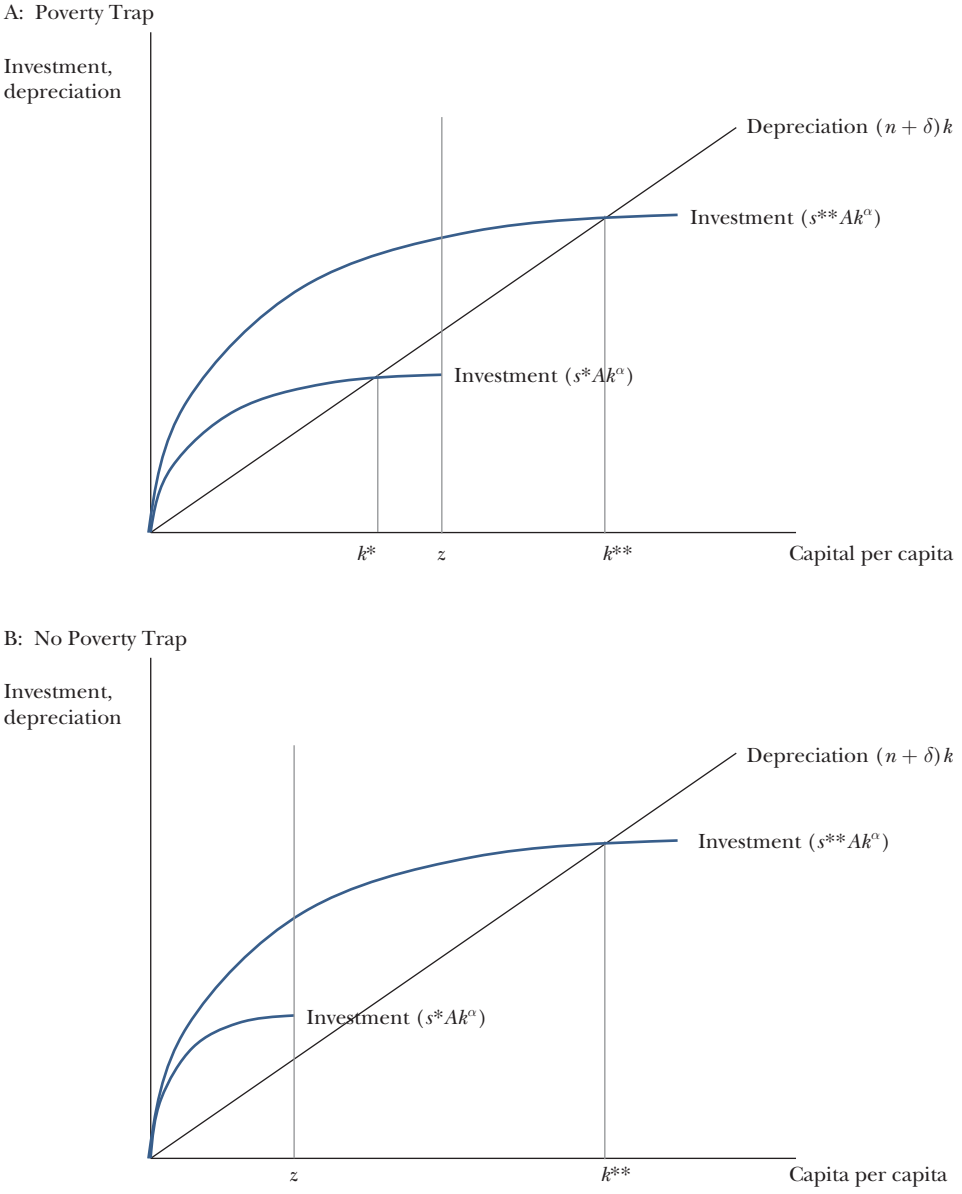
Economists have long turned to theories of poverty traps to account for the persistence of poverty over extended periods of time. One commonly invoked mechanism is that a country that is poor will remain poor because it will not be able to accumulate sufficient capital per capita for incomes to rise. This basic mechanism is captured in the standard diagram that adorns many textbook discussions of poverty traps, shown in the first panel of Figure 1.

Countries have access to an aggregate per capita production function $y = Ak^{\alpha}$ in which output (y) depends on capital per capita (k) and the level of technology (A). The economy is closed, and investment is financed by saving a fraction of income. At low levels of development, the saving rate s^* is low; the intuition often given for this assumption is it is difficult for those with low incomes to save much. At higher levels of development, the saving rate is also higher, at $s^{**} > s^*$.

The two curved lines show the amount of investment available when the saving rate is low and when the saving rate is high. At any given level of the capital stock and output, investment matches the savings: that is, investment is high when the saving rate is high, and investment is low when the saving rate is low. Because saving is a fixed fraction of output, the amount of investment increases with output and the capital stock but at a decreasing rate due to diminishing returns in the aggregate production function.

The rate of population growth is given by n and the rate of depreciation of capital stock by δ. If the amount of new saving and investment exactly offsets the decline in the per capita capital stock due to depreciation and population growth, then the economy will enter a steady-state equilibrium in which the level of capital and output remains constant over time. At low levels of development, saving and

Figure 1
Growth Models With and Without Poverty Traps



investment are low and there is a stable low-level equilibrium at k^* beyond which the country cannot grow. However, if a country manages to accumulate capital greater than some threshold z, the country then can move to a high saving rate $s^{**} > s^*$. This results in a high steady-state capital stock at $k^{**} > k^*$ and a higher level of per capita income. This framework encapsulates the basic idea that self-perpetuating

mechanisms, such as low saving at low levels of development, can keep countries trapped in poverty.

Moreover, this framework can also easily be used to illustrate a variety of other mechanisms for poverty traps. Any set of assumptions that generates a positive relationship between the level of development and growth fundamentals, such as higher productivity levels (A), higher saving rates (s), or lower population growth rates and/or depreciation ($n + \delta$), can in principle generate a diagram similar to Figure 1, with multiple steady states and a low-level poverty trap. For example, the two curves of Figure 1 could correspond to A^* and A^{**} rather than to s^* and s^{**}, then the figure represents a poverty trap story in which countries remain poor because they do not have access to high-productivity technologies. The menu of options grows further if we interpret A broadly, for example to include institutions. Similarly, one could interpret k broadly to include human capital, opening the door to models of poverty traps based on underinvestment in human as opposed to physical capital.

The appeal of such simple models of poverty traps is difficult to overstate. At a conceptual level, it seems eminently plausible that many low-income countries may lack good growth fundamentals supporting high saving rates and productivity gains. From a policy perspective, poverty trap models are a call to action, laying out a clear case for interventions to spring countries from such poverty traps. For example, if investment rates and capital accumulation are low because countries are poor and cannot afford to save, foreign aid can help to finance investment until countries develop to a point to the right of the threshold z where their saving rates climb high enough to place them on a trajectory to the high steady state at k^{**}.

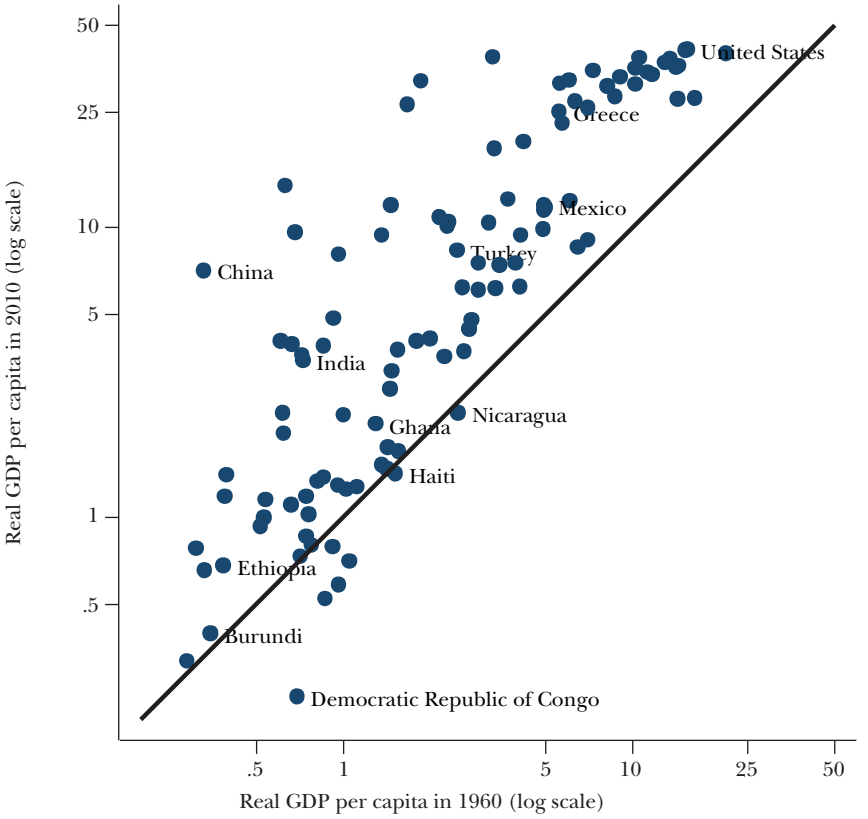
The key ingredient here is that some fundamental—such as the saving rate—is itself affected by how poor the country is. However, this is a necessary, but not sufficient, condition for a poverty trap to occur. For example, the second panel of Figure 1 is identical to the first except that the threshold z at which saving rates increase from s^* to s^{**} is shifted slightly to the left. Although saving rates rise with the level of development, no poverty trap occurs. A country that starts out poor gradually accumulates capital until it reaches the threshold value of z. At this point, its saving rates increase, and over time it reaches the new high steady state at k^{**}.

Finally, note that poverty can be persistent without there being a poverty trap in the sense the term is used here. For example, in the second panel of Figure 1, the unique steady state at k^{**} may still be at a very low level of income, due to low productivity A. In such a case, there is no scope for a short-term policy action to jump-start the economy out of poverty unless that policy is able to change the fundamentals of the economy, such as by changing the level of technology A. Simply investing in more capital will not raise the long-run steady-state level of income.

Evidence on the Persistence of Poverty

Models of poverty traps at the country level imply that countries stuck in traps will suffer from stagnant income levels over long periods of time. Similarly, poverty traps imply income dynamics at the individual or household level that are very

Figure 2
Absolute Income Stagnation is Rare



Source: Penn World Tables, Version 7.1.
Note: Real GDP per capita is in thousands of 2005 US dollars adjusted for differences in purchasing power.

different for those who are initially rich and those who are initially poor. In this section, we discuss the reduced-form evidence for the persistence of poverty, at both the "macro" and "micro" levels. This distinction is relevant, because it is in principle possible to see persistently low incomes at the country level but considerable income mobility at the individual level. Conversely, rising incomes at the country level can coincide with pockets of stagnation at the individual level within the country.

At the aggregate level, the cross-country evidence over the last half-century or so on the persistence of poverty across countries—the core pattern which many models of poverty traps aim to explain—is not particularly compelling. The truly stagnant income levels predicted by standard models of poverty traps are in fact quite rare in the cross-country aggregate data. Figure 2 puts the experience of Haiti, Nicaragua, and Burundi noted in the introduction into a broader cross-country context. It plots the log level of real GDP per capita in 2010 (on the vertical axis) against the log level of real GDP per capita in 1960 (on the horizontal axis) for a set

Table 1
Real GDP per Capita Growth 1960–2010, by Quintiles of Initial Income Distribution

	Mean real GDP per capita growth 1960–2010 (percent per year)	*Standard deviation*	*Number of countries*
Poorest quintile	2.2	2.0	22
Second quintile	0.9	1.5	22
Third quintile	2.0	1.6	22
Fourth quintile	2.4	1.2	22
Richest quintile	2.1	0.6	22

Source: Penn World Tables Version 7.0.
Note: Table 1 reports summary statistics on average growth performance from 1960–2010 for 110 countries, grouped by quintiles of the initial income distribution in 1960, updating calculations by Easterly (2006).

of 110 countries for which data is available. While it is true that a few initially poor countries cluster along the 45-degree line, and even a few such as the Democratic Republic of Congo have experienced negative growth over this period, the vast majority of countries at all initial income levels have experienced positive growth over the past 50 years, as reflected in their position well above the 45-degree line.

The observation that low-income countries as a group do not seem to show a greater propensity to no growth or even substantially slower growth can be seen more clearly in Table 1, which reports summary statistics on average growth performance from 1960–2010 for 110 countries, grouped by quintiles of the initial income distribution in 1960, updating calculations by Easterly (2006). Across all countries, average annual growth in real per capita GDP was 1.9 percent per year. The key point for the present discussion is that significantly positive growth performance can be found across all quintiles of the initial income distribution, including the poorest. Indeed, among the poorest 20 percent of countries, average per capita income growth was 2.2 percent per year, slightly higher than the global average, and also slightly higher than the average of the initially richest 20 percent of countries. Even if we further disaggregate the poorest quintile of countries, average growth among those in the poorest decile was 1.8 percent per year, while those in the second decile grew at 2.6 percent per year.

The fact that even the initially poorest 10 percent of countries has grown at a rate similar to the historical growth rate of the United States over the past 50 years—and indeed, the US growth rate over the past 200 years—is difficult to square with models of poverty traps, for two main reasons. First, poverty trap models imply stagnant incomes at low levels of development, while the post-World War II growth experience suggests that stagnant incomes are both rare and not systematically associated with initial levels of development.

Second, poverty trap models typically feature some sort of threshold that, once crossed, leads to an exit from the poverty trap and an acceleration of growth.

The fact that even the poorest countries have on average shown positive growth means that at least some of them should have crossed the relevant threshold for the poverty trap that might otherwise have constrained their development. As an illustration, per capita GDP of the median country in the poorest quintile in 1960 was about $550; suppose we take as the relevant threshold log per capita income of the richest country in the second quintile, which was about $1,500. At a growth rate of 2.2 percent per year, it would take just 46 years for the median country in the poorest quintile to reach this hypothetical threshold. More generally, unless the thresholds below which countries are trapped are very high indeed, most countries should simply grow their way out of any poverty trap.[3]

A further piece of evidence comes from looking directly at accelerations in growth rates, which poverty trap models predict should occur as countries cross the relevant threshold of per capita income. Hausmann, Pritchett, and Rodrik (2005) study episodes of substantial accelerations in real GDP growth rates. While their emphasis is not on poverty traps in particular, they document that there is no stable relationship over time between countries' income levels and the likelihood of a growth acceleration. On the other hand, models of poverty traps suggest that poor countries emerging from traps should see growth accelerations.

A number of studies have also searched for evidence of poverty traps by examining whether there is an S-shaped relationship in income or asset dynamics at the individual or household level. The typical approach involves estimating a nonlinear model relating income or assets today to those in the previous period, and examining the estimated dynamic relationship. For example, Jalan and Ravallion (2004) find no evidence for this kind of poverty trap at the household level using six years of income in China, and Lokshin and Ravallion (2004) find no evidence of poverty traps over four-to-six year intervals in Hungary and Russia. Naschold (2013) likewise finds no evidence of multiple-equilibria poverty traps over three-to-five year periods using panel data from rural Pakistan and Ethiopia. In contrast, Barrett et al. (2006) do find evidence of multiple equilibria in asset dynamics in some of their more remote sites in small samples from rural Kenya and Madagascar, as do Adato, Carter, and May (2006) with asset dynamics over a five-year period in South Africa.

However, the persuasiveness of this household-level evidence is mixed, whatever conclusion is reached. Most developing countries lack long-term longitudinal panels, so tracking income and assets over time is difficult and prone to measurement error. One potential solution has been to track cohorts rather than individuals over time, with Antman and McKenzie (2007) finding no evidence for a poverty trap using 58 quarters of repeated cross-sections from Mexico. Administrative data might allow a researcher to track the incomes of individuals and their children over long periods of time with less measurement error, as has been possible in several developed

[3] While most of the models of poverty traps we are aware of feature fixed absolute thresholds, it is possible to construct arguments for poverty traps based on relative thresholds. For example, consider poverty traps based on subsistence consumption (which we discuss in the following section). If the level of consumption required for subsistence rises with average incomes, then it may take much longer to grow out of a trap. One suggestive piece of evidence for this possibility is the fact that, across countries, national poverty lines tend to increase with the level of development (Chen and Ravallion 2011).

country studies (for example, Lucas and Pekkala Kerr 2013), but most of the poor in developing countries work outside the formal system, which means that they appear only sporadically in official administrative records, if at all. At least until improved data becomes available, a more promising approach is to investigate the evidence for specific mechanisms through which poverty traps are theorized to occur.

Inspecting the Mechanisms

The overall evidence suggests that poverty traps are not widespread at either the country or individual level, with most countries experiencing some growth and poor individuals not appearing to have dramatically different income dynamics from those who earn more. However, certain types of nonlinearities in these dynamics may appear in situations of crisis or famine, when markets and institutions break down (Ravallion 1997). Barrett and Carter (2013) offer a number of other reasons why direct testing of income or asset dynamics may struggle to find poverty traps even in cases where they exist. For example, poverty traps might apply for certain country circumstances or particular individuals but not others, and failing to account for this heterogeneity by averaging across these groups could mislead. Therefore, we examine the evidence behind some of the main theoretical mechanisms thought to give rise to the types of poverty traps that motivate many arguments for foreign aid. We begin with perhaps the two most well-known macro mechanisms, savings and coordination failure, then discuss two of the most well-known micro mechanisms, nutrition and lumpy investments, before turning to behavioral and geographic poverty traps.

Saving-Based Poverty Traps

Saving-based poverty traps were among the first to be developed into formally articulated models. The basic idea is straightforward: if countries (or individuals) are too poor to save, they cannot accumulate capital, and thus their incomes can only grow at the rate of total factor productivity growth. If this productivity growth is low or zero, then incomes will be stagnant. While the theoretical argument has been understood for a long time, empirical evidence concerning this mechanism is much more recent.

Kraay and Raddatz (2007) calibrate two standard growth models to match key features of sub-Saharan African countries in order to study saving-based poverty traps. The first is a Solow-style economy with an exogenously given saving rate that rises with income. A necessary condition for the existence of a poverty trap is that saving rates are an S-shaped function of the level of development, starting out flat when countries are poor, increasing sharply over some intermediate range, and then leveling off again. Kraay and Raddatz estimate a nonlinear empirical relationship between saving rates and incomes observed in the cross-country data and use this to calibrate the exogenous saving function in the Solow economy. While saving rates do increase with the level of development, it turns out that they increase sharply at quite low levels of development such that a stable low-level equilibrium corresponding to a poverty trap

does not emerge. In short, in the data, the world looks more like panel B of Figure 1 than panel A.

Second, Kraay and Raddatz (2007) consider a growth model with endogenous optimal saving decisions, coupled with a subsistence consumption constraint. They find that at low levels of development, growth reflects the balance of two forces. On the one hand, at low levels of development, the marginal utility of consumption is very high and this lowers optimal saving rates and slows growth. On the other hand, the marginal product of capital starts comparatively high because there are low levels of capital investment and diminishing returns have not yet had a forceful effect, which encourages saving, investment, and growth. For plausible calibrations of the key parameters, the attractiveness of saving and investment dominates, and countries quickly grow out of the subsistence constraint. Only when countries are very close to subsistence levels does the parameterized model suggest that saving and investment would be so low that growth could be expected to stagnate for extended periods of time. But this insight poses a challenge to the view that poverty traps are a widespread phenomenon given that income levels vary considerably among poor countries. Returning to the countries noted in the introduction, in order to argue that stagnant incomes in Burundi and Nicaragua are due to a subsistence consumption-type poverty trap, it is necessary to provide a rationale for why the relevant subsistence level of consumption is seven times higher in Nicaragua than it is in Burundi. Otherwise, theory predicts that Nicaragua would long since have outgrown subsistence constraints.

Big-Push Models of Poverty Traps

Another possible mechanism behind the poverty trap comes from the view that there are increasing returns to scale in the "modern" sector of the economy (for simplicity, usually thought of as manufacturing), and constant returns in the traditional sector (usually thought of as agriculture). If the economy devotes most of its resources to production in the traditional sector, wages will be equalized at a low level across sectors. If on the other hand the economy makes a "big push" to allocate most of its resources to the modern sector, it can realize the benefits of increasing returns and enjoy high wages in both sectors and a high overall income level. Direct empirical evidence on the importance of such mechanisms is scarce, most notably because it is difficult to disentangle the effects of policy efforts to promote "modern" sectors and activities from other factors driving aggregate growth. In the absence of such direct evidence, the literature has produced suggestive calibration exercises that point to somewhat conflicting conclusions.

On the one hand, Graham and Temple (2006) develop an innovative calibration strategy that allows them to deduce whether a second equilibrium level of output could exist for a country in addition to the one actually observed in the data. Their methodology also allows them to determine whether the second equilibrium, if it exists, implies a higher or lower income level than the one actually observed for the country. Applying this methodology to 127 countries, they find that for many of them, the existence of a second equilibrium is compatible with the data. Moreover, about a quarter of countries, predominantly poor ones, are in the lower of the two equilibria. For the typical country in the low-level

equilibrium, the corresponding high-level equilibrium involves an income level two to three times higher than the actual one. Such income differences resulting from countries finding themselves in a "bad" equilibrium certainly are nontrivial, although they are only a small fraction of the observed income differences between low-income and even middle-income countries.

On the other hand, Caucutt and Kumar (2008) study a "big push"–type model with a coordination failure arising from the fact that agents find it optimal to invest in labor-saving technologies only if other agents also do so. This coordination failure leads to a poverty trap when all agents fail to invest in the better technology. However, they argue that the existence of the low-level equilibrium is not particularly robust to small changes in parameter values, and particularly for more empirically plausible parameters of the model. Moreover, they find that even when a trap does exist, the one-time subsidy required to induce agents to invest and switch to the high equilibrium is only about 5 percent of income. If the "cure" to poverty is such a small intervention, this begs the question of why the much larger volumes of aid to poor countries have not succeeded in springing countries free from this particular trap.

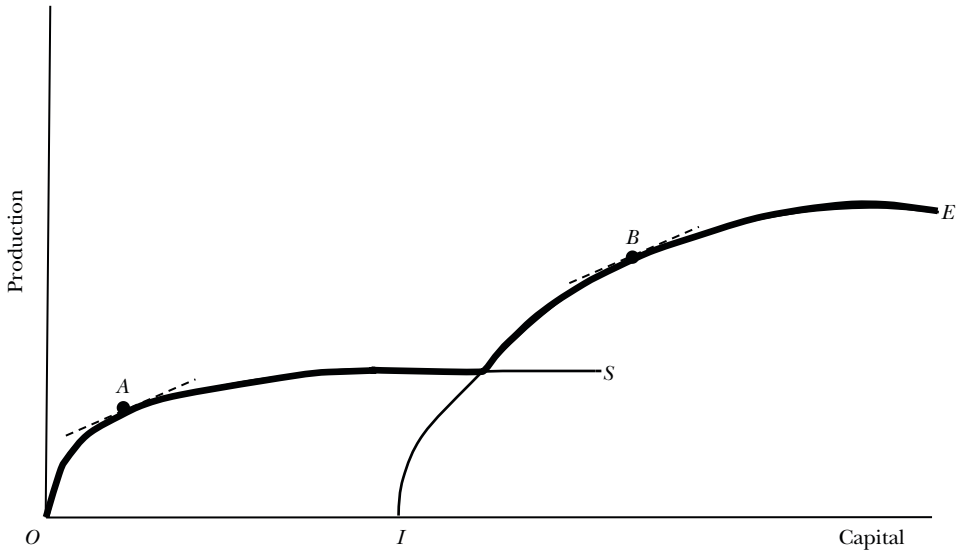
Nutritional Poverty Traps

One of the earliest examples of an S-shaped poverty trap at the individual level is based on nutrition levels (for example, Mazumdar 1959, formalized in Dasgupta and Ray 1986). In this model, poverty can be self-reinforcing because poor individuals are too malnourished to physically be able to do productive work, thereby not earning enough or producing enough food to alleviate this malnourishment. The idea is that the link between food intake and physical work capacity is nonlinear with increasing returns when one starts from a low consumption level.

However, such traps appear unlikely for most people. First, calories are too cheap in most of the world for many of the poor to be in a nutritional poverty trap. For example, Subramanian and Deaton (1996) calculated that the cost of the calories needed for a day's activity in rural Maharashtra were less than 5 percent of the daily wage, while more recently Banerjee and Duflo (2011) calculate that 2,400 daily calories can be obtained in the Philippines for as little as 21 cents (in purchasing power parity terms). There may be specific situations where these dynamics can arise: for example, during famines where markets break down and prices rise (Ravallion 1997) or in cases of continual fecal-oral contamination in which the body is unable to absorb many nutrients, effectively raising the price per absorbed calorie (Ravallion forthcoming).

Second, while both calories and micronutrients do affect productivity, a variety of studies summarized by Strauss and Thomas (1998) suggest that the shape of this relationship tends not to follow the S-shaped pattern needed for poverty trap dynamics. Of course, these points do not mean there is no scope for policy efforts to ease malnutrition; indeed, a large literature shows lasting impacts of nutritional deficiencies in early childhood and thus implies long-term benefits of early interventions to overcome these deficiencies (for example, Behrman, Alderman, and Hoddinott 2004). But the rationale for such policies should not rest on pushing people over some critical threshold in a poverty trap.

Figure 3
Production Non-convexity Arising from Choice between Two Technologies

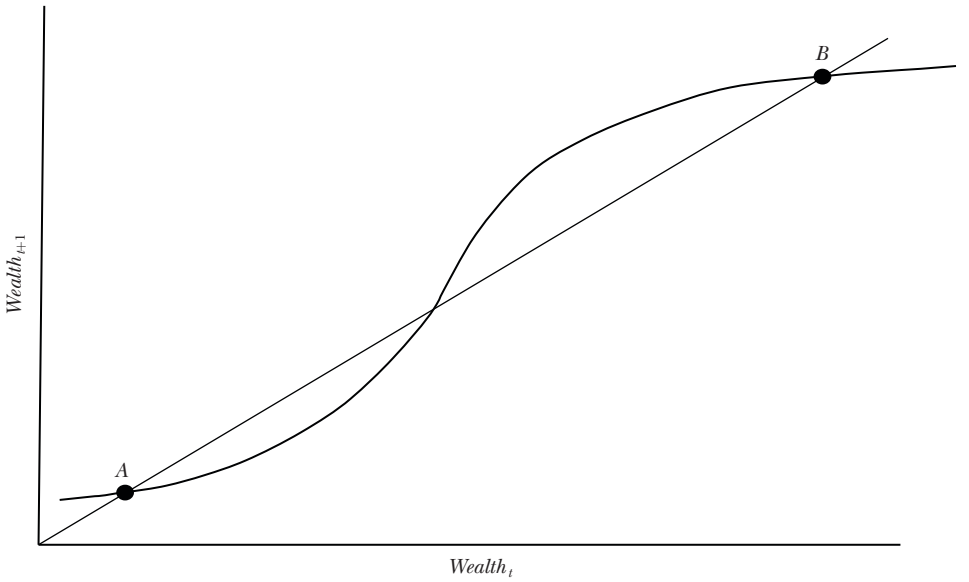


Note: Following Banerjee and Duflo (2011), Figure 3 then shows the subsistence production technology (curve *OS*) and entrepreneurial activity (curve *IE*), and the combined nonconvex production set *OE* (the darkened curve).

Lumpy Investments Coupled with Borrowing Constraints

A number of models of poverty traps are based on an interaction between capital market imperfections, which restrict the amount individuals can borrow, and the idea that the production technology is nonconvex—that is, there is a range where investing a little has low returns and investing a lot more has much higher returns. For example, in the models of Banerjee and Newman (1993) and Aghion and Bolton (1997), individuals may choose between a subsistence activity that requires no fixed capital investment and an entrepreneurial activity that requires a minimum outlay of capital *I*. Following Banerjee and Duflo (2011), Figure 3 then shows the subsistence production technology (curve *OS*) and entrepreneurial activity (curve *IE*), and the combined nonconvex production set *OE* (the darkened curve). Individuals with low initial wealth who cannot borrow begin in the subsistence production activity, while individuals with high initial wealth become entrepreneurs. In the absence of borrowing, individuals choosing between consumption and investment in their business will choose a production level such that the marginal return on investment equals the marginal rate of substitution between consumption today and consumption in the future. This gives rise to two steady-state production points given by *A* and *B* in Figure 3, and to the S-shaped wealth dynamics in Figure 4. Individuals who begin with too little wealth will be stuck in the poverty trap at *A*—returns to capital are too low for them to bootstrap

Figure 4
Poverty Trap Corresponding to Dual Technology Choice

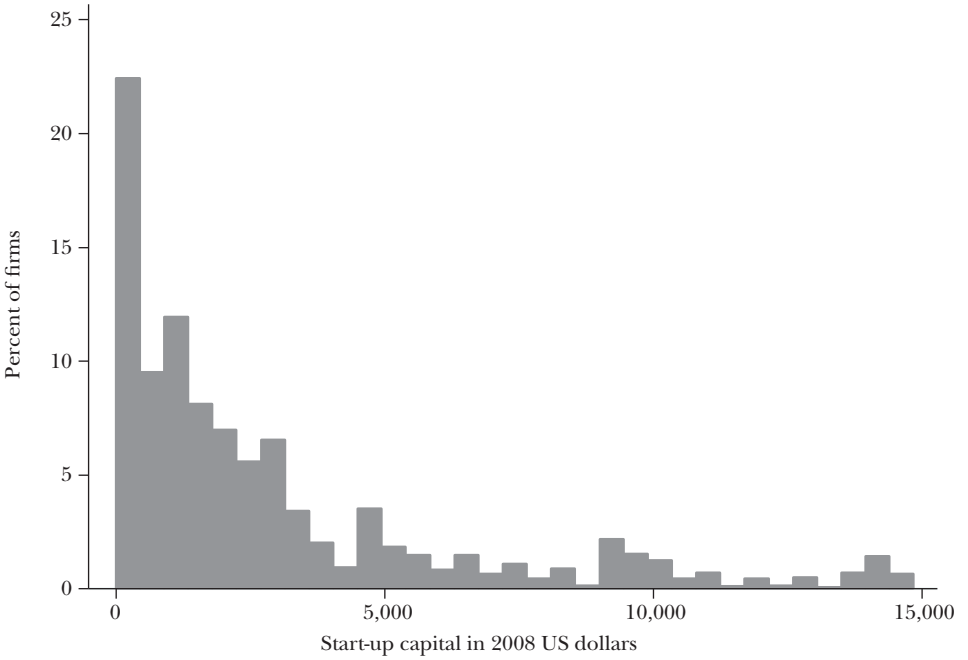


their way out of poverty, while borrowing constraints prevent them reaching the scale needed for the entrepreneurial activity.

How realistic is this model? The closest practical analog to this theoretical case can be found in studies of pastoralists in rural Ethiopia conducted by Lybbert, Barrett, Desta, and Coppock (2004) and Santos and Barrett (2011), and by Barrett et al. (2006) among rural pastoralists in Kenya. Raising cattle is the only production activity for these individuals, and looking at cattle herd size dynamics, these studies find evidence for a process like that in Figure 4, whereby individuals who start with a small herd size appear trapped around a very low level (A in the figure), whereas there is also a much higher-level equilibrium (B). They suggest that a minimum herd size is required in order to be able to undertake the migratory herding necessary to deal with variation in water availability and forage, and those with smaller herds than this size must stay near their base camps, where land degradation means only a small herd size can be supported. Moreover, Santos and Barrett (2011) show that these asset dynamics are known to the farmers, and those too close to the low-level steady state are excluded from informal credit markets. So in this remote location, with only a single nonconvex production technology and no access to credit, some pastoralists appear trapped in poverty as theory would suggest. The key question, to which we will return, then becomes why more people don't leave such areas.

In contrast, in most of the world, individuals face multiple production technologies. Individuals have a choice over many different types of sectors—for example, they might make clothing, sell food, or sell other goods—and, within a sector, a

Figure 5
The Empirical Distribution of Start-up Capital Shows There Are Many Production Choices



Source: Sri Lankan Longitudinal Survey of Enterprises, Baseline data.
Notes: There are 2,255 microenterprises in the survey, and the graph is for the 2,019 of them in the bottom 90 percent in terms of start-up capital.

choice over which items to make and sell. Empirically this gives rise to several fact patterns that, taken together, are difficult to reconcile with models of poverty-traps like the one in Figures 3 and 4.

First, in practice, it seems individuals don't need a lot of capital to start a business; many business owners start with very low levels of capital, and some start with zero capital. Second, the amount of capital needed to start a business appears relatively continuous—it does not seem that business owners only have a small number of production technologies they can choose amongst. As an example, Figure 5 plots the distribution of start-up costs for 2,019 urban Sri Lankan microenterprises in the baseline of the Sri Lanka Longitudinal Survey of Enterprises (SLLSE) (de Mel, McKenzie, and Woodruff 2009). Start-up capital includes land, buildings, machinery, and working capital. The data show a wide range of initial start-up levels of capital, with 10 percent of firms starting with US$100 or less and 23 percent with $500 or less. McKenzie and Woodruff (2006) find a similar range of different starting capital levels for Mexican microenterprises, with many starting with low amounts.

When many different production technologies are available, each with a different initial starting cost, then the overall production technology appears

convex. Imagine Figure 3 with ten rather than two production technologies. The outer curve (shown in bold in Figure 3) would still have small areas of locally increasing returns, but the curve becomes much smoother than with two technologies. In these kinds of figures, nonconvexity is the tell-tale sign of a potential poverty trap, because it will be difficult for those with low incomes to jump across the nonconvexity; but as one moves to many production choices across a range of levels of start-up capital, the overall production choice set becomes more convex.[4]

If individuals face many different production technologies and can choose among them (for example, by adding another product line to their retail store, or selling off one business and using the proceeds to start another), then a multiple-equilibria-based poverty trap should not arise. Even if individuals are unable to borrow, they should be able to start small, reinvest some of their earnings, and slowly bootstrap their way towards a higher production level. Moreover, to the extent that individuals are liquidity constrained and far from their optimal production levels, additional small capital investments should have high returns. This prediction is in sharp contrast to the dual technology poverty traps model in which individuals who start small will have very low returns on capital unless they make a very big investment.

A third fact that casts doubt on technology-based poverty trap models is that many small-scale businesses do in fact have high returns to capital. Several studies, including randomized experiments that give one-time grants of $100 or $200 to microenterprise owners find high returns to capital when starting at a relatively low level of capital stock. In Sri Lanka, de Mel, McKenzie, and Woodruff (2008, 2012) find a return to capital of 5 percent per month on average. In Mexico, using nonexperimental methods, McKenzie and Woodruff (2006) find marginal returns to capital averaging 15 percent per month for male business owners with less than $200 in capital, with a follow-up experiment finding returns of 20 percent per month or more. Finally, in Ghana, Fafchamps, McKenzie, Quinn, and Woodruff (2014) also find returns averaging 20 percent per month and evidence that the impact is higher when capital is given in-kind than as an unrestricted cash grant.[5]

Finally, if all that was holding individuals back from moving from a low-level steady-state like *A* in Figure 4 to operating at a higher level that is out of poverty, like *B*, was access to finance, then one would expect microfinance loans to have large effects on poverty—at least if individuals could borrow enough to finance the lumpy production technology needed to move them out of a trap. However, a number of randomized experiments have now offered microfinance to randomly selected individuals or communities, and none of them have found strong evidence of impacts on business growth or poverty (see Duflo, Banerjee, Glennerster, and Kinnan, 2013,

[4] A distinct issue that can then arise is whether there are nonconvexities further up the scale distribution that prevent small firms from turning into medium firms, or medium firms into large firms. Hsieh and Olken discuss this issue in this symposium.

[5] One disconcerting thread in this line of research is that in some studies the higher returns for small investments appear for men but not for women. For example, de Mel, McKenzie, and Woodruff (2008, 2012) find returns that are higher for men and near zero for female owners. Similarly, Fafchamps et al. (2014) find no evidence of returns for women running very small businesses.

for an example, and their last section for an overview of several recent studies). At least in most of the low-income world, it seems unlikely that the combination of a limited number of lumpy productive investments, coupled with constraints on borrowing, are causing poverty traps.

While these facts taken together help to rule out multiple-equilibria poverty traps, some individuals may remain persistently in poverty because their underlying characteristics may be such that their single steady-state level of production is extremely low (Barrett and Carter 2013). Regardless of the amount of financing available, individuals with poor skills and facing other constraints may be unable to produce at more than a subsistence level. With this possibility in mind, a number of policy interventions have sought to combine assets with intensive skills training and savings promotion. Examples include the BRAC ultra-poor program evaluated by Bandiera et al. (2013) among rural women in Bangladesh; the Bardhan hard-core program evaluated by Banerjee, Duflo, Chattopadhyay, and Shapiro (2011) in rural West Bengal, India; and the SKS ultra-poor program evaluated by Morduch, Ravi, and Bauchet (2012) in rural Andhra Pradesh, India. These programs were aimed at extremely poor households, with eligibility requirements for households being that they have no male workers, not own productive assets, have limited land, and not be microfinance clients. The Bandiera et al. (2013) study finds this combination intervention has lasting effects over a four-year horizon, with an increase in earnings that is large in relative terms, but small in absolute value (approximately US$0.07 per day higher). Banerjee et al. (2011) find similar results over an 18-month horizon, with program households receiving an approximate US$0.06 per day increase in per capita earnings. In contrast, Morduch, Ravi, and Bauchet (2012) find no overall effect on income growth, which they attribute to the presence of opportunities for wage labor that opened up at the same time as the ultra-poor program and enabled the control group to grow at the same rate.

These programs demonstrate that the combination of asset support, training, and other assistance can take some individuals from being extremely poor to being just poor. It is difficult to discern from this evidence whether the relatively small gains, in absolute terms, involve escaping a poverty trap or just a standard dynamic of greater financial and human capital leading to higher income. In either case, the individuals remain quite poor, although it is possible that as more time passes, the income gains from the intervention will increase.

Behavioral Poverty Traps

For individuals to be able to grow out of poverty from a low initial level of capital, they need to save and reinvest continually. However, recent work in behavioral economics suggests that poverty may be self-reinforcing because of the way it affects decision making. Banerjee and Mullainathan (2010) provide a theoretical example of a poverty trap. In this model, individuals allocate spending between temptation and nontemptation goods, and the fraction of the marginal dollar spent on temptation goods declines with total amount spent. This model leads to poor individuals being present-biased and unwilling to take on small high-return investments, keeping them poor. Shah, Mullainathan, and Shafir (2012) provide another potential explanation in which scarcity causes individuals in poverty to devote more

mental effort to meeting daily needs, leaving less attentional resources for other problems—such as perhaps thinking about how to grow their firm. A consequence, in these theories, is that poor business owners may lack the self-control to reinvest cash in their business, preventing them from making investments that would slowly allow them to grow out of poverty.

Some suggestive evidence supports the potential importance of this form of the poverty trap. Ghana, Fafchamps, McKenzie, Quinn, and Woodruff (2014) find that a one-time grant given as cash had much lower effects on business profitability than the same size grant given in kind among female microenterprise owners, with some evidence to suggest this is driven by self-control. Schaner (2013) finds persistent effects on nonfarm business growth of short-term incentives to save among Kenyan households, which she attributes to them setting up "mental accounts" geared towards the business. The control group of firms in the Sri Lankan grant experiment of de Mel, McKenzie, and Woodruff (2012) still have not caught up to the treatment group over five years later, although the treatment group largely makes divisible investments of working capital. It appears that small business owners are not taking on some investments that are divisible and profitable, limiting their ability to reinvest their way towards a high steady-state. Indeed, Kremer, Lee, Robinson, and Rostopshova (2013) and Duflo, Kremer, and Robinson (2011) provide additional examples of small business owners not undertaking small investments with high marginal returns. Nevertheless, these same types of small firms do appear to grow over time in a growing economy (de Mel, McKenzie, and Woodruff 2013), suggesting that any behavioral constraints on firm growth do not prevent the owners responding over time to rising demand. This suggests that small positive shocks may be enough to shift the equilibrium business size, so any "trap" would still not be tied to an absolute size threshold.

Micro Meets Macro: Geographic Poverty Traps

The evidence most consistent with poverty traps comes from poor households in remote rural regions—whether it is the work among East African pastoralists or the suggestive evidence about ultra-poor programs breaking some individuals out of poverty traps. In remote rural areas, isolation reduces the number of available production technologies, which means the choice between lower-income and higher-income outcomes may be a more difficult discrete step (as illustrated back in Figure 3). Jalan and Ravallion (2002) define a "geographic poverty trap" as occurring when the characteristics of a geographic region are such that a household's consumption cannot rise over time while an otherwise identical household living in a different, better-endowed area would enjoy a rising standard of living. They use farm household panel data from rural China to estimate a dynamic consumption model that allows fixed or slowly changing geographic characteristics to influence household consumption and find evidence to support the idea that, all else equal, living in a poor area lowers the productivity of a farmer's own investments, with areas with insufficient geographical capital, such as few rural roads, potentially in geographic poverty traps.

Why don't more people move out of poor areas? In the Chinese context, mobility restrictions such as the household registration (*hukou*) system under which households

are only entitled to social services like health and education if they are living in the area of their registration, along with thin rural land markets and uncertain land property rights, might explain low mobility (Jalan and Ravallion 2002). But many people remain in poor rural areas throughout much of the world, despite high returns to internal migration. Cost can be one reason—the same credit market failures that prevent investing in more productive assets can also prevent households financing the costs of moving. Bryan, Chowdhury, and Mobarak (2013) conduct experimental work in Bangladesh to try to understand why more households don't undertake seasonal migration, which they show to have high returns; part of the explanation, they suggest, is that households close to subsistence are unwilling to take the risk of migration; but they become more willing to do so if insured against this risk. The result is that policies to facilitate more internal migration may help lift some people out of poverty traps. However, concerns about congestion and the development of urban slums, along with other political concerns, often make policymakers reluctant to encourage more internal migration in many low-income countries.

More broadly, in our view, the strongest evidence for a poverty trap at an individual level is the one based on country of residence. The same individual can be in persistent poverty or earning substantially more depending solely on the country they happen to work in. For example, using a visa lottery, McKenzie, Gibson, and Stillman (2010) estimate that Tongan workers experience a 263 percent increase in their (purchasing-power-adjusted) income within a few months of migrating to New Zealand. Clemens, Montenegro, and Pritchett (2008) estimate that a male worker with nine years of education would earn 7.8 times as much working in the United States as in Haiti. Effectively, the different productive technologies in Figure 3 operate in different countries, and poor people are unable to afford the cost of movement that would allow them to earn substantially higher incomes.

To make this concrete, consider two individuals in Mexico of identical abilities but different initial wealth levels. The cost of migrating to the United States, where they could be substantially more productive and earn multiples of their Mexican incomes, is not affordable to a liquidity-constrained poor individual. In contrast, the wealthier individual can pay for the recruiting fees, skills training, certification costs, and transportation costs necessary for different types of legal migration through a work visa; pay the costs needed to migrate through an investor visa; or pay to hire a *coyote* (smuggler) to enable them to migrate illegally. The result is that migrants will be positively selected on wealth (McKenzie and Rapoport 2007), with poor individuals unable to pay the migration costs and therefore remaining poor, while slightly wealthier individuals who can pay these costs will be able to migrate and then move to the higher-equilibrium income.

Implications for Policy

Overall, our view of the existing literature finds no strong evidence for many of the common mechanisms theorized to give rise to poverty traps. Coupled with the fact that even poor countries have managed to grow at a rate similar to the historical

growth rate of the United States over the past 200 years, we suggest that one should be skeptical of claims that countries need a "big push" of aid or loans to take them over some threshold at which their growth prospects will shift dramatically. Likewise, one should be skeptical of claims that microfinance will be the solution to poverty traps caused by the interaction of financial constraints and lumpy investment technologies or that there are many workers for whom receiving more calories will be the difference between them being stuck working at low productivity or not.

Of course, this rather mixed evidence for poverty traps does not imply that there is no economic case for improving nutrition or for improved access to finance through microfinance at the individual level. Even if households are not in a poverty trap, they can be in a situation where they are converging only very slowly to a steady state and thus still benefit from aid. Aid projects that reduce premature mortality are an obvious example where a lack of poverty traps need not preclude large benefits from aid. Many other successful aid projects like conditional cash transfers based on school attendance or health care targets, deworming, school reforms, and others have outcomes that improve the well-being of recipients in important ways even though, in our view, such programs are unlikely to move that person from one equilibrium to another—or at least to a very different one.

The same is true at the aggregate level. While we have not seen much evidence in support of mechanisms that would imply that a large scaling-up of aid is *necessary* to deliver positive growth effects, this does not mean that aid cannot have positive effects on aggregate growth. Even if a country is not specifically in a poverty trap, people in that country may be persistently poor due to poor fundamentals, and aid-financed investments can certainly help to improve these fundamentals, thereby leading to higher growth. However, it is difficult to argue that such aid programs are likely to lead to a sharp acceleration of growth at the aggregate level as a country breaks free from a poverty trap.

While the evidence indicates that poverty traps are rare, this does not mean they can never exist. The clearest evidence for traps appears to come from people being trapped in low-productivity locations—whether this be remote rural regions within a country, or in low-productivity countries. Policy efforts to lower the barriers to internal and international mobility therefore appear to offer large potential payoffs in terms of taking people out of poverty.

We conclude with two important qualifications. First, a key feature of the empirical evidence reviewed in this paper is that individual studies have focused on documenting specific mechanisms in isolation. In reality the world is of course much more complicated, and it is plausible that many trap-like forces might be simultaneously at play, at both the individual and the country level. If there are important interactions across different mechanisms, then these trap-like mechanisms might jointly impede development, even if in isolation they do not appear to matter all that much. More theoretical and empirical work on such interactions could be a promising direction for future research. Second, although we argue that many models of poverty traps do not seem to be that prevalent empirically, we do not think that economists' efforts over the past several decades to develop these models and confront them with the data have been a waste of time. Rather, models

of poverty traps, arguably, have provided an intellectually coherent and rigorous framework for efforts to understand the process of economic development. While both the reduced-form relationships as well as efforts to document particular mechanisms have yielded rather mixed evidence about the importance of poverty traps, these models and the associated empirical efforts to test them have greatly enhanced our understanding of a wide range of important development issues, such as how markets function in developing countries; what the returns are to a range of capital, educational, and nutritional investments; and what other barriers may be holding back growth.

■ *We thank the editors, Chris Barrett, Luis Serven, and Martin Ravallion for helpful comments. The views expressed here are the authors' and do not reflect those of the World Bank, its Executive Directors, or the countries they represent.*

References

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

Acemoglu, Daron, and Fabrizio Zilibotti. 1997. "Was Prometheus Unbound by Chance? Risk, Diversification, and Growth." *Journal of Political Economy* 105(4): 709–751.

Adato, Michelle, Michael R. Carter, and Julian May. 2006. "Exploring Poverty Traps and Social Exclusion in South Africa using Qualitative and Quantitative Data." *Journal of Development Studies* 42(2): 226–47.

Aghion, Philippe, and Patrick Bolton. 1997. "A Theory of Trickle-Down Growth and Development." *Review of Economic Studies* 64(2): 151–72.

Antman, Francisca, and David McKenzie. 2007. "Poverty Traps and Nonlinear Income Dynamics with Measurement Error and Individual Heterogeneity." *Journal of Development Studies* 43(6): 1057–83.

Azariadis, Costas, and John Stachurski. 2005. "Poverty Traps." Chap. 5 of *Handbook of Economic Growth,* Vol. 1, Part A, edited by Philippe Aghion and Steven N. Durlauf. Elsevier.

Bandiera, Oriana, Robin Burgess, Narayan Das, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. 2013. "Can Basic Entrepreneurship Transform the Economic Lives of the Poor?" STICERD—Economic Organisation and Public Policy Discussion Paper 43.

Banerjee, Abhijit V., and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* New York, NY: Public Affairs.

Banerjee, Abhijit, Esther Duflo, Raghabendra Chattopadhyay, and Jeremy Shapiro. 2011. "Targeting the Hard-Core Poor: An Impact Assessment." Unpublished paper. http://www.povertyactionlab.org/publication/targeting-hard-core-poor-impact-assessment.

Banerjee, Abhijit V., and Andrew F. Newman. 1993. "Occupational Choice and the Process of Development." *Journal of Political Economy* 101(2): 274–98

Banerjee, Abhijit, and Sendhil Mullainathan. 2010. "The Shape of Temptation: Implications for the Economic Lives of the Poor." NBER Working Paper 15973.

Barrett , Christopher B., Paswel Phiri Marenya, John McPeak, Bart Minten, Festus Murithi, Willis Oluoch-Kosura, Frank Place, Jean Claude Randrianarisoa, Jhon Rasambainarivo, and Justine Wangila. 2006. "Welfare Dynamics in Rural Kenya and Madagascar." *Journal of Development Studies* 42(2): 248–77.

Barrett, Christopher B., and Michael R. Carter. 2013. "The Economics of Poverty Traps and Persistent Poverty: Empirical and Policy Implications." *Journal of Development Studies* 49(7): 976–90.

Behrman, Jere R., Harold Alderman, and John Hoddinott. 2004. "Hunger and Malnutrition." Copenhagen Consensus Challenge Paper, http://greencardeb5expert.com/PDF/Hunger_and_Malnutrition.pdf.

Ben-David, Dan. 1998. "Convergence Clubs and Subsistence Economies." *Journal of Development Economics* 55(1): 155–71.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak. 2013. "Escaping Famine through Seasonal Migration." Yale Economics Department Working Paper 124, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2348362.

Caucutt, Elizabeth, and Krishna B. Kumar. 2008. "Africa: Is Aid an Answer?" *BE Journal of Macroeconomics: Advances* 8(1): Article 32.

Chen, Shaohua, and Martin Ravallion. 2011. "Weakly Relative Poverty." *Review of Economics and Statistics* 93(4): 1251–61.

Clemens, Michael. 2012. "New Documents Reveal the Cost of 'Ending Poverty' in a Millennium Village: At Least $12,000 per Household." Center for Global Development, http://www.cgdev.org/blog/new-documents-reveal-cost-%E2%80%9Cending-poverty%E2%80%9D-millennium-village-least-12000-household.

Clemens, Michael, Claudio E. Montenegro, and Lant Pritchett. 2008. "The Place Premium: Wage Differences for Identical Workers across the U.S. Border." Center for Global Development Working Paper 148.

Dasgupta, Partha, and Debraj Ray. 1986. "Inequality as a Determinant of Malnutrition and Unemployment: Theory." *Economic Journal* 96(384): 1011–34.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2008. "Returns to Capital in Microenterprises: Results from a Field Experiment." *Quarterly Journal of Economics* 123(4): 1329–72.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2009. "Innovative Firms or Innovative Owners? Determinants of Innovation in Micro, Small, and Medium Enterprises." World Bank Policy Research Working Paper 4934.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2012. "One-time Transfers of Cash or Capital Have Long-lasting Effects on Microenterprises in Sri Lanka." *Science* 335(6071): 962–66.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2013. "What Generates Growth in Microenterprises? Experimental Evidence on Capital, Labor and Training." http://economics.mit.edu/files/8666.

Duflo, Esther, Abhijit Banerjee, Rachel Glennerster, and Cynthia Kinnan. 2013. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." NBER Working Paper 18950.

Duflo, Esther, Michael Kremer, and Jonathan Robinson. 2011. "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya." *American Economic Review* 101(6): 2350–90.

Easterly, William. 2006. "Reliving the 1950s: The Big Push, Poverty Traps, and Takeoffs in Economic Development." *Journal of Economic Growth* 11(4): 289–318.

Fafchamps, Marcel, David McKenzie, Simon Quinn, and Christopher Woodruff. 2014. "Microenterprise Growth and the Fly-paper Effect: Evidence from a Randomized Experiment in Ghana." *Journal of Development Economics* 106(C): 211–26.

Galor, Oded, and David N. Weil. 2000. "Population, Technology and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond." *American Economic Review* 90(4): 806–28.

Ghatak, Maitreesh, Massimo Morelli, and Tomas Sjostrom. 2001. "Occupational Choice and Dynamic Incentives." *Review of Economic Studies* 68(4): 781–810.

Graham, Bryan S., and Jonathan Temple. 2006. "Rich Nations, Poor Nations: How Much Can Multiple Equilibria Explain?" *Journal of Economic Growth* 11(1): 5–41.

Hausman, Ricardo, Lant Pritchet, and Dani Rodrik. 2005. "Growth Accelerations." *Journal of Economic Growth* 10(4): 303–29.

Jalan, Jyotsna, and Martin Ravallion. 2002. "Geographic Poverty Traps? A Micro Model of Consumption Growth in Rural China." *Journal of Applied Econometrics* 17(4): 329–46.

Jalan, Jyotsna, and Martin Ravallion. 2004. "Household Income Dynamics in Rural China." In *Insurance Against Poverty,* edited by Stefan Dercon, 108–124. Oxford: Oxford University Press.

Kraay, Aart, and Claudio Raddatz. 2007. "Poverty Traps, Aid, and Growth." *Journal of Development Economics* 82(2): 315–47.

Kremer, Michael, Jean Lee, Jonathan Robinson, and Olga Rostapshova. 2013. "The Return to Capital for Small Retailers in Kenya: Evidence from Inventories." Unpublished paper, Harvard.

Lokshin, Michael, and Martin Ravallion. 2004. "Household Income Dynamics in Two Transition Economies." *Studies in Nonlinear Dynamics and Econometrics* 8(3): Article 4.

Lopez, Humberto, and Luis Servén. 2009. "Too Poor to Grow." World Bank Policy Research Working Paper 5012.

Lucas, Robert E. B., and Sari Pekkala Kerr. 2013. "Intergenerational Income Immobility in Finland: Contrasting Roles for Parental

Earnings and Family Income." *Journal of Population Economics* 26(3): 1057–94.

Lybbert, Travis, Christopher Barrett, Solomon Desta, and D. Layne Coppock. 2004. "Stochastic Wealth Dynamics and Risk Management among a Poor Population." *Economic Journal* 114(498): 750–77.

Mazumdar, Dipak. 1959. "The Marginal Productivity Theory of Wages and Disguised Unemployment." *Review of Economic Studies* 26(3): 190–97.

McKenzie, David, John Gibson, and Steven Stillman. 2010. "How Important is Selection? Experimental vs. Non-experimental Measures of the Income Gains from Migration." *Journal of the European Economic Association* 8(4): 913–45.

McKenzie, David, and Hillel Rapoport. 2007. "Network Effects and the Dynamics of Migration and Inequality: Theory and Evidence from Mexico." *Journal of Development Economics* 84(1): 1–24.

McKenzie, David, and Christopher Woodruff. 2006. "Do Entry Costs Provide an Empirical Basis for Poverty Traps? Evidence from Mexican Microenterprises." *Economic Development and Cultural Change* 55(1): 3–42.

McKenzie, David J., and Christopher Woodruff. 2008. "Experimental Evidence on Returns to Capital and Access to Finance in Mexico." *World Bank Economic Review* 22(3): 457–82.

Morduch, Jonathan, Shamika Ravi, and Jonathan Bauchet. 2012. "Failure vs. Displacement: Why an Innovative Anti-Poverty Program Showed No Net Impact." PRIMCED Discussion Paper Series, No. 32.

Naschold, Felix. 2013. "Welfare Dynamics in Pakistan and Ethiopia—Does the Estimation Method Matter?" *Journal of Development Studies* 49(7): 936–54.

Nelson, Richard R. 1956. "A Theory of the Low-Level Equilibrium Trap in Underdeveloped Economies." *American Economic Review* 46(5): 894–908.

Quah, Danny T. 1997. "Empirics for Growth and Distribution: Stratification, Polarization, and Convergence Clubs." *Journal of Economic Growth* 2(1): 27–59.

Ravallion, Martin. 1997. "Famines and Economics." *Journal of Economic Literature* 35(3): 1205–42.

Ravallion, Martin. Forthcoming. "The Idea of Antipoverty Policy." In *Handbook of Income Distribution*, Vol. 2. Elsevier.

Santos, Paulo, and Christopher B. Barrett. 2011. "Persistent Poverty and Informal Credit." *Journal of Development Economics* 96(2): 337–47.

Schaner, Simone. 2013. "The Persistent Power of Behavior Change: Long-run Impacts of Temporary Savings Subsidies for the Poor." http://www.dartmouth.edu/~sschaner/main_files/Schaner_LongRun.pdf.

Shah, Anju, Sendhil Mullainathan, and Eldar Shafir. 2012. "Some Consequences of Having Too Little." *Science* 338(6107): 682–85.

Strauss, John, and Duncan Thomas. 1998. "Health, Nutrition, and Economic Development." *Journal of Economic Literature* 36(2): 766–817.

Subramanian, Shankar, and Angus Deaton. 1996. "The Demand for Food and Calories." *Journal of Political Economy* 104(1): 133–62.

UNCTAD. 2006. "Economic Development in Africa: Doubling Aid: Making the 'Big Push' Work." Geneva: United Nations Conference on Trade and Development.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 149–168

Page Limits on Economics Articles: Evidence from Two Journals[†]

David Card and Stefano DellaVigna

Forty years ago the top journals in economics published relatively short articles. The median length of the papers in the *American Economic Review, Econometrica,* the *Journal of Political Economy,* the *Quarterly Journal of Economics,* and the *Review of Economic Studies* in the early 1970s was under 20 pages. As shown in Figure 1, the typical length of the articles published in these journals has nearly *tripled* since then to around 50 pages.[1] The trends have been similar at all five journals and also across fields, leading to widespread concern about the allocation of journal space and the readability of articles.

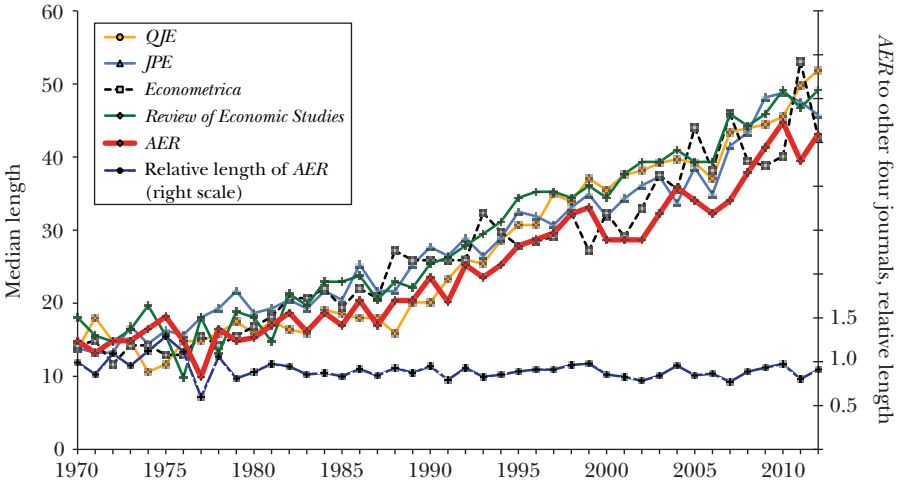
In response to these concerns the *American Economic Review (AER)* introduced a page limit for submissions in 2008, becoming the first and still the only one of the top five economics journals to do so: as explained below, the *AER* page limit is 40 pages for papers with 1.5-line spacing and 50 pages for double spacing. The policy change was significant: over 40 percent of submissions in the previous year had exceeded what became the new limit. In March 2009, the *Journal of the European*

[1] We use a "standardized" page length. The *AER* is relatively dense: current articles are about 30–40 pages, which is about 50 of our standardized pages. The *QJE* lays out a page in a more readable format; their articles are 50+ pages on average.

■ *David Card is the Class of 1950 Professor of Economics and Stefano DellaVigna is the Daniel Koshland, Sr. Distinguished Professor of Economics and Professor of Business Administration, both at the University of California, Berkeley, California. Card is Director of the Labor Studies Program and DellaVigna is a Faculty Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are card@econ.berkeley.edu and sdellavi@econ.berkeley.edu.*

[†] To access the online Appendix, data Appendix, and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.3.149 doi=10.1257/jep.28.3.149

Figure 1
Standardized Lengths of Papers Published by Top-Five Journals



Notes: The measures of page length shown in Figure 1 are standardized to take into account the differences over time and between journals in the number of characters per page of text. Shorter papers, comments, replies, and notes are also excluded. 2012 page lengths are based on articles published before November 2012.

Economic Association (JEEA) followed suit, adapting the *AER* policy nearly verbatim. Within a few months *JEEA* abandoned the submission limit, though it remains in effect at the *AER*.

We use anonymized submissions to *AER* and *JEEA* to evaluate the costs and benefits of the policy change at the two journals. On the benefit side, shorter submissions may lead to shorter published papers, freeing up space for additional articles. Space in the top economics journals is very scarce. Despite some increase in total pages published, the top five journals currently publish fewer articles per year than they did in the 1970s and have far lower acceptance rates (Card and DellaVigna 2013). Shorter submissions may also be easier for editors and referees to evaluate, helping to reverse the trend toward longer adjudication times (Ellison 2002). On the cost side, the introduction of page limits may cause some authors to choose another outlet for their paper rather than spend the time to shorten it. The loss of longer manuscripts is a concern if longer articles are more likely to be cited, as "impact factors" based on citations are widely used to compare journals. Shorter submissions may also be harder to read if authors use formatting tricks to meet the page limit threshold or to suppress important details of their work.

The fraction of authors who respond to a page limit by diverting their work to other outlets provides direct information on the elasticity of supply to a particular journal and the degree of market power the journal enjoys. We find that the

introduction of page limits at the *AER* led to an immediate drop in the number of longer submissions and the emergence of a spike in the distribution of page lengths centered around the 40-page limit. Comparing the inflow of new submissions that were at or above the threshold to the inflow of papers below the page limit, we conclude that there was no loss of longer papers at the *AER*. Instead, we infer that a typical author was willing to shorten his or her paper by *at least* 22 pages to meet the *AER's* guidelines. This implies that the *AER* is viewed as highly differentiated from other journals.

How did authors accomplish these cuts? On average, the policy led to a four-page shortening of submitted manuscripts. Two pages were eliminated by formatting changes; 1.5 pages were cut by reducing the length of appendices submitted for publication; and 0.5 pages were eliminated by cutting tables and figures. Controlling for formatting changes, we find *no change* in the number of pages of basic text material submitted by authors. Moreover, we find no significant effect on the length of final accepted manuscripts at the *AER,* suggesting that the new policy did nothing to relieve the competition for journal space at the *AER*. This lack of response is confirmed by the absence of any trend in the relative length of published papers in the *AER* compared to the other top-five journals, shown by the bottom line in Figure 1.

We find a much different pattern of authors' responses to the page-length policy at the *Journal of the European Economic Association*. In particular there was no spike in the distribution of submissions following the introduction of page limits at *JEEA*. Instead, virtually all longer submissions were diverted to other outlets—a pattern that led the *JEEA* editors to reverse the policy after only a few months. The willingness of authors to "go elsewhere" points to a highly competitive market for economics articles at the level below the top five journals.

In the final section of the paper, we return to the interpretation of our findings and their implications for the design of editorial policies. In particular, we consider the question of whether shorter papers are indeed better. We show that prior to the page-length policy at the *AER* longer papers had a *higher* likelihood of receiving a revise-and-resubmit verdict than shorter papers. This is the opposite of a warning posted on the *AER* submission page in 2007 stating that longer papers were "rarely accepted for publication." It is consistent, however, with the pattern of substantially *higher* citation rates to longer papers published in the top five economics journals over the past four decades (Card and DellaVigna 2013) and with evidence on citations for submissions to the *JEEA,* which shows higher citations for longer submissions irrespective of whether they were ultimately accepted or rejected for publication in the *JEEA*.

In light of this evidence, we conclude that page length policies may be counterproductive for journals that face more elastic supplies of manuscripts, like *JEEA*. For journals with substantial market power, like *AER*, page length policies appear to impose fewer costs on the journal, but it is an open question whether the social costs to authors of shortening their papers are offset by other benefits. The most obvious potential benefit—of making published papers shorter—has not been realized.

Submission Behavior in the Presence of a Page Limit

The author of a paper perceives some payoff from submitting it to a particular journal, reflecting the likelihood it will be accepted for publication and the value of having the paper appear in that journal. If the highest-payoff journal imposes a binding page limit, the author faces a choice: shorten the paper, or submit it to the next best alternative. Assuming that the cost of shortening the paper is increasing in the size of the cuts needed to meet the limit, the author will shorten the paper if its length falls below a threshold that depends on the author's *match surplus*—the gap in payoffs between submission to the journal in question and the payoff to the next best alternative outlet (for more details, see the working paper version of Card and DellaVigna 2012). Depending on how page lengths and the journal-specific match surpluses are distributed, the imposition of a page limit at a journal will cause larger or smaller losses in submissions to that journal. At one extreme, when the match surpluses associated with a journal are large for nearly all potential authors, that journal may be able to impose a page limit and only lose a few very long papers. At the other extreme, when the match surpluses associated with a journal are small, most papers exceeding the journal's page limit will be diverted to other outlets, and only those requiring minimal cuts will be modified to meet the submission limit.

This simple model suggests that after the imposition of a binding page limit, the distribution of page lengths among submissions to the journal will exhibit a spike at the page limit reflecting the fraction of longer papers that have been cut to meet the threshold. If the policy commands complete compliance, there will be no submissions above the threshold, and the fraction of lost papers can be estimated by comparing the size of the spike to the number of papers that would have been above the threshold in the absence of the policy.

Actually conducting such a comparison is complicated by two factors. First, in the absence of a page-length policy, authors have wide latitude in how they format their paper and may use extra pages to make the paper easier to read—for example, by placing each figure on a separate page or by using wide margins. Once a length policy is introduced, authors of longer papers can often shorten their paper by making small (and legitimate) format changes. Thus, to actually compare page length distributions before and after a page-length policy one needs to develop standardized measures of page length. Once page lengths are standardized, the predicted "spike" in page lengths at exactly the page limit becomes fuzzy, making it harder to measure its size. A second concern is that pre–post comparisons of standardized page length distributions make no allowance for trends. As explained below, we therefore use comparisons between submission rates of papers of different lengths to conduct a more robust analysis.

Page Limits at the *American Economic Review*

Prior to 2008, the *AER* sought to discourage longer submissions but had no formal page-length policy. The instructions on the submission page in 2007

suggested that authors submit papers in double-spaced format using 12-point type, and noted: "Manuscripts longer than 50 pages are rarely accepted for publication." A sterner warning that "Manuscripts should not exceed 50 pages" was added in April 2008 but was not enforced by the editorial office.

In August 2008, Robert Moffitt, Editor of the *AER*, instructed the staff to enforce the existing page limit. A new policy with explicit page limits was posted in September 2008, with these key features:

1. All manuscripts must be formatted with 1.5 line spacing and must not exceed 40 pages (50-page limit applies to double-spaced manuscripts). This limit includes reference lists, figures, and tables.
2. Manuscripts must use 12-point font. Margins must be one inch top, bottom, and sides. Please use Times New Roman or similar font. These font, margin, and line-spacing requirements also apply to reference lists and tables.
3. You must include the words "Not for Publication" at the beginning of any lengthy appendix. The 40-page limit can be exceeded by an appendix if it is clearly marked as such.

The *AER* editorial office sent an email informing AEA members of the new policy on September 29, 2008.[2]

Comparing Manuscript Lengths Before and After

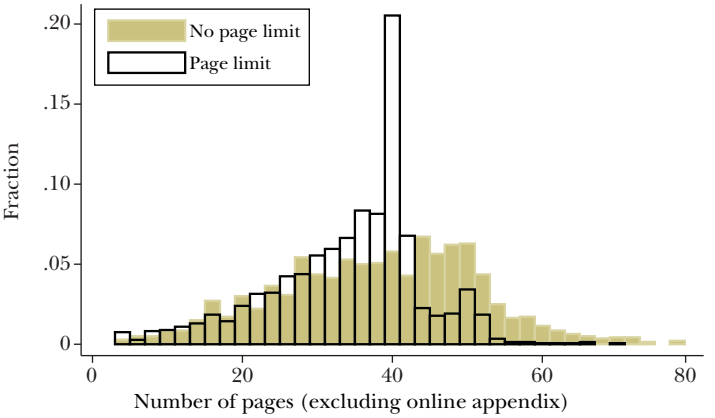
We collected data on new submissions to the *AER* from the year before and the year after the page-length policy change. Following the advice of the Managing Editor of the journal, we omit the transitional period from July 1, 2008, to September 30, 2008, and focus on the 1,406 manuscripts submitted from July 1, 2007, to June 30, 2008, and the 1,462 manuscripts submitted from October 1, 2008, to September 30, 2009. Detailed information on these 2,868 papers was collected by a temporary worker at the *AER* office using submission records from the Manuscript Central system.[3] (We drop a small number of manuscripts that had errors in submission or were otherwise invalid). We also collected data on the *final accepted* versions of all papers that were submitted during our sample period and accepted for publication by July 2012—a total of 224 manuscripts.

Figure 2 overlays the distributions of (unadjusted) page lengths among manuscripts submitted to the *AER* in the year before and the year after the introduction of the page limit policy. In the before period, long papers were relatively common,

[2] On July 1, 2011, the margin spacing rule was increased from 1 inch to 1.5 inches.

[3] Specifically, we measured the number of pages, the font, the margins, the lines per page, and the number of characters per page and per line. These detailed measures allow us to conduct a variety of accuracy checks. For example, we collected direct information on the number of characters per page (from a sample manuscript page) and also collected information on font size, line spacing, and margin sizes that can be used to estimate characters per page. Figure 2b in the online Appendix shows a scatterplot of the two measures, which have a correlation coefficient of 0.96 in both the *AER* and *JEEA* datasets. These reliability checks were also used to detect and correct errors in data coding. Table A1 in the online Appendix presents summary statistics for the datasets. Detailed instructions to the coders are available in the online Appendix available with this paper at http://e-jep.org.

Figure 2
Distribution of Manuscript Lengths at *AER* Before and After Page Limit Policy



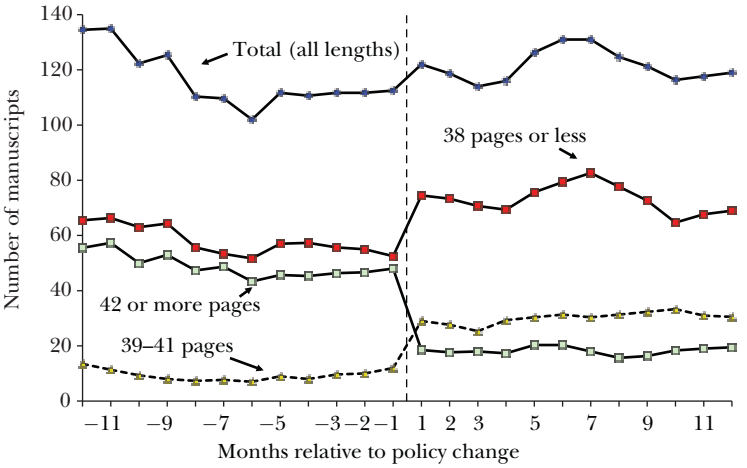
Notes: The sample includes all manuscripts submitted to the *AER* from July 1, 2007, to June 30, 2008 (the "pre" period with no page limits), and from October 1, 2008 to September 30, 2009 (the "post" period of page limits). Manuscript length includes the count of pages of text, tables, figures, and any appendices submitted for publication and excludes the cover page(s) and any online appendices.

with 44 percent of manuscripts longer than 40 pages, and 15 percent longer than 50 pages. In the after period, only 3 percent of papers are longer than 50 pages, indicating a small number of exceptions to the stated length policy. Interestingly, the number of 41–50 page manuscripts also declines very substantially, reflecting the fact that most authors adhered to the 40-page limit for 1.5-spaced manuscripts, rather than the 50-page limit for double-spaced papers. As expected, there is a large spike in post-policy submissions at 39–40 pages, representing 21 percent of submissions compared to just 6 percent in the pre-page limit period. This spike suggests that many authors reformatted or shortened their papers to submit to the *AER*.

A concern with the simple contrast in Figure 2 is that there could be underlying trends in submission behavior that confound the pre–post comparison. To address this concern, Figure 3 plots the number of submissions per month in different length categories before and after the introduction of the page limit policy. The total number of submissions per month, shown in the top line of the figure, is relatively stable over our two-year sample period at around 120 per month. There is no evidence of a decline in the total number of submissions after the introduction of page limits, as would be expected if a significant number of longer papers were diverted to other outlets. The introduction of page limits did lead to a sharp decrease in the number of submissions of 42 pages or longer, coupled with a rise in the number of papers around the limit (39–41 pages). Both of these shifts appear to have been fully realized by the first month of full implementation of the policy in October 2008.

Many of the changes in the distribution of page lengths documented in Figures 2 and 3 are attributable to changes in formatting of the papers, rather than

Figure 3
Number of *AER* Submissions by Month and Length



Notes: Manuscript counts are smoothed using 3-month moving average except at the policy change. Manuscript length includes pages of text, tables, figures, and appendices submitted for publication and excludes the cover pages(s) and any online appendices.

to reductions in the actual size of the submissions. Prior to the page limit policy, for example, about a third of the papers were double-spaced, whereas after the introduction of the policy the large majority of papers adopted 1.5-line spacing. An author with a double-spaced manuscript of 42 pages who simply switched to 1.5 spacing would end up submitting a 32 page paper. There is also a shift toward narrower margins after the page-length policy was adopted.[4]

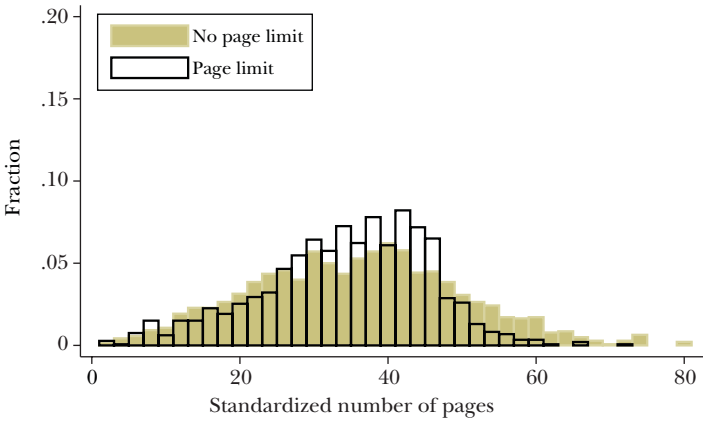
We summarize the impacts of these formatting changes using information on the number of characters per page. Prior to the page-length policy, the average number of characters per page was 2,250—a little over the expected density of a double-spaced 12-point font paper. After the policy, the modal density increased to 2,560 characters—close to the expected density of a 1.5-spaced document using 12-point font with no offset equations or major gaps. Interestingly, 13 percent of manuscripts submitted after the policy change have more than 3,000 characters/page, substantially above the limit that is technically permissible.

To facilitate comparisons of manuscripts before and after the policy change we construct a standardized measure of length based on the average number of characters per page in the document. Specifically, we define:

$$\text{standardized length} = \text{actual length} \times \text{characters per page}/2550.$$

[4] We document these changes in online Appendix Figures 4a–d, available with this paper at http://e-jep.org.

Figure 4
Distribution of Standardized Manuscript Lengths at *AER* Before and After Page Limit Policy



Notes: See note to Figure 2. In this figure, page lengths are standardized assuming 2,550 characters per page, which is the expected density assuming 12-point font, 1.5-line spacing, and 1-inch margins.

We assume a standardized page has 2,550 characters, which is the expected density of a page formatted at 1.5 spacing with 12-point font and 1-inch margins, assuming 91.5 characters per line times 31 lines per page, and allowing a 10 percent reduction for partial lines associated with paragraph endings and section headings. We then add the number of pages of tables and figures, plus any pages of appendix materials included for publication, plus any abstract or title pages to obtain the total standardized length of each submission.

Figure 4 shows the distribution of standardized page lengths before and after the policy change. Compared to the pre-period, the post-period has fewer longer papers and more papers of intermediate length (30–45 pages). There is also a small decrease in the fraction of papers between 20 and 25 pages, offset by a small rise in the fraction between 26 and 30 pages. In contrast to the distribution of unadjusted page lengths (in Figure 2), the spike at 40 pages is no longer visible in the standardized distribution, reflecting the wide variation in page densities among submissions of 39–40 pages.

Under the assumption that submission rates would trend together in the absence of any policy change, and that submissions of shorter papers are unaffected by the policy, a difference-in-differences approach allows us to infer the change in the number of longer papers caused by the policy. We use a threshold of 30 standardized pages to classify shorter ($\leq$30 pages) and longer ($>$30) papers. Among the 361 papers submitted in the post-period with an unadjusted length of 39–41 pages (that is, papers at the page-length spike) only five have a standardized length of 30 pages or less, so we believe the number of submissions with $\leq$30 standardized pages is unlikely to have been much affected by the page length policies.

Table 1

Difference in Differences Analysis of Impact of Page Limit Policy on Submissions per Week of Longer versus Shorter Manuscripts, *American Economic Review*

	Year before limits imposed (1)	*Year after limits imposed (2)*	*Difference: After − Before (3)*
Number of shorter manuscripts (30 pages or less)	10.21 (0.50)	10.02 (0.55)	–0.19 (0.74)
Number of longer manuscripts (31 pages or more)	16.79 (0.66)	18.10 (0.61)	1.31 (0.90)
Difference: Longer − Shorter	6.58 (0.73)	8.08 (0.76)	1.50 (1.05)

Notes: Standard errors are in parentheses. The sample includes 52 weekly observations for one year prior to the page limit policy (July 2007–June 2008) and 52 weekly observations for one year after (October 2008–September 2009). Manuscript lengths are measured using standardized page lengths. See text.

We estimate the following difference-in-differences specification

$$n_{L,t} - n_{S,t} = \alpha + \beta d_{PL,t} + \varepsilon_t,$$

where $n_{L,t}$ represents the number of longer submissions in week t; $n_{S,t}$ is the number of shorter submissions in the same week; $d_{PL,t}$ is an indicator equal to 1 for observations from the period during which page limits are in effect; and ε_t is an error term. The coefficient β measures the relative change in the number of longer versus shorter submissions in the post-period.

The underlying components of the difference in differences are presented in Table 1. Comparing columns 1 and 2, there was essentially no change in the number of shorter papers submitted per week following the adoption of the page limit policy by the *AER:* that is, the number of shorter papers submitted per week was 10.21 in the pre-period and 10.02 in the post-period. By comparison, the number of longer manuscripts submitted per week increased slightly, from 16.79 to 18.10, leading to a difference in differences of $\hat{\beta} = 1.50$ manuscripts per week, with a standard error of 1.05. Relative to the pre-policy submission rate of 16.8 per week, this implies an 8.9 percent *increase* in longer submissions (standard error = 6.3 percent).[5]

[5] A possible concern with these estimates is that the number of submissions per week is serially correlated, leading an ordinary least squares procedure to understate the standard error of the estimated difference in differences. In fact, the residuals from the equation are essentially uncorrelated (first order serial correlation = –0.01; second order correlation = –0.01) so quasi-differencing the data to remove serial correlation and re-estimating leads to an estimated coefficient and standard error that are essentially identical to the corresponding ordinary least squares estimates. We present results using the log of the number of submissions in online Appendix Table 2.

While this point estimate suggests that the *AER*'s page limit policy had no negative effect on the submission rate of longer papers, the sampling error means we can only rule out estimates of β smaller than -0.6 papers per week, a worst case loss of 2.2 percent of the total weekly submission flow. Our preferred estimate is that page limits caused no loss of papers at the *AER*.

Let us assume that authors who are considering submitting a paper to the *AER* have a certain fixed threshold of page length (for simplicity, we will assume the threshold is the same across authors) beyond which they are unwilling to consider shortening the paper and will instead submit their paper to another journal. In the original pre-period distribution of papers, the 97.8 percentile of page lengths was a paper of 65 pages. Given that the worst-case scenario involves losing 2.2 percent of submissions, it follows that authors are willing to cut their papers from 65 to 40 pages—a reduction of 25 pages—to submit to the *AER*.[6]

If there was any loss of longer papers, what types of papers were more likely to be diverted to other journals? One important characteristic is the quality of papers, which we measure by whether a paper receives a revise-and-resubmit decision. Figure 5A shows a nonparametric regression of the revise-and-resubmit indicator on the number of (standardized) pages among papers submitted prior to the page limit policy. The figure provides strong evidence that the quality of papers increases with the number of pages. Indeed, 20 percent of the 221 papers longer than 50 (normalized) pages received a revise-and-resubmit decision, compared to only 6.9 percent of the 1,185 papers shorter than 50 pages. Hence, any loss of longer papers could have negative impact on quality.

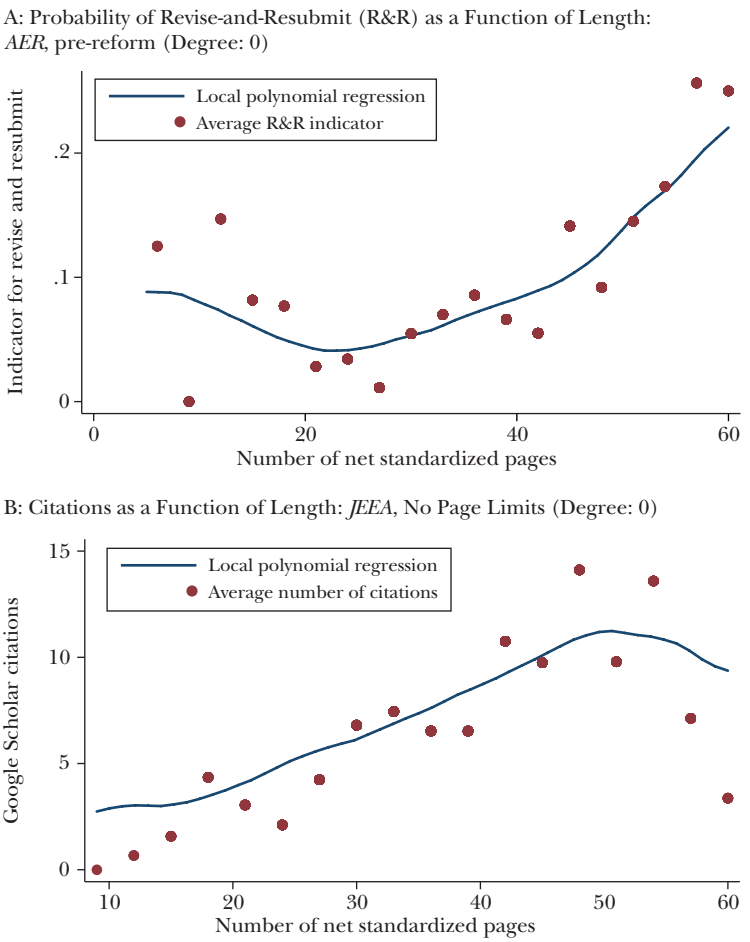
A second measure of quality of manuscripts is the number of citations. While we were unable to construct citation information for submissions to the *AER*, we obtained Google Scholar citations for all the manuscripts submitted to *JEEA*. Limiting attention to papers submitted to *JEEA* in the periods with no page limits, Figure 5B plots a local polynomial regression of the number of citations as a function of the normalized paper length. The number of citations is generally increasing in the length of the manuscript. The average citation count for the 87 papers longer than 50 (standardized) pages is 12.6, compared to 7.3 for the 636 papers shorter than 50 pages, confirming that longer manuscripts are on average more important contributions.

The positive correlation between pages and citations in Figure 5B is also present across published papers in top economics journals. In Card and DellaVigna (2013), we show that papers in the upper quintile of the length distribution published in

[6] We investigate the robustness of our estimate in a series of alternative specifications shown in online Appendix Table A2, available with this paper at http://e-jep.org. The results are very similar if we use a quasi-differenced specification in which we include the number of shorter papers submitted in a week as a control variable with an unrestricted coefficient (rather than assume a coefficient of 1.0 as in the baseline model) or if we add a linear trend to the quasi-differenced model, allowing different long-run trends in submission rates of shorter and longer papers. The results are more sensitive to shifting the threshold for defining shorter and longer papers from 30 to 35 standardized pages, which would suggest a modest loss of papers.

Figure 5

Relationship between Page Length and Quality Measures (*AER* and *JEEA*)



Notes: Figure 5A shows nonparametric regression of the probability of receiving a revise–and–resubmit verdict for manuscripts submitted to *AER* in the year before page limits were adopted, as a function of standardized manuscript length. Figure 5B shows nonparametric regression of the number of Google Scholar citations (as of July/August 2012) to manuscripts submitted to *JEEA* in the no–page–limit period.

the top five journals from 1970 to 2012 have *50 percent more citations* than those in the middle quintile, controlling for many other features of the papers. Similar specifications fit to articles from the *AER* alone imply 40 percent higher citations for papers in the top versus the middle of the length distribution. (Both these exercises excluded notes and replies, as well as the annual Papers and Proceedings issue of the *AER*).

We also examined the impact of the page limit policy across different fields. Specifically, we used the (up to three) *JEL* codes provided by authors to count the number

of submissions in eight subfields: theory; laboratory experiments; macroeconomics (including international macro); applied microeconomics (public economics, labor, micro-development, and law and economics); industrial organization; econometrics; finance; and all other fields.[7] This analysis does not include the *25 percent* of submissions that do not report a *JEL* code. (Papers can be counted in up to three of these categories, depending on the number of *JEL* codes provided). We then conducted a difference-in-differences analysis similar to Table 1 by subfield (with details in online Appendix Table 4). Across all fields but macro, there was an increase in overall submissions to the *AER* in the year after the page-length policy was introduced. In all fields except applied microeconomics and econometrics, the increase in the submission rate of longer papers was greater than for shorter papers. Interestingly, there was a relatively large loss of longer papers (a decline of 7 percent) in empirical micro—the field that had typically submitted the largest fraction of long papers prior to the policy. We cannot reject the hypothesis, however, that the difference in differences for empirical micro is the same as for any of the other fields.[8]

Benefits of Page Limits?

The primary motivation for the page limit policy at the *AER* was to shorten the length of published papers. In the first panel of Table 2, we compare the average length (excluding online appendices) of all first submissions in the year before and after the page limit introduction. The first row shows that there was a significant 3.9 page reduction in the average lengths of submitted manuscripts after the policy in the post period. Taking account of formatting changes by examining standardized page lengths (second row), however, the reduction in page lengths is only 1.6 pages. The difference is a little bigger (−4.2 pages) in a regression-adjusted comparison that controls for editor and field fixed effects. The last column in Table 2 reports the difference after accounting for such controls.

Which portions of a paper changed in length? The next rows in the upper panel show that the page limit policy did not lead to any shortening in the number of pages of text, but led to a half-page reduction in the number of pages of tables and figures and a 1.5 page reduction in the length of appendices intended for publication. The latter was offset by a 2.4 page increase in the length of online appendices (that is, appendices not intended for publication and not counted in our measure of page length).

Overall the page limit policy led to a modest shortening of submitted papers. What about the papers that were ultimately accepted for publication? The second panel of Table 2 focuses on the final accepted versions of papers that were originally

[7] We assign the fields using JEL codes as follows: theory = D, C7; laboratory experiments = C9; macroeconomics = E, F, I, O4, O5, O11; applied micro = H, I, J, K, and O except O4, O5, O11; IO = L; econometrics = C, except C7 and C9; finance = G.

[8] We also considered whether the page limit policy at the *AER* led people to write shorter working papers in the anticipation of having to comply with the policy. Using measures of the length of NBER working papers written between 2007 and 2012, we find no evidence of a shift in length following the page-length policy. Details are in online Appendix Figure 7, available with this paper at http://e-jep.org.

Table 2
Impact of Page Limits on Manuscript Length—*AER*

	No policy in place (1)	*Policy in place (2)*	*Difference (3)*	*Regression-adjusted difference (4)*
All first submissions:				
Mean number of pages	37.8	33.9	−3.9 (.5)***	−4.2 (.5)***
Mean number of pages, standardized	35.5	33.9	−1.6 (.5)***	−2.2 (.5)***
Mean number of pages of text, standardized	25.6	26.1	0.5 (.4)	0.0 (.4)
Mean number of pages of tables and figures	4.8	4.2	−0.5 (.2)**	−0.5 (.2)**
Mean number of pages of appendix	4.0	2.5	−1.5 (.2)***	−1.6 (.2)***
Mean number of pages of online appendix	0.6	3.0	2.4 (.2)***	2.6 (.2)***
Sample size	1,406	1,462		
Final accepted version of papers invited for revision[a]				
Mean number of pages	40.6	39.4	−1.2 (1.9)	−2.2 (2.4)
Mean number of pages, standardized	40.5	39.8	−0.7 (2.0)	−1.9 (2.4)
Change in number of pages, standardized, from 1st submission	−1.2	-0.8	0.4 (2.0)	0.1 (2.6)
Number of rounds of revision	1.7	1.9	0.19 (0.9)**	−0.04 (.10)
Sample size	114	110		
Final accepted manuscripts, expanded sample[b]				
Mean number of pages	40.8	39.1	−1.6 (1.4)	−1.3 (2.0)
Mean number of pages, standardized	40.2	39.9	−0.3 (1.4)	−0.7 (2.0)
Change in number of pages, standardized, from 1st submission	−2.2	0.3	2.5 (1.4)*	3.3 (2.1)
Sample size	218	211		

Notes: Entries in the final column are regression-adjusted with controls for editor/coeditor and field of submission. And see notes to Table 1.
[a] Includes papers accepted in first round without revision.
[b] See text for description of expanded sample.
*, **, and *** indicate statistical significance at 10 percent, 5 percent, or 1 percent.

submitted during the sample window of plus-or-minus one-year from the policy change.[9] In this sample of 224 papers, we find a very small and statistically insignificant reduction in the standardized length of accepted papers (–0.7 pages, standard error = 2.0).

To check the robustness of this conclusion, in the last panel of Table 2, we extend the sample of accepted papers by including submissions that received an

[9] This sample does not include all papers that received a "revise and resubmit" verdict because about 10 percent of these papers were ultimately rejected and another 10 percent were still under review as of June 2012.

initial revise-and-resubmit verdict between January 2006 and June 2007 and were accepted for publication between July 2007 and June 2012. (Papers accepted before July 2007 are not included in the *AER*'s Manuscript Central database and cannot be tracked.) We also include submissions from October 2009 to September 2010 that received a revise-and-resubmit and were accepted by June 2012. In this extended sample of 429 papers, we confirm the main finding that the (standardized) length of published papers remained nearly unchanged after the introduction of page limits, with an estimated shortening of only 0.3 pages, or 0.7 pages after regression adjustment.

One of the reasons for the smaller effect on accepted papers is that while the revision process led to a shortening of papers in the pre-policy period (−2.2 standardized pages), this tendency was actually reversed in the post-policy period, leading to a *lengthening* of papers during the revision process (+0.3 standardized pages). (The pattern is qualitatively similar but muted in the smaller sample shown in the second panel). This reversal suggests that referees and editors were previously asking for papers to be cut, but under the new policy were asking for additional material to be added to the manuscript.

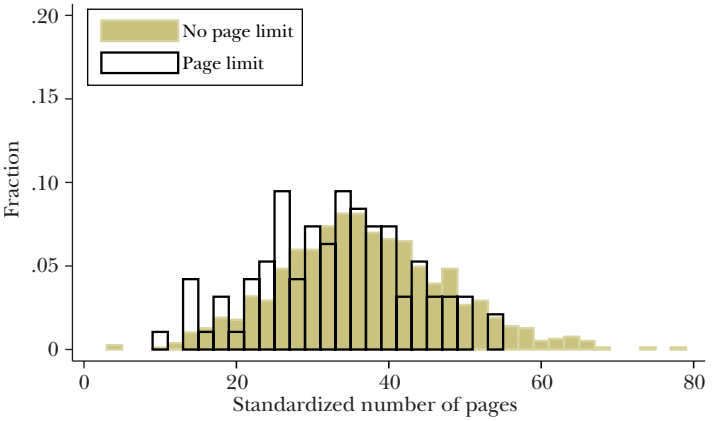
To put these changes in length in perspective, recall from Figure 1 that the typical paper published in the top five journals increased in length by about 30 pages from 1970 to 2012, with about a 12 page increase between 2002 and 2012. Relative to these trends the estimated impact of the page limit policy is small. Figure 1 also shows the ratio of the median length of papers in the *American Economic Review* to the average of papers in the other top-five journals. This ratio has remained between 80 percent and 100 percent over the period from 2005 to 2012, with no trend. We conclude that the page limit policy did *not* have much effect on the length of published papers in the journal.

Finally, it is possible that, while not affecting the length of accepted papers by much, the page limit made the revision process faster. In the middle panel of Table 2 we show the number of rounds of revision required for accepted papers in the baseline sample. We find no evidence of a speed up. Indeed, the number of rounds actually rose slightly, but after adjusting for editor fixed effects, the change is very close to zero.

Page Limits at the *Journal of the European Economic Association*

Until March 2009, the *Journal of the European Economic Association* had no page length restrictions on new submissions. On March 25, 2009, the editors introduced a page limit policy modeled on the *AER* policy: that is, 1.5-line spacing, a 40-page limit (including figures, tables, and references), and 12-point font. Within two months after the introduction of the policy, the editorial team became concerned that the page limit policy was hurting the number of submissions. Following an internal debate, enforcement of the page-length policy halted on July 8, 2009, and the policy was removed from the *JEEA* website on August 5, 2009.

Figure 6
Distribution of Standardized Manuscript Lengths at *JEEA* in Periods With and Without Page Limit Policy



Note: Page lengths are standardized assuming 2,550 characters per page, which is the expected density assuming 12-point font, 1.5-line spacing, and 1-inch margins.

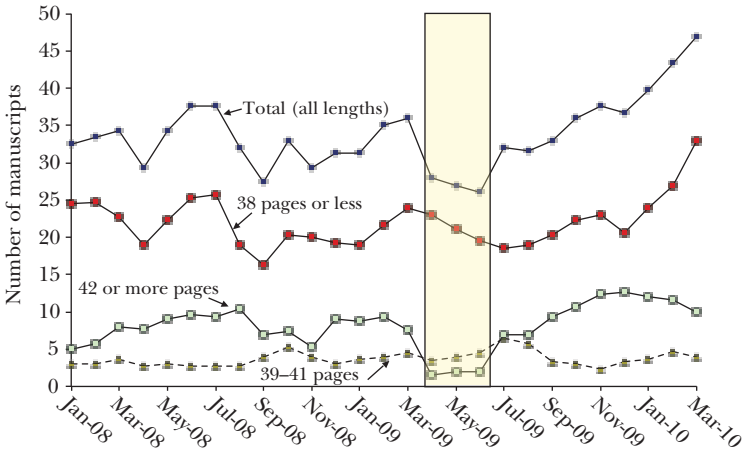
We collected data from electronic records of submissions to the *JEEA* for a period of 15 months before the introduction of the policy (January 1, 2008, to March 24, 2009); for the 3.5 months period during which the page limit policy was in place (March 25, 2009, to July 8, 2009); and for approximately nine months following the official removal of the page limit (August 5, 2009, to March 31, 2010). As in our analysis of the *AER* policy change, we omit from most of the analysis the transitional period, in this case from July 8, 2009 to August 5, 2009.[10] Our sample contains 883 newly submitted manuscripts, of which 95 were submitted while page limits were in effect.

We compare papers submitted during the page limit period to those submitted before or after. As at the *AEA*, the page-length policy at *JEEA* coincided with the introduction of new guidelines for manuscript formatting. Using the same procedures we developed for the *AER*, we estimate standardized page lengths for all submissions to the *JEEA*. The distributions of standardized page lengths with and without the page limit policy are shown in Figure 6. In contrast to the corresponding figure for the *AER* (Figure 4), there is no rise in the density of papers around the page limit.

Figure 7 presents time series evidence on the number of monthly submissions for *JEEA* before, during, and after the page-length policy was in effect. Several features of this graph are quite different from the corresponding graph for the *AER*. Most importantly, the drop in submissions of longer papers in the policy period of

[10] We include the transitional period only in the time-series graph in Figure 7.

Figure 7
Number of *JEEA* Submissions by Month and Length



Notes: Manuscript counts are smoothed using 3-month moving average except at the policy change. Manuscript length includes pages of text, tables, figures, and appendices submitted for publication and excludes the cover page(s) and any online appendices.

April–June 2009 is matched by a parallel drop in total submissions. Taken together with the fact that there was no increase in the fraction of papers close to the page limit, this pattern suggests that *JEEA* authors responded to the page limit policy by diverting most of the papers that were above the page limit to other journals. After the removal of the page limit policy, submissions appear to have quickly returned to their pre-page-limit rates.

We again carried out a simple difference-in-differences analysis, splitting papers into shorter papers and longer papers using a threshold of 30 standardized pages. The results, presented in Table 3, show that during the page limit period, *JEEA* received slightly more shorter submissions (2.6 per week versus 2.4) but significantly fewer longer submissions (3.7 per week versus 5.7). The difference-in-differences estimator implies a loss of 2.15 longer manuscripts per week, which is statistically different from 0.[11] The implied loss of 2.15 papers per week represents a 38 percent reduction in the inflow of longer papers.[12] Interpreted though the lens of a version of our simple model in which all authors have the same surplus from submitting to *JEEA*, the loss of 2.15 longer manuscripts per week suggests that nearly all authors of longer manuscripts were unwilling to shorten their paper, and preferred instead to submit elsewhere.

[11] The residuals from this weekly regression model are very slightly positively correlated (first order correlation = 0.049). Quasi-differencing the data and re-estimating we obtain a point estimate of −2.16, with an estimated standard error of 0.94.

[12] Online Appendix Tables 3 and 5, available with this paper at http://e-jep.org, present a series of alternative specifications for the difference-in-differences model that probe the robustness of this estimate.

Table 3

Difference-in-Differences Analysis of Impact of Page Limit Policy on Submissions per Week of Longer versus Shorter Manuscripts, *Journal of the European Economic Association*

	Period with no page limits *(1)*	*Period with page limits* *(2)*	*Difference: After − Before* *(3)*
Number of shorter manuscripts (30 pages or less)	2.38 (0.18)	2.60 (0.34)	0.22 (0.49)
Number of longer manuscripts (31 pages or more)	5.66 (0.31)	3.73 (0.53)	−1.93 (0.81)
Difference: Longer − Shorter	3.29 (0.33)	1.13 (0.70)	−2.15 (0.89)
Number of weeks	**98**	**15**	**113**

Notes: Standard errors are in parentheses. Sample includes weekly data for 64 weeks prior to page limit policy (January 1, 2008, to March 24, 2009), 15 weeks while page limit policy was in effect (March 25, 2009, to July 8, 2009), and 34 weeks after page limit policy was removed (August 2009–March 2010). Manuscript lengths are measured using standardized page lengths. See text.

Concluding Thoughts

How do authors respond to page limits on new submissions? Our analysis shows that the answer depends on the competitive position of the journal that imposes the limit. Although the policies adopted by the *Journal of the European Economic Association* and the *American Economic Review* were identical, authors responded to the *JEEA* limit by diverting all or nearly all of their longer manuscripts to other outlets, whereas authors responded to the *AER* by reformatting and shortening their manuscripts. A simple model of author behavior suggests that these very different responses can be explained by differences in the perceived surplus associated with a publication at *JEEA* versus the *AER*. *JEEA*—a respected but relatively new journal—faces substantial competitive constraints on its policy choices. In contrast the *AER*—a top journal with a reputation built over a century—has market power and can raise the cost of submission with little or no loss of supply.

We suspect that these findings also apply to other dimensions of journal policy. Specifically, we conjecture that journals below the top five face a relatively elastic manuscript supply and can raise or lower submission rates by systematically varying the expected time to first decision or the quality of refereeing. While we have no direct evidence on this response, it is anecdotally consistent with a near-doubling in submissions to *JEEA* from 2009 to 2012 following a sharp reduction in the average time-to-first-decision. It is also consistent with the successful emergence of the four *AEJ* journals from the American Economic Association which offer a generally fast turn-around to submissions. In contrast, data presented by Ellison (2002) show

a relatively weak relationship between changes in submission rates and trends in time-to-first-decision at the top five journals over the 1980s and 1990s.

Even at the *AER*, where the page limit policy led to little or no loss in submissions, it also appears to have had few benefits. In particular, despite the goal of reducing the length of published papers, the average length of accepted manuscripts was not significantly impacted. Arguably, a policy that forces hundreds of authors each year to spend time shortening papers without any obvious benefits should be reconsidered.

More generally this paper highlights the importance of evaluating editorial policy choices. Such evaluation is rare in economics, with the notable exception of Blank's (1991) evaluation of double-blind refereeing, the evaluation by Brogaard, Engelberg, and Parsons (2014) of conflict of interest rules, and the study by Chetty, Saez, and Sandor in this issue on referee deadlines and incentives.

Returning to the issue of manuscript length, there may be a legitimate concern about the trend shown in Figure 1 toward ever-longer papers in the top five journals. Perhaps a more aggressive page-length policy, such as a 30-page limit, could shorten the length of published papers. We suspect that such a policy would come at some cost, even to a top journal like the *AER*. More importantly, since journals are very concerned about citations and longer articles garner more citations, it is unclear whether the goal of reducing page lengths is even justified.

One could argue that the observed correlation between citations and article length is driven by the willingness of editors to allow extra space to papers they believe will be influential. But evidence such as we have assembled in Figures 5A and 5B on probability of acceptance, and on citation rates for all submissions (regardless of publication status), support the view that longer papers are "better" papers. Of course, it may be possible to preserve the quality of the longer papers while shortening them at the margin, though we are unaware of any causal evidence on the benefits (or costs) of shortening a given paper.

Instead of seeking to place limits on papers in existing journals, perhaps the economics profession would benefit from an alternative journal with a focus on shorter papers. There may be an interesting parallel in the field of social psychology. The top journal in this field, the *Journal of Personality and Social Psychology*, publishes relatively long articles, as do other influential journals in the discipline. In 1988, however, a new journal, *Psychological Science,* was created to mirror the format of *Science*. Research papers submitted to *Psychological Science* can be no longer than 4,000 words. By comparison, this paper, which is not long by economics standards, is around 6,500 words long. *Psychological Science* has quickly emerged as a leading journal in its area. In social psychology, journals publishing longer articles coexist with journals specializing in shorter, high-impact articles.

We speculate that authors who wanted to submit to an *Economic Science* journal with very tight page limits would adopt a different form of writing, with less space devoted to expansive introductions, model development, and specification testing. Editors and referees of such a journal would also have to adopt different standards, placing more weight on actual findings and less on the framing and interpretation

of the results. Whether publications in *Economic Science* would be as highly valued by the economics profession as those currently published in the top five journals is an open question.

■ *We are very grateful to Penny Goldberg, Robert Moffitt, Stephanie Raimander, Steve Stelling, and Fabrizio Zilibotti for their assistance in this project. We also thank Laura Ristau for her outstanding work in coding thousands of manuscripts in the AEA office, a team of undergraduates at Berkeley (Noah Dreymann, Kaiji Gong, Sam Johnson, Ki Sung Kim, Wonsoon Kim, Sunny Lee, Seongjoo Min, Zi Peng, Tim Semenov, Eileen Tipoe, Brian Wheaton, and Janice Zhou), as well as Alex He, Xuan Li, and Jeff Sorenson for excellent research assistance. We thank Esther Duflo, Liran Einav, Penny Goldberg, Hilary Hoynes, Larry Samuelson, Ivo Welch, Justin Wolfers, and audiences at the 2012 NBER Summer Institute (Labor Studies), at Harvard University, at Mannheim University, and at UC, Santa Cruz for useful comments.*

References

Blank, Rebecca M. 1991. "The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from *The American Economic Review*." *American Economic Review* 81(5): 1041–67.

Brogaard, Jonathan, Joseph Engelberg, and Christopher Parsons. 2014. "Networks and Productivity: Causal Evidence from Editor Rotations." *Journal of Financial Economics* 111(1): 251–70.

Card, David, and Stefano DellaVigna. 2012. "Revealed Preferences for Journals: Evidence from Page Limits." *NBER Working Paper* 18663.

Card, David, and Stefano DellaVigna. 2013. "Nine Facts about Top Journals in Economics." *Journal of Economic Literature* 51(1): 144–61.

Ellison, Glenn. 2002. "The Slowdown of the Economics Publishing Process." *Journal of Political Economy* 110(5): 947–93.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 169–188

What Policies Increase Prosocial Behavior? An Experiment with Referees at the *Journal of Public Economics*[†]

Raj Chetty, Emmanuel Saez, and László Sándor

The peer review process familiar to all academic researchers offers a classic example of the positive externalities from prosocial behavior: the reviewer bears the costs from submitting a high-quality referee report quickly, while the gains to the authors of the paper and to society from the knowledge produced are potentially large. We evaluate the impacts of economic and social incentives on peer review using an experiment with 1,500 referees at the *Journal of Public Economics*. The specific aim of the experiment is to understand how to improve the speed and quality of peer review, an issue of particular importance to the economics profession given the slowdown of the publishing process (Ellison 2002). Our broader objective is to evaluate commonly used methods of increasing prosocial behavior and to test the predictions of competing theories.

In our experiment, we randomly assign referees to four groups: a control group with a six-week (45 day) deadline to submit a referee report, a group with a four-week (28 day) deadline, a cash incentive group rewarded with $100 for meeting a four-week deadline, and a social incentive group in which referees were told that their turnaround times would be publicly posted. The experiment yields four sets of results.

First, shortening the deadline from six weeks to four weeks reduces median review times from 48 days to 36 days. Because missing the deadline has no direct

■ *Raj Chetty is Professor of Economics, Harvard University, Cambridge, Massachusetts. Emmanuel Saez is Professor of Economics at the University of California, Berkeley, California. László Sándor is a PhD candidate in Economics, Harvard University, Cambridge, Massachusetts. Their email addresses are chetty@fas.harvard.edu, saez@econ.berkeley.edu, and sandor@fas.harvard.edu.*

[†] To access the Appendix, data Appendix, and disclosure statements, visit http://dx.doi.org/10.1257/jep.28.3.169

doi=10.1257/jep.28.3.169

consequence, we believe the shorter deadline acts primarily as a "nudge" (Thaler and Sunstein 2008) that changes the default date at which referees submit reports. Second, providing a $100 cash incentive for submitting a report within four weeks reduces median review times by an additional eight days. Third, the social incentive treatment reduces median review times by approximately 2.5 days—which is intriguing given that the degree of social pressure applied here is relatively light. We also find that social incentives have much larger effects on tenured professors, but in contrast, tenured professors are less sensitive to deadlines and cash incentives than untenured referees.

Finally, we evaluate whether the treatments have an impact on other outcomes besides review time.[1] Economic models of multitasking (for example, Holmstrom and Milgrom 1991) predict that referees will prioritize the incentivized task (in this application, submitting a report quickly) at the expense of other aspects of performance (in this case, potentially the quality of reviews). We find that the shorter deadline has no effect on the quality of the reports that referees submit, as measured by whether the editor follows the recommendation of the referee or the length of referee reports. The cash and social incentives induce referees to write slightly shorter referee reports, but do not affect the probability that the editor follows the referee's advice. We also find little evidence of negative spillovers across journals: the treatments have no detectable effects on referees' willingness to review manuscripts and review times at other Elsevier journals.

We conclude that small changes in journals' policies could substantially improve the peer review process at little cost. Shorter deadlines appear to be an essentially costless means of expediting reviews. Cash and social incentives are also effective, but have monetary and psychic costs that must be weighed against their benefits.

A large body of evidence from the lab has considered the determinants of prosocial behavior and altruism (for example, Ledyard 1995; Fehr and Fischbacher 2003; Vesterlund forthcoming). Our study provides evidence from the field, which has been considerably more limited. Prior work concerning prosocial behavior has often debated whether extrinsic incentives such as cash payments are effective in increasing prosocial behavior because they may crowd out intrinsic motivation (Titmuss 1971; Bénabou and Tirole 2006). In our application, if referees submit reviews to be recognized by editors for their service to the profession, the provision of monetary incentives could potentially erode this signal and have a negative impact on review times. However, our analysis shows that, at least in this context, price incentives, nudges, and social pressure are all effective and complementary methods of increasing prosocial behavior.

[1] The cash incentive increases the fraction of referees who agree to review a manuscript. The social incentive reduces agreement rates, while the shorter deadline has no impact. We show that the selection effects induced by these changes in agreement rates are modest and are unlikely to explain the observed changes in review times.

Table 1
Description of Treatment Groups

	Group			
	6 Week	*Social*	*4 Week*	*Cash*
Deadline	6 weeks (45 days)	6 weeks (45 days)	4 weeks (28 days)	4 weeks (28 days)
Incentives	None	Review time posted online at end of year	None	$100 Amazon gift card if deadline met
Duration of intervention	Feb. 15, 2010 to Oct. 26, 2011	Feb. 15, 2010 to Oct. 26, 2011	Feb. 15, 2010 to Oct. 26, 2011	Feb. 15, 2010 to May 9, 2011

Notes: This table describes the four treatment groups to which referees were randomly assigned. Every referee was assigned permanently to one group; referees never changed groups. Referees were notified about the conditions of the review request upon invitation and were sent a reminder one week before the deadline. Examples of these invitation and reminder emails are shown in online Appendices A and B. Cash incentives were stopped for invitations after May 9, 2011; after that point, referees assigned to the cash incentive group simply faced a four-week deadline, with no incentives. The other treatments were implemented without any changes for the full duration of the experiment, from February 15, 2010 to October 26, 2011.

Experimental Design

We conducted the experiment over a 20-month period, from February 15, 2010, to October 26, 2011. All referees for the *Journal of Public Economics* during this period were randomly assigned to one of four groups. These assignments were permanent for the duration of the experiment: referees never switched groups. The coeditors in charge of handling each new submission chose referees to review the paper without seeing the group to which the referee was assigned.

Some key features of the four groups are shown in Table 1.[2] All deadlines for the differing groups were defined relative to the date at which the invitation was sent—not the date at which the referee accepted the invitation—to eliminate incentives to delay agreement.

The *control* or what we will refer to as the *six-week* group actually faced a 45-day deadline for submitting a referee report, the deadline that was in place at the journal before the experiment began. The deadline was described using the

[2] An online Appendix available with this paper at http://e-jep.org includes the details of the experiment. Appendix Figure 1 presents a flow chart for the entire experiment. Appendix A shows our invitation emails. Online Appendix B shows our reminder and thank-you emails. Appendix C includes more detail on data sources and variable definitions. Appendix Table 1 presents summary statistics for the primary experimental period (referee invitations between February 15, 2010 and May 9, 2011). Appendix D describes the reweighting methodology behind Figure 2B. Appendix E presents the hazard model estimates of treatment effects on review times. Appendix F provides a list of other journals used to assess spillover effects. Appendix G presents a summary of all the appendix tables and figures. A de-identified version of the 3,397 observation dataset is available at the *JEP* website: http://e-jep.org.

following language in the invitation letter: "If you accept this invitation, I would be very grateful if you would return your review on or before **July 21, 2010** (6 weeks from now)."

The *four-week* group faced a 28-day deadline for submitting a report. The email they received was identical to that sent to referees in the control group, except for the due date.

The *cash incentive* group faced a 28-day deadline and received a $100 Amazon gift card for submitting a report before the deadline. In addition to the standard text describing the deadline, the invitation letters in the cash incentive group included the following text: "As a token of appreciation for timely reviews, you will receive a **$100** Amazon.com® Gift Card if you submit your report on or before the due date. The Journal of Public Economics will automatically email you a gift card code within a day after we get your report (no paperwork required)."

Finally, the *social incentive* group faced a six-week (45 day) deadline and was told that referee times would be publicly posted by name at the end of the calendar year. In addition to the standard text describing the deadline, the invitation letters in the social incentive group included the following text: "In the interest of improving transparency and efficiency in the review process, Elsevier will **publish referee times by referee name**, as currently done by the Journal of Financial Economics at this website. The referee times for reports received in 2010 will be posted on the Journal of Public Economics website in January 2011. Note that referee anonymity will be preserved as authors only know the total time from submission to decision (and not individual referee's times)."

One week prior to their deadlines, referees who had not yet submitted reports received emails reminding them that their reports were due in a week. For the social incentive and cash incentive groups, these emails included language reminding referees of the treatments they faced. We also sent overdue reminders five days, 19 days, and 33 days after the due date. Referees in the cash, four-week, and six-week groups were simply informed their reports were past due. Referees in the social incentive group were again reminded that their referee times would be publicly posted. After the referees submitted reports, they received a thank you email. Referees in the cash incentive group received an Amazon gift card code in this thank you email if they submitted before the 28-day deadline. Those in the social incentive group received information on the number of days it took for them to submit the report.

To study the impact of monetary payments on intrinsic motivation after cash incentives are withdrawn, we stopped cash payments on May 9, 2011, roughly six months before we ended the other treatments. Referees in the cash incentive group continued to face a four-week deadline after this point, and received the same invitation and reminder emails as those in the four-week group. All other treatments continued until the end of the experiment on October 26, 2011, at which point all referees were reverted back to the six-week (45-day) deadline.

We analyze the effects of the experiment using information from two sources. We obtain information on referee assignments, review times, and other related outcomes at the *Journal of Public Economics,* as well as other Elsevier journals from

Elsevier's editorial database. We obtain information on referee characteristics—an indicator for holding an academic position, tenure status, gender, and an indicator for working in the United States—from curricula vitae posted online.

Each observation in our analysis dataset corresponds to a single referee invitation sent between February 15, 2010, and October 26, 2011. During this period, 3,397 invitations were sent out to 2,061 distinct referees. We include all observations in the referee report–level dataset in our analysis, so referees who are invited multiple times contribute multiple observations.

In our baseline analysis, we restrict attention to referee invitations sent between February 15, 2010, and May 9, 2011, the period when the cash reward was offered. We term this period the *primary* experimental period. During this period, we sent 2,423 invitations, of which 66.2 percent were accepted. Among these referees, 93.7 percent submitted a report before the editor made a decision. The median turnaround time for those who submitted reports was 41.0 days. Among the 1,157 referees who agreed to review a manuscript during the primary experimental period, 74.9 percent of referees agreed to review one manuscript during the experiment, 16.4 percent agreed to review two manuscripts, and the rest agreed to review three or more manuscripts.

To verify the validity of our experimental design, we also calculated these summary statistics by treatment group for referee assignments from November 1, 2005, to February 15, 2010, *before* the experiment began. As expected, given randomization, we find no statistically significant differences across the control group or the three treatment groups in these predetermined characteristics (details in online Appendix Table 2a). Hence, differences in performance across the four groups during the experimental period can be interpreted as causal effects of the treatments.

Four Sets of Outcomes

We analyze four sets of outcomes: 1) agreement to submit a review, 2) time taken to submit the review, 3) report quality, and 4) performance at other journals.

Outcome 1: Acceptance of Referee Invitation

Table 2 shows the percentage of referee invitations accepted by treatment group. We structure this and all subsequent tables as follows. The four columns correspond to the four experimental groups: six-week, social, four-week, and cash. For each group, we report the point estimate and associated standard error in parentheses. We cluster standard errors by referee to account for the fact that some referees review multiple papers. We also report p-values for the null hypothesis that agreement rates are the same in each treatment group and its corresponding control group. For the social incentive and four-week deadline groups, the control group is defined as the six-week deadline group. For the cash incentive group, the control group is defined as the four-week deadline group, which is the relevant comparison because the cash incentive group also faced a four- week deadline.

Table 2
Fraction of Referees Who Accept Review Invitation by Treatment Group

	Group:			
	6 Week	*Social*	*4 Week*	*Cash*
Percent who accept invitation	67.6%	61.1%	64.1%	72.0%
	(2.14)	(2.43)	(2.23)	(2.17)
p-value for equality with control		0.045	0.252	0.010
Observations	639	568	626	590

Notes: This table shows the percentage of referees who accept invitations to review in each treatment group. We restrict the sample to invitations sent between February 15, 2010, and May 9, 2011, the time period when the cash reward was offered. Standard errors, clustered at the referee level, are reported in parentheses. We also report *p*-values for the null hypothesis that agreement rates are the same in each treatment group and its corresponding control group. For the social and four-week groups, the control group is defined as the six-week deadline group. For the cash incentive group, the control group is defined as the four-week deadline group, which is the relevant comparison because the cash incentive group also faced a four-week deadline. The number of observations (referee report invitations) is reported in the last row.

Table 2 shows that 67.6 percent of the referee invitations are accepted in the six-week group. The acceptance rate is slightly lower at 61.1 percent in the social incentive group, a difference that is marginally statistically significant ($p = 0.045$). The acceptance rate in the four-week deadline group is 64.1 percent, not significantly different from the acceptance rate in the six-week group. Lastly, the acceptance rate in the cash incentive group is 72.0 percent, which is significantly higher than the acceptance rate in the four-week deadline group ($p = 0.010$).

Consistent with this statistical evidence, the journal received a few emails showing that the treatments influenced the decisions by some referees to review papers. For example, a referee assigned to the social incentive group wrote, "I was surprised to receive an email stating the journal is posting referee times by names. . . . I would like to withdraw my agreement to referee this paper. Sorry about that. I would have been happy to send in a report on time under a different policy." Other referees' emails explain why cash incentives increase acceptance rates. For instance, a referee in the control group wrote, "I am sorry to have to decline this 'invitation' to work for free . . . Can't Elsevier offer a better reward for the time they ask to devote to this screening?"

Overall, our results on acceptance rates allay the concern that pushing referees to submit reviews quickly will make it difficult to find referees who are willing to submit reviews.

Outcome 2: Review Time

We now turn to the central outcome our treatments were designed to change: the time that referees take to submit their reviews. Naturally, we can only observe review times for referees who agree to submit reviews. Because the referees who

accept invitations may differ across the treatment groups, differences in review times across groups reflect a combination of selection effects (changes in the composition of referees) and behavioral responses (changes in a given referee's behavior). For instance, referees who expect to be unable to submit a review quickly might be less likely to agree to review a paper under the shorter four-week deadline. This would reduce average review times in the four-week group via a selection effect even if referee behavior did not change.

Distinguishing between selection and changes in behavior is not critical for a journal editor seeking to reduce average review times, because it does not matter whether improvements come from getting faster referees or inducing a given set of referees to work faster. For the broader objective of learning about how incentives affect prosocial behavior, however, it is important to separate selection from behavioral responses. We therefore begin by assessing selection and then present estimates of treatment effects on review times both with and without adjustments for selection.

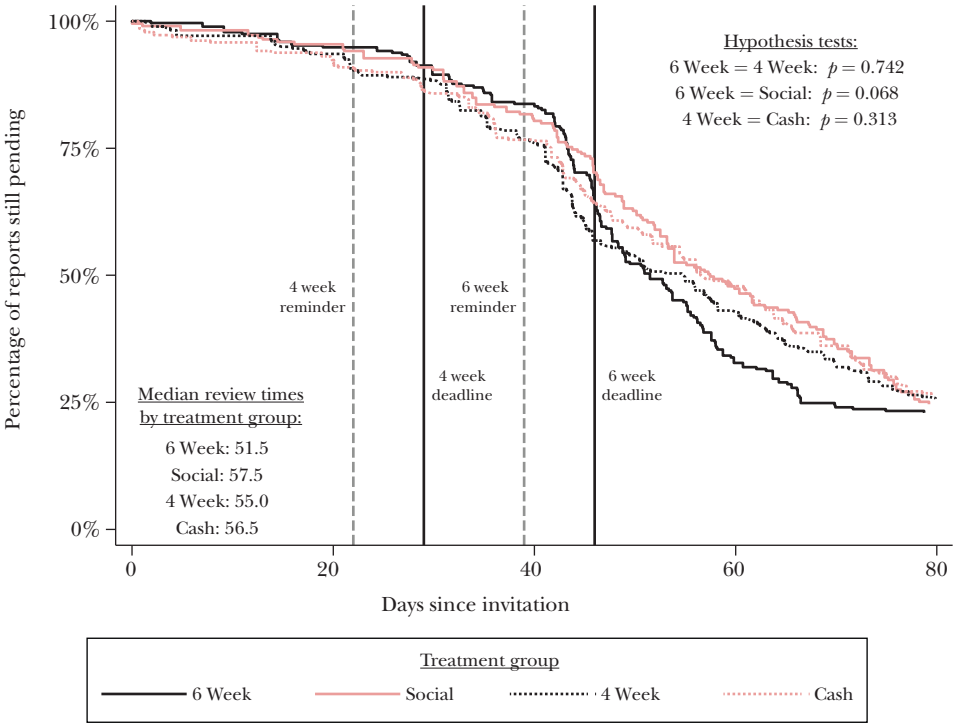
We evaluate the magnitude of selection effects in two ways. First, we compare predetermined referee characteristics, such as tenure status and nationality, across the four groups. We find that these characteristics are generally quite similar across referees who accept invitations in the four groups (details available in online Appendix Table 2b).

Second, we compare the *pre-experiment* review times of referees who agreed to review papers in each of the four experimental groups. For this analysis, we focus on the 67 percent of referees in our primary experimental sample who reviewed a manuscript for the journal before the experiment began (from November 2005 to February 15, 2010). All of these pre-experiment reviews were subject to a six-week deadline. Figure 1 plots survival curves for review times according to the treatment group to which the referees were later assigned, using data from the most recent review before the experiment began. These survival curves show the fraction of reviews that are still pending after a given number of days.[3]

The survival curves in the cash, four-week deadline, and six-week deadline groups are all very similar. Referees who agreed to submit a review under a shorter deadline or cash incentive treatment are no faster than those in the control group based on historical data. Nonparametric (Wilcoxon) tests for equality of the survival curves uncover no differences in review times across these three groups. We find marginally significant evidence ($p = 0.068$) that referees who agree to review papers in the social incentive group are slightly *slower* than those in the six-week control group. Hence, if anything, the social incentive treatment appears to induce slightly unfavorable selection in terms of referee speed. One explanation may be that diligent referees tend to be more concerned and anxious about their reputation and are hence less likely to accept the invitation with the social treatment. Overall, this evidence indicates that selection effects are modest and that differences in outcomes

[3] We include reviewers who do not submit reviews in these and all subsequent survival curves by censoring their spells at the point when editors make a decision on the paper.

Figure 1
Pre-Experiment Review Times for Referees Who Accept Invitations during Experiment, by Treatment Group



Notes: This figure plots survival curves that show the distribution of *pre-experiment* review times by treatment group. The sample consists of referees who accepted invitations between February 15, 2010, and May 9, 2011, the period when the cash reward was offered. Among these referees, 67.3 percent accepted a review invitation before the experiment began (from November 2005 to February 15, 2010); we use their data to construct this figure. For referees who reviewed multiple papers, we use the most recent pre-experimental review. Each survival curve plots the percentage of reports still pending versus the number of days elapsed since the referee received the invitation. Before the experiment, all referees faced the six-week deadline and reminders were not sent systematically. The solid vertical lines depict the six-week deadline (45 days) and the four-week deadline (28 days) that were used during the experiment. The dashed vertical lines depict the reminders sent one week before each deadline during the experiment. We report median review times, defined as the point at which the fraction of reports pending is 50 percent, for each group. We also report *p*-values from nonparametric Wilcoxon tests for the hypothesis that the pre-experiment review times are the same in each treatment group and its corresponding control group. We compare the four-week and social incentive groups to the six-week group. We compare the cash group to the four-week group because the cash group also faced a four-week deadline. We truncate the x-axis at 80 days in the figure for scaling purposes, but use all available data for the hypothesis tests.

across the groups during the experiment are likely to be driven primarily by changes in referee behavior, with the possible exception of the social incentive group.

Figure 2 presents our main results on the impact of the treatments on review times during the primary experimental period. Panel A plots raw survival curves

for reviews by treatment group. In panel B, we adjust for selection using propensity score reweighting as in DiNardo, Fortin, and Lemieux (1996). We reweight the four-week, cash, and social incentive groups to match the six-week group on pre-experiment review times (including an indicator for having no pre-experiment data) using the procedure described in online Appendix D. We report median survival times (the point at which 50 percent of reports have been submitted) and nonparametric (Wilcoxon) tests for the equality of the survival curves in each figure (see online Appendix Table 3 for details).

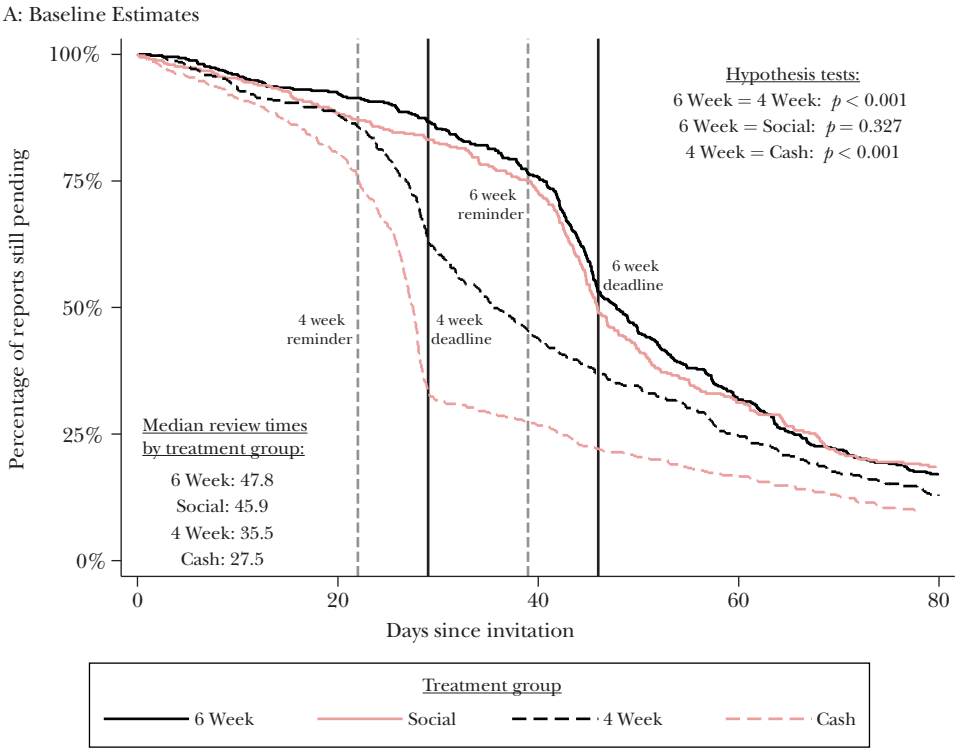
In contrast with the survival curves in Figure 1, the survival curves in Figure 2 diverge sharply, showing that the treatments induced substantial changes in review times. Adjusting for differences in prior review times (Figure 2B titled "Reweighted Estimates"), does not affect the results substantially, indicating that most of the change in review times is driven by changes in referee behavior rather than selection effects. We discuss next the impacts of each of the treatments in detail, starting with the shorter deadline and then turning to the cash and social incentives.

Shortening the deadline from six weeks (45 days) to four weeks (28 days) reduces median review times by 12.3 days, based on the baseline estimates Figure 2A. Hence, we estimate that shortening the deadline by one day reduces median review times by $12.3/(45-28) = 0.72$ days. The effect is so large because nearly 25 percent of referee reports are submitted in the week between the reminder email and the deadline, and the shorter deadline simply shifts these reports forward. Before week three (shown by the first dashed line in Figure 2), the number of pending reports in the four-week and six-week groups is not very different; however, in week four, the survival curve for the four-week deadline group drops sharply relative to the six-week group. The four-week deadline thus appears to act as a nudge that makes referees work on their reports in the fourth week rather than the sixth week.

Providing a $100 cash incentive for submitting a report within four weeks reduces median review times by an additional eight days relative to the four-week deadline. The cash incentive has powerful effects especially after referees receive the reminder email: nearly 50 percent of referees submit a report in the window between the reminder email and the deadline for receiving the cash payment. Missing the four-week deadline simply postpones writing the report by a few weeks but costs $100. Consistent with what one would predict based on a standard model of intertemporal optimization, the survival curve is much flatter immediately after the four-week deadline, as very few referees submit reports immediately after the cutoff for the cash payment. Nevertheless, because so many referees make an effort to meet the four-week deadline, there are fewer reports pending even 10 weeks after the initial invitation in the cash incentive group relative to all the other groups.

The strong response to the cash incentive in the week before the deadline also supports the view that the cash incentive changes referee behavior rather than the selection of referees who agree to review, as selection effects would be unlikely to generate such nonlinear responses. Indeed, the response to the cash treatment is so large that one can show using a nonparametric bounding approach, as in Lee (2009), that selection effects account for very little of the impact. Recall from Table 2 that

Figure 2
Review Times by Treatment Group during Experiment



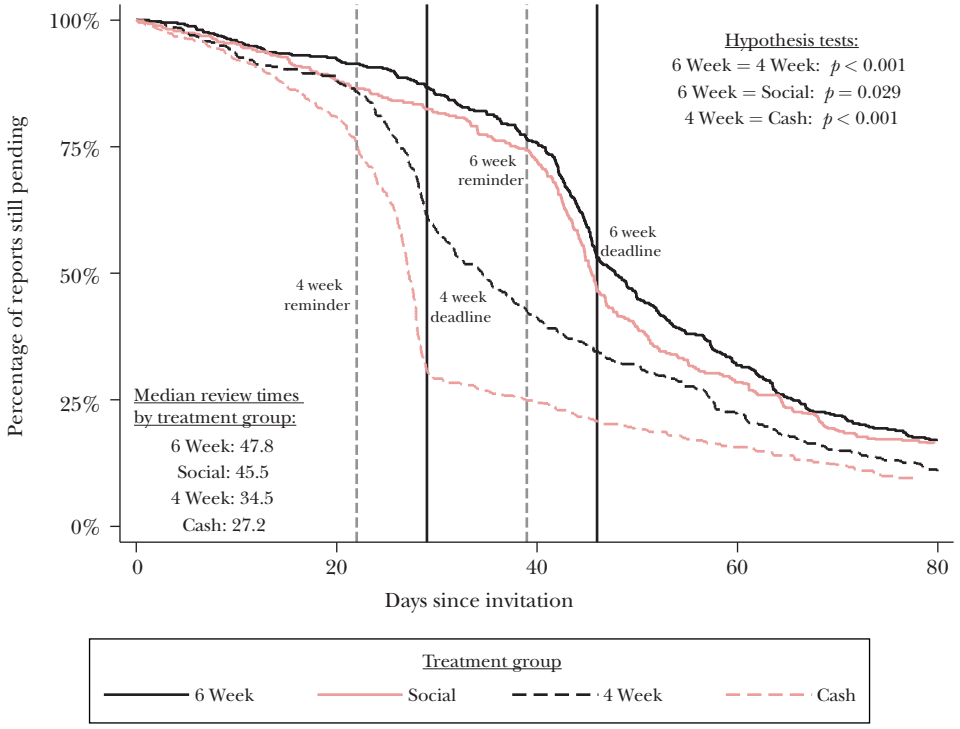
(continued on next page)

referees in the cash group are $12.3 = (72.0/64.1 - 1)$ percent more likely to accept review invitations than referees in the four-week group. Assuming that referees who accept the four-week invitation would also have accepted the (more attractive) cash invitation, we can bound the selection effect by considering the worst-case scenario in which the additional referees who accept the cash invitation have the shortest spells. For example, 66 percent of referees in the cash group submit their report within 28 days. If we exclude the 12.3 percent fastest referees in the cash group, we obtain a selection-adjusted lower bound of $(66 - 12.3)/(100 - 12.3) = 61$ percent submitting within 28 days. This remains well above the 36 percent of referees who submit a report within 28 days in the four-week group, showing that the difference in review times between the two groups cannot be caused by selection. A similar bounding exercise implies that the difference in review times between the four-week and six-week groups also cannot be due to selection.

Figure 2 demonstrates that the direct incentive effect of money outweighs any crowd-out of intrinsic motivation to submit referee reports in a timely manner. To investigate the impact of monetary incentives on intrinsic motivation more directly,

Figure 2 (continued)

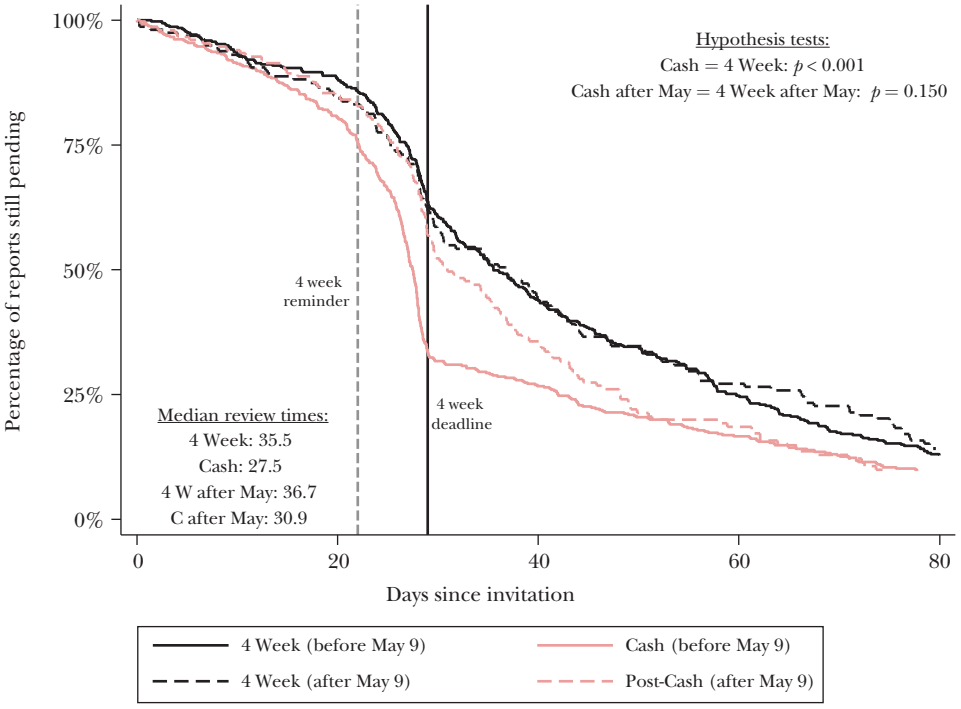
B: Reweighted Estimates



Notes: This figure plots survival curves showing the distribution of review times by treatment group during the primary experimental period, February 15, 2010, to May 9, 2011 (when the cash reward was offered). In panel A, each survival curve plots the percentage of reports still pending versus the number of days elapsed since the referee received the invitation. Panel B replicates panel A, reweighting the observations in the three treatment groups to match the distribution of pre-experiment review times in the six-week group (see online Appendix D for details). The solid vertical lines depict the six-week deadline (45 days) and the four-week deadline (28 days). The dashed vertical lines depict the reminders sent one week before each deadline. We report median review times, defined as the point at which the fraction of reports pending is 50 percent, for each group. We also report *p*-values from nonparametric Wilcoxon tests for the hypothesis that review times are the same in each treatment group and its corresponding control group. We compare the four-week and social groups to the six-week group. We compare the cash group to the four-week group because the cash group also faced a four-week deadline. We truncate the x-axis at 80 days in the figures but use all available data for the hypothesis tests.

we study the behavior of referees for the six months *after* the cash incentive ended on May 9, 2011. A long literature in social psychology starting with the classic work of Deci (1971) predicts that cash rewards have negative long-run effects on prosocial behavior by eroding intrinsic motivation. Existing evidence for this effect is based primarily on lab experiments (Deci, Koestner, and Ryan 1999; Frey and Jegen 2001; Kamenica 2012). Our experiment offers a new test of this hypothesis in the field, which complements earlier work on economic incentives and prosocial behavior

Figure 3
Review Times Before versus After the End of Cash Reward



Notes: This figure plots survival curves showing the distribution of review times in the four-week and cash treatment groups before versus after May 9, 2011. On May 9, cash rewards were stopped for those in the cash treatment group, and referees in this group were subsequently treated identically to those in the four-week group. Hence, the cash (after May 9) group includes referees who previously received cash rewards but no longer do, while the cash (before May 9) group includes referees receiving cash incentives. The four-week group faced the same treatment both before and after May 9. Each survival curve plots the percentage of reports still pending versus the number of days elapsed since the referee received the invitation. The solid vertical line depicts the four-week deadline (28 days). The dashed vertical line depicts the reminder sent one week before the deadline. We report median review times, defined as the point at which the fraction of reports pending is 50 percent, for each group. We also report *p*-values from nonparametric Wilcoxon tests for the hypothesis that review times are the same in the cash and four-week groups before and after May 9.

in other settings (for example, Gneezy and Rustichini 2000; Gneezy, Meier, and Rely-Biel 2011; Lacetera, Macis, and Slonim 2013).

In our application, the prediction from theories in which monetary payments crowd-out intrinsic motivation is that referees who had previously received cash incentives should become slower after they stop receiving cash payments—at least relative to referees in the four-week deadline group, who never received cash payments. We test this hypothesis in Figure 3, which plots survival curves for referees assigned to the four-week and cash incentive groups using data before May 9 versus

after May 9, when cash payments ended.[4] The survival curves for the four-week group are similar for invitations before and after May 9, indicating that review times do not vary significantly by invitation date. Referees assigned to the cash incentive group are much less likely to meet the 28-day deadline after May 9 than before May 9, when they were receiving cash rewards. However, there is no evidence that these referees become *slower* than those in the four-week comparison group, which is what one would expect if intrinsic motivation had been eroded. If anything, it appears that the cash treatment leads to some persistent improvements even after the incentive is removed, perhaps because referees have gotten in the habit of submitting reports slightly sooner.[5] We conclude that the temporary provision of monetary incentives does not have detrimental subsequent effects in the case of peer review.

Next, we turn to the social incentive treatment (where turnaround times were to be publicly posted). We find a statistically significant difference between the social incentive and control group survival curves when reweighting on pre-experiment durations in Figure 2B. The difference between the unweighted social and control survival curves in Figure 2A is smaller and statistically insignificant. This is because the social incentive treatment appears to induce slightly slower referees to accept review invitations, as shown in Figure 1. Once we adjust for this selection effect, we find that the social incentive treatment induces referees to work significantly faster, although the magnitude of the impact remains small. Based on the reweighted survival curves, we estimate that the social incentive reduces the median review time by 2.3 days.[6]

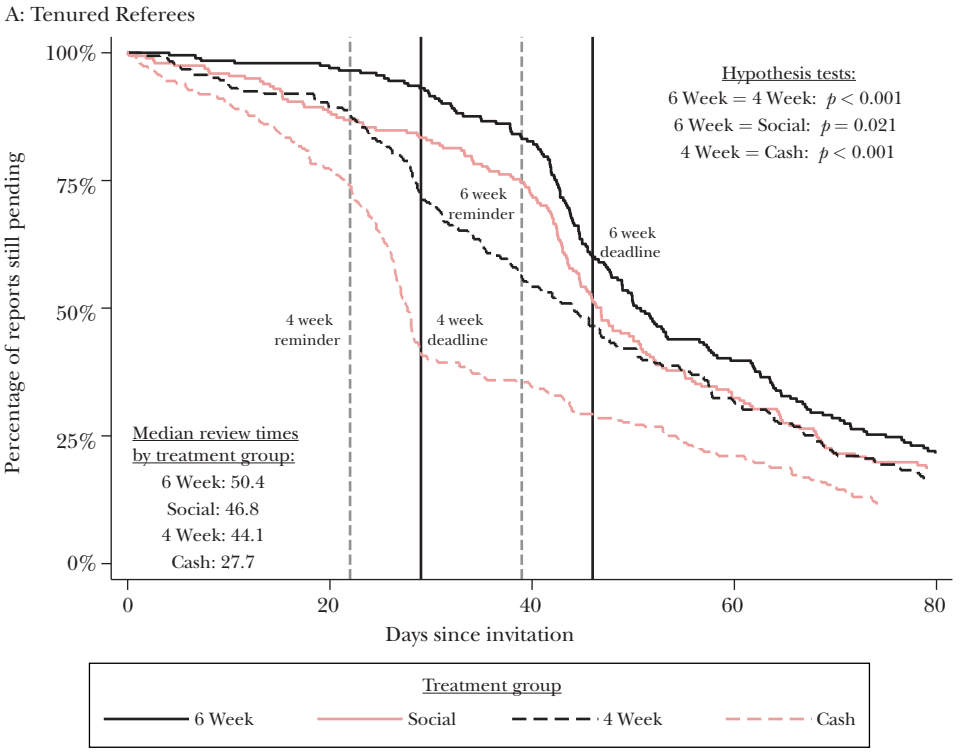
Finally, we explore the heterogeneity of the treatment effects by referee characteristics. We find no significant heterogeneity in treatment effects by several of

[4] Of the referees who were assigned to the cash incentive group and accepted a review invitation after May 9 (after the cash rewards had ended), 47 percent did not receive an invitation to review a manuscript before May 9. To minimize selection effects, we include these referees in Figure 3 even though they never received the cash incentive treatment. The estimates in Figure 3 should therefore be interpreted as intent-to-treat estimates. Restricting the sample to the selected subset of referees who received prior invitations yields very similar results.

[5] One might be concerned that referees did not recognize that the cash incentive had stopped after May 9, biasing our comparisons in Figure 3. Two facts allay this concern. First, if referees mistakenly thought the cash reward was still in place after May 9, one would expect to see the "Post-cash" survival curve in Figure 3 to drop steeply in the week before the four-week deadline. This does not occur: the "Post-cash" survival curve tracks the four-survival curves almost perfectly prior to the deadline. Second, the cash incentive increased agreement rates from 64.1 percent (in the four-week group) to 72.0 percent prior to May 9, as shown in Table 2. This difference also disappears after May 9: 64.1 percent of referees previously assigned to the cash incentive group agree to do the review after May 9, compared with 65.4 percent in the four-week group during the same period.

[6] We evaluate the robustness of the treatment effect estimates using semiparametric Cox hazard models in online Appendix E. Consistent with the graphical evidence in Figure 2, we find that the cash incentive and four-week deadlines substantially increase hazard rates of report submission, particularly in the week before the deadline. The social incentive treatment reduces review times significantly when controlling for differences in pre-experiment review times. These results, which are reported in online Appendix Table 4, are robust to changes in the control vector and sample specifications. In online Appendix Figure 2, we use all the data through the end of the experiment (October 26, 2011) rather than restricting the sample to the point at which cash treatments were stopped (May 9, 2011). The point estimates remain similar, but as expected, we obtain more precise estimates when using all the data.

Figure 4
Heterogeneity in Treatment Effects by Tenure Status

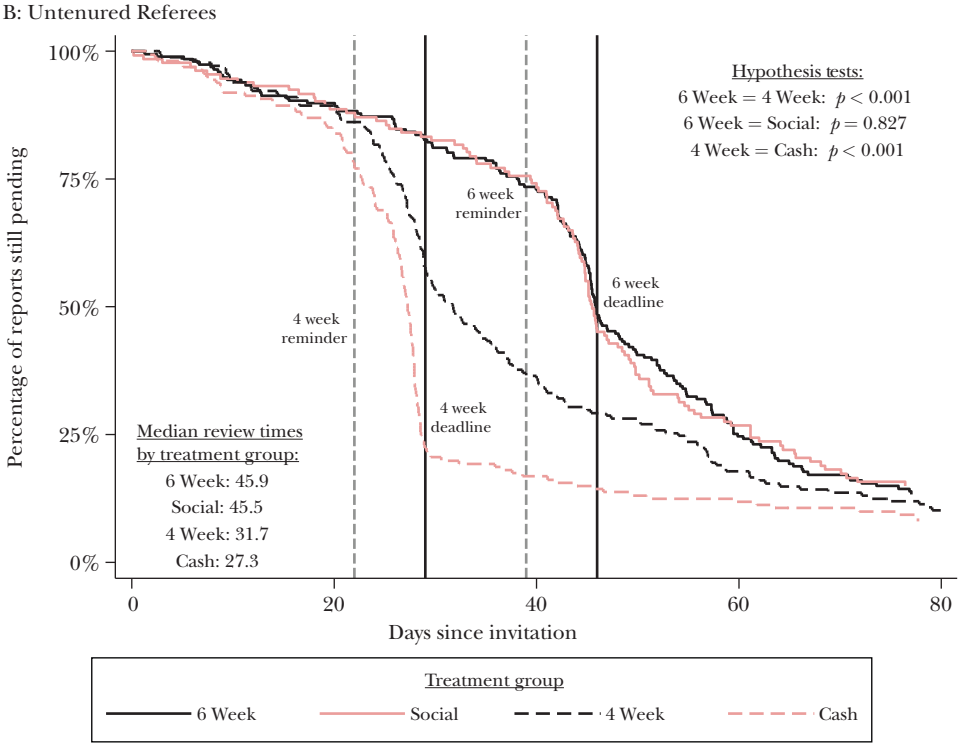


(continued on next page)

the referee characteristics we collected: an indicator for holding an academic position, gender, and an indicator for working in the United States. However, we do find substantial heterogeneity in treatment effects between tenured and untenured referees, as shown in Figure 4. This figure replicates Figure 2A, dividing the sample into referees who had tenure at the time they were invited to review the manuscript (Figure 4A) and those who were not tenured at that time (Figure 4B). The shorter deadline has a significantly larger effect on untenured referees than tenured referees. Untenured referees make a clear effort to submit reports before the deadline, as evident from the sharp drop in the survival curve in Figure 4B just before the deadline for the four-week group. In contrast, tenured referees are not as sensitive to the shorter deadline.

The cash incentive improves performance substantially in both groups, but again the impact is larger among untenured referees: 78 percent of untenured referees submit reports before the deadline to receive the cash reward, whereas only 58 percent of tenured referees do so. While the cash incentive and shorter deadline have smaller effects on tenured referees, the social incentive has *larger* effects on

Figure 4 (continued)



Notes: Figure 4 replicates Figure 2A, splitting the sample between tenured referees (panel A) and untenured referees (panel B). Tenure status is measured during the experiment based on information from CVs posted online (see online Appendix C for details); referees whose tenure status could not be identified are excluded from this figure. In both panels, the sample consists of referees who accepted invitations between February 15, 2010, and May 9, 2011, the period when the cash reward was offered. Each survival curve plots the percentage of reports still pending versus the number of days elapsed since the referee received the invitation. See notes to Figure 2 for further details.

tenured referees. Figure 4B shows that review times are almost identical in the social incentive and control groups for untenured referees. In contrast, tenured referees in the social incentive group submit reports significantly earlier than those in the control group, as shown in Figure 4A.

One explanation for why the social incentive treatment is more effective among tenured referees is that untenured referees are already concerned about their reputation with coeditors, who are typically senior colleagues in their field. In contrast, tenured referees might become more concerned about their professional reputation when they face social pressure.[7] Regardless of whether the heterogeneous

[7] Consistent with this explanation, we find that tenured referees are considerably slower than untenured referees in the control group, but behave like untenured referees in the social incentive group, as shown in online Appendix Figure 3.

effects are driven by this mechanism, the findings in Figure 4 suggest that social incentives can usefully complement other policy instruments by improving behavior among groups who are less responsive to cash incentives and nudges.

Outcome 3: Review Quality

Models of multitasking predict that if an agent is given an incentive to perform better in one aspect of a job (such as production speed), performance in other aspects of the job (such as quality) might deteriorate. Might the treatments that induce referees to submit reports more quickly also lead referees to submit lower-quality reviews?

We measure the quality of reviews in two ways. The first is an indicator for whether the editor follows the referee's recommendation with regard to whether the manuscript should be accepted, rejected, or revised and resubmitted. The second is the length of the referee report. While length is not equivalent to quality, one natural way in which referees might submit a report more quickly is by providing less-detailed comments to authors, especially because only the editor knows the referee's identity.

Table 3, which is constructed in the same way as Table 2, shows the fraction of cases in which the editor follows the referee's recommendation (panel A) and the median length of the referee report (panel B) by treatment group. We find no statistically significant differences across the groups in the rate at which editors follow the referee's advice. We do, however, find that referees write shorter reports to authors under the social and cash incentive treatments. The median report is approximately 100 words shorter (11 percent shorter) in the social and cash groups relative to the six-week and four-week groups. These findings suggest that referees who rush to submit a report earlier because of explicit cash or social incentives might cut back slightly on the level of detail in their comments to authors. Interestingly, referees do not write shorter reports to meet the four-week deadline, consistent with the view that many referees begin writing reports only in the week after they receive a reminder.

Overall, we conclude that one can induce referees to submit reviews more quickly without reducing the quality of reviews significantly. Shorter deadlines have no adverse effect on either measure of quality, while cash and social incentives induce referees to write slightly shorter reports but do not affect the quality of the review as judged by the editor's ultimate decision.

Outcome 4: Spillover Effects on Other Journals

A natural concern with interventions that improve referee performance at one journal is that they may have negative spillover effects at other journals. Do referees who submit reviews more quickly at the *Journal of Public Economics* prioritize them over other referee reports? In this case, changes in journal policies might not improve the overall efficiency of the review process.

We test for such spillover effects using data from 20 other Elsevier journals in related subfields, such as the *Journal of Health Economics* and the *Journal of Development Economics* (see online Appendix F for a complete list). We analyze referee invitations from other journals that are received 1) *after* referees have received an

Table 3
Measures of Review Quality by Treatment Group

	Group:			
	6 Week	*Social*	*4 Week*	*Cash*
Agreement between editor decision and referee's recommendation				
Editor follows referee's recommendation	77.9%	76.2%	77.5%	76.2%
	(2.00)	(2.34)	(2.20)	(2.15)
p-value for equality with control		0.585	0.884	0.686
Observations	403	324	373	404
Length of referee report				
Median number of words in referee report	877	757	864	786
	(29.1)	(32.5)	(30.3)	(29.2)
p-value for equality with control		0.006	0.757	0.064
Observations	401	321	369	399

Notes: This table shows the effects of the treatments on review quality. The sample includes all referees who received invitations sent between February 15, 2010, and May 9, 2011 (the period when the cash reward was offered) and submitted a report. In panel A, the outcome is the fraction of reports in which the editor's decision (reject versus accept/revise-and-resubmit) matches the referee's recommendation. We report standard errors in parentheses. Standard errors are clustered by referee in panel A (but not panel B). We also report *p*-values for the null hypothesis that the percentages are the same in each treatment group and its corresponding control group. For the social and four-week groups, the control group is defined as the six-week deadline group. For the cash incentive group, the control group is defined as the four-week deadline group, which is the relevant comparison because the cash incentive group also faced a four-week deadline. The number of observations (referee reports submitted) is reported in the last row. In panel B, the outcome is the median number of words in the referee report. Standard errors are reported in parentheses and the *p*-values are for hypothesis tests analogous to those in panel A. The number of observations is the number of submitted reports for which we were able to obtain automated word counts of report length.

invitation from the *Journal of Public Economics* during the primary experimental period and 2) before December 31, 2011.

Specifically, we test whether referees' propensities to review manuscripts and their review times at other journals vary across our four treatment groups. Each observation in this analysis is a referee invitation at another journal. The mean agreement rate is approximately 60 percent in all four groups, with no statistically significant differences across the groups (see online Appendix Table 5). Median review times are approximately 56 days in all four groups, again with no statistically significant differences across the groups (see online Appendix Figure 4).[8]

Of course, referees must postpone some activity to prioritize submitting referee reports. The social welfare impacts of our treatments depend on what activities get

[8] The similarity across the four groups in performance at other journals supports the view that the treatment effects at the *Journal of Public Economics* during the experimental period are driven by changes in referee behavior rather than selection effects.

postponed. If referees postpone activities with pure private benefits such as leisure, social welfare may increase because referee reports have positive externalities. If on the other hand referees postpone working on their research or on other prosocial tasks, expediting referee reports could reduce welfare. If small delays in these other activities have little social cost, the welfare costs from such delays would be modest. Understanding the nature of crowd-out across different forms of prosocial behavior is an interesting question that we defer to future research.

Lessons for the Peer Review Process

Our results offer three lessons for the design of the peer review process at academic journals.

First, shorter deadlines are extremely effective in improving the speed of the review process. Moreover, shorter deadlines generate little adverse effect on referees' agreement rates, the quality of referee reports, or performance at other journals. Indeed, based on the results of the experiment, the *Journal of Public Economics* now uses a four-week deadline for all referees.

Second, cash incentives can generate significant improvements in review times and also increase referees' willingness to submit reviews.[9] However, it is important to pair cash incentives with reminders shortly before the deadline. Some journals, such as the *American Economic Review*, have been offering cash incentives without providing referees reminders about the incentives; in this situation, sending reminders would improve referee performance at little additional cost.

Third, social incentives can also improve referee performance, especially among subgroups such as tenured professors who are less responsive to deadlines and cash payments. Light social incentives, such as the *Journal of Financial Economics* policy of posting referee times by referee name, have small effects on review times. Stronger forms of social pressure—such as active management by editors during the review process in the form of personalized letters and reminders—could potentially be highly effective in improving efficiency. It would be useful to test this hypothesis in future work using an experiment in which editors are prompted to send personalized reminders to referees at randomly chosen times.

More generally, our findings show that it is possible to improve the efficiency of the peer review process substantially with relatively low-cost interventions, demonstrating the value of studying the peer review process empirically (as in Card and DellaVigna, this issue). Our results reject the view that the review process in

[9] These findings contrast with the results of Squazzoni, Bravo, and Takács (2013), who argue that monetary rewards decrease the quality and efficiency of the review process based on a lab experiment designed to simulate peer review. Our results might differ because the peer review process requires referees to invest considerable time to read papers and write referee reports, unlike the investment game studied in this lab experiment.

economics is much slower than in other fields, such as the natural sciences, purely because economics papers are more complex or difficult to review.

Lessons for Increasing Prosocial Behavior

Beyond the peer review process, our results also offer some insights into the determinants of prosocial behavior more broadly.

First, attention matters: reminders and deadlines have significant impacts on behavior. Nudges that bring the behavior of interest to the top of individuals' minds are a low-cost way to increase prosocial behavior, consistent with a large literature in behavioral economics (Thaler and Sunstein 2008).

Second, monetary incentives can be effective in increasing some forms of prosocial behavior. We find no evidence that intrinsic motivation is crowded out by financial incentives in the case of peer review, mirroring the results of Lacetera, Macis, and Slonim (2013) in the case of blood donations. While crowd-out of intrinsic motivation could be larger in other settings, these results show that one should not dismiss corrective taxes or subsidies as a policy instrument simply because the behavior one seeks to change has an important prosocial element.

Finally, social incentives can be effective even when other policy instruments are ineffective. This result echoes findings in other settings—such as voting (Gerber, Green, and Larimer 2008), campaign contributions (Perez-Truglia and Cruces 2013), and energy conservation (Allcott 2011)—and suggests that social incentives are a useful complement to price incentives and behavioral nudges.

■ *We thank Nava Ashraf, David Autor, Stefano DellaVigna, Hilary Hoynes, Damon Jones, Emir Kamenica, Lawrence Katz, Henrik Kleven, Ulrike Malmendier, Monica Singhal, Timothy Taylor, and numerous seminar participants for helpful discussions and comments. We thank Jenny Henzen and the Elsevier staff for supporting the project and providing data and Liz Anderson for implementing the experiment. Greg Bruich, Jessica Laird, Keli Liu, Alex Olssen, and Heather Sarsons provided outstanding research assistance. Financial support from the Lab for Economic Applications and Policy at Harvard and the National Science Foundation is gratefully acknowledged. A de-identified version of the referee-level dataset used in this paper can be found at the JEP website, with the paper.*

References

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95(9–10): 1082–95.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96(5): 1652–78.

Deci, Edward L. 1971. "Effects of Externally Mediated Rewards on Intrinsic Motivation." *Journal of Personality and Social Psychology* 18(1): 105–15.

Deci, Edward L., Richard Koestner, and Richard M. Ryan. 1999. "A Meta-analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin* 125(6): 627–68.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach." *Econometrica* 64(5): 1001–1044.

Ellison, Glenn. 2002. "The Slowdown of the Economics Publishing Process." *Journal of Political Economy* 110(5): 947–93.

Fehr, Ernst, and Urs Fischbacher. 2003. "The Nature of Human Altruism." *Nature* 425(6960): 785–91.

Frey, Bruno S., and Reto Jegen. 2001. "Motivation Crowding Theory: A Survey of Empirical Evidence." *Journal of Economic Surveys* 15(5): 589–611.

Gerber, Alan S., Donald P. Green, and Cristopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-scale Field Experiment." *American Political Science Review* 102(1): 33–48.

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25(4): 191–210.

Gneezy, Uri, and Aldo Rustichini. 2000. "A Fine is a Price." *Journal of Legal Studies* 29(1): 1–17.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal–Agent Analyses: Incentives Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics and Organization* 7(Special Issue): 24–52.

Kamenica, Emir. 2012. "Behavioral Economics and Psychology of Incentives." *Annual Review of Economics* 4: 427–52.

Lacetera, Nicola, Mario Macis, and Robert Slonim. 2013. "Economic Rewards to Motivate Blood Donations." *Science* 340(6135): 927–28.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In *Handbook of Experimental Economics*, edited by John H. Kagel and Alvin E. Roth, 111–94. Princeton University Press.

Lee, David. S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76(3): 1071–1102.

Perez-Truglia, Ricardo, and Guillermo Cruces. 2013. "Social Incentives in Contributions: Field Experiment Evidence from the 2012 U.S. Presidential Campaigns." Unpublished paper.

Squazzoni, Flaminio, Giangiacomo Bravo, Károly Takács. 2013. "Does Incentive Provision Increase the Quality of Peer Review? An Experimental Study." *Research Policy* 42(1): 287–94.

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* New Haven: Yale University Press.

Titmuss, Richard M. 1971. *The Gift Relationship.* London: George Allen and Unwin.

Vesterlund, Lise. Forthcoming. "Voluntary Giving to Public Goods: Moving Beyond the Linear VCM." In *Handbook of Experimental Economics,* Vol. 2, edited by John H. Kagel and Alvin E. Roth. Princeton University Press.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 189–204

The Effects of an Anti-Grade-Inflation Policy at Wellesley College[†]

Kristin F. Butcher, Patrick J. McEwan, and Akila Weerapana

Average grades in colleges and universities in the United States are markedly higher than they were several decades ago. In 1960, the average grade point average for all private and public institutions was about 2.4, or a little above a C+. By 2006, this number was about 3.0, or roughly a B, and even higher in private institutions (Rojstaczer and Healy 2010). Courses in the humanities usually have the highest grades, science and math courses have the lowest grades, and social sciences fall somewhere in the middle (Rojstaczer and Healy 2010). Since the highest available grade is usually an A, this means that grade inflation has gone hand-in-hand with compression at the top of the distribution, particularly in the humanities.

If grades are the fundamental way in which students, administrators, graduate schools, and employers receive information about an individual's absolute and relative abilities, then grade inflation and compression masks valuable information and distorts choices. Based in part on grades, students make choices about how hard to work (Babcock 2009), courses (Sabot and Wakemann-Linn 1991), majors, and careers. Administrators make choices about where to allocate academic support services, graduate schools make choices about whom to admit (Wongsurawat 2009), and employers make choices about whom to hire (Chan, Hao, and Suen 2007).

In the early 2000s, the faculty and administration at Wellesley College—a selective, small, private, women's liberal arts college outside of Boston,

■ *Kristin F. Butcher is the Marshall I. Goldman Professor of Economics, Patrick J. McEwan is Professor of Economics, and Akila Weerapana is Associate Professor of Economics, all at Wellesley College, Wellesley, Massachusetts. Butcher is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Butcher is the corresponding author at kbutcher@wellesley.edu.*

[†] To access the data Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.3.189

doi=10.1257/jep.28.3.189

Massachusetts—concluded that grade inflation and compression was causing a number of these problems, potentially undermining the institution's credibility and reputation. Thus, the College implemented the following policy in Fall 2004: average grades in courses at the introductory (100) level and intermediate (200) level with at least 10 students should not exceed a 3.33, or a B+. The rule has some latitude. If a professor feels that the students in a given section were particularly meritorious, that professor can write a letter to the administration explaining the reasons for the average grade exceeding the cap. Grades by department are reported to administrators and faculty during Academic Council meetings, the main governing body at the institution, so that peers can see if some departments are regularly violating the policy. Penalties are left to the discretion of the administration. The grading policy is detailed on the Registrar's web page, and is known to students, prospective students, and alums.

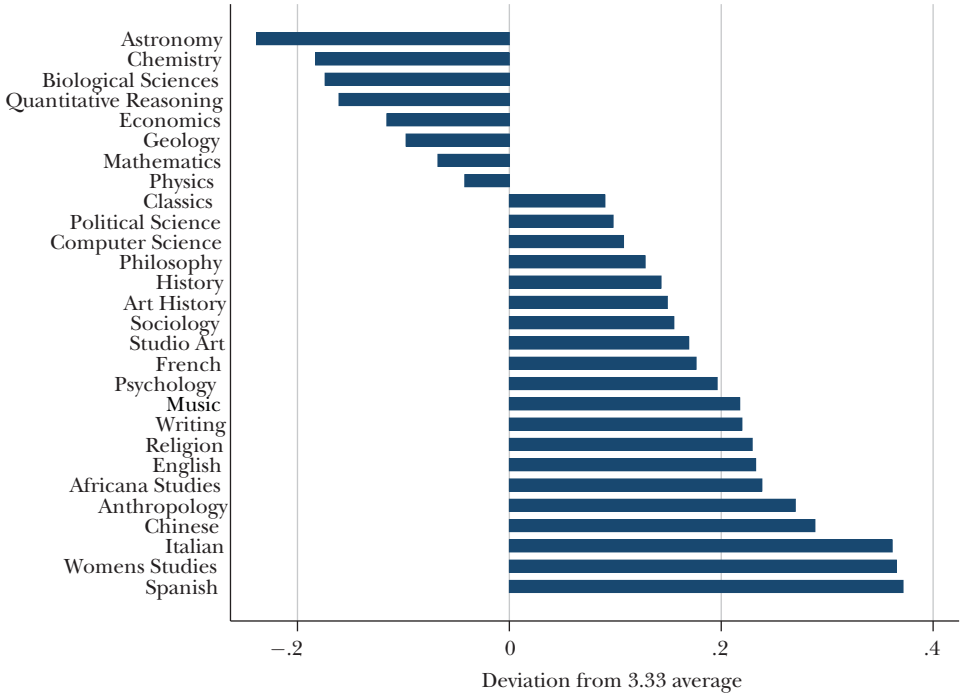
Since grading patterns at Wellesley College were similar to those across academia, only courses in high-grading departments in the humanities and social sciences (except economics) needed to change grading practices in order to comply. In this paper, we evaluate the consequences of the policy by comparing outcomes in departments that were obligated to lower their grades with outcomes in departments that were not. The policy had an immediate effect, bringing average grades down in the previously high-grading departments. Faculty complied by reducing compression at the top of the grade distribution, but there is little evidence that they increased the use of very low grades. We also examine the impact of the change in grading policy for different subgroups. For African-American students and students with low initial test scores, the gap in grade point average with other students increased in the departments where grades were reduced. There is little evidence of a change in sorting by student quality—as measured by students' initial test scores—into capped and uncapped courses. However, the number of students enrolled in courses in capped departments and the number of majors in these departments show relative declines. It also appears that students reduced their evaluations of their professors' performance in response to the change in the grading policy. The implications of these changes will be discussed in the concluding section.

How the Cap Affected Grades

Figure 1 shows how department and program average grades deviated from the policy target in the years prior to the adoption of the 3.33 cap (in 100- and 200-level courses with more than 10 students, from Fall 1998 to Spring 2003). At Wellesley, as at other institutions, it is largely the sciences that were low-grading and other departments that were high-grading.[1] Average grades were below the cap

[1] Given that these patterns persist across institutions, it is plausible that there are inherent differences in the cost of giving lower grades across disciplines. Achen and Courant (2009) and Franz (2010) have suggested that if typical assessment mechanisms are costlier in the humanities than in the sciences—for

Figure 1
Pre-policy Grade Differences across Departments



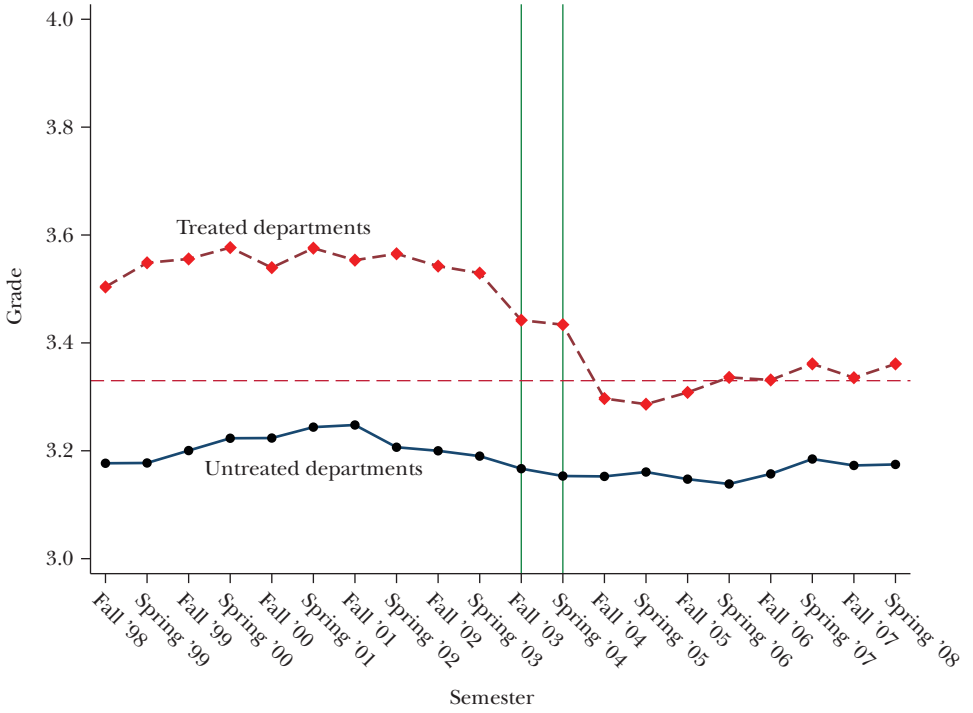
Source: Authors using student-transcript-level data from Wellesley College.
Note: Figure 1 shows how department and program average grades at Wellesley deviated from the policy target in the years prior to the adoption of the 3.33 cap (in 100- and 200-level courses with more than 10 students, from Fall 1998 to Spring 2003).

in Astronomy, Chemistry, Biological Sciences, Quantitative Reasoning, Economics, Geology, Mathematics, and Physics. Other departments had grades above the cap, with at least three departments exceeding it by a third of a letter grade.

Figure 2 shows what happened to average grades in "treated" and "untreated" departments from Fall 1998 to Spring 2008. Grades dropped a bit in the soon-to-be-treated departments in Fall 2003 when the policy was publicly debated (and so data from the academic year 2003–2004 are dropped from subsequent analyses). After the adoption of the policy, treated departments lowered their grades enough to comply.

example, if it is more difficult both to see that an argument in a paper is unsupported than to see that an answer to a math problem is incorrect, and to communicate that to students—then faculty in some disciplines will have more incentive to inflate grades. Smaller classes, and the nature of the work, may make a low grade feel more personal, and thus more difficult to both give and receive, in a humanities class than in a science class. In addition, when small enrollments may lead to a program being reduced or eliminated by an institution's administration, faculty are under pressure to maintain enrollment levels, potentially by giving higher grades.

Figure 2
Change in Grading Patterns over Time
(Fall 1998 to Spring 2008)



Source: Authors using student-transcript-level data from Wellesley College.
Notes: The two vertical lines indicate the semesters when the policy was introduced and then implemented. The horizontal line shows the 3.33 average grade that the policy used as the new standard.

There is some evidence of backsliding in treated departments: since Spring 2006, average grades in the treated departments slightly exceeded the cap. Nonetheless, by and large, the policy was effective in its basic goal of lowering average grades in high-grading departments to the agreed-upon target.

To assess the effect of the policy on academic outcomes, we use a standard difference-in-differences methodology: that is, we compare the change in outcomes in the treated departments to the change in outcomes in the untreated departments. Under the assumption that factors affecting outcomes—other than the policy change—are the same in the treated and untreated departments, the approach will estimate the effect of the grading policy on a given outcome. It is a problem for this methodology if, during this period, there are differential changes across departments affecting the outcomes of interest that are separate from the grading policy. For example, if Wellesley attracted a particularly talented group of young scientists for the Fall 2004 term, then the underlying quality of students in the different

departments would be changing at the same time as the policy, and these changes may have their own independent effect on grades and other outcomes.

Alternatively, if the policy resulted in differential sorting across departments—perhaps because some students are more grade-sensitive than others—then there would, again, be an underlying change in the quality of students that might exacerbate the changes in outcomes across departments.[2] To the extent that these potential changes in student quality across departments are observable—for example, in baseline test scores—we can control for them by estimating the following regression model:

$$Y_{idt} = \beta_0 + \beta_1 PostPolicy_t + \beta_2 Treated_d + \beta_3 PostPolicy_t{*}Treated_d + X_{idt}\Gamma + \varepsilon_{idt},$$

where Y is the outcome of interest; i indexes the individual, d the department, and t the semester or year; *PostPolicy* is a dummy for Fall 2004 and beyond; and *Treated* is a dummy variable equal to 1 for those departments with average grades above the cap prior to the policy. The coefficient of interest is β_3, because it indicates whether the outcomes for students in courses in the treated departments changed differentially before and after the policy was implemented.[3] X_{idt} is a vector of fixed and time-varying individual, course, and department characteristics. Standard errors are clustered at the department level.

Our main dataset is transcript-level data on student grades and courses from Fall 1998 to Spring 2008.[4] Wellesley College has a diverse and high-quality student body. Over the time period, 45 percent of the students identified as white, 22.8 percent identified as Asian, 5 percent as African American/black, and 5 percent as Latina. About 10 percent of students are first-generation college students, and about 10 percent are legacy students, related to an alum of the institution. Average SAT scores are high, consistent with Wellesley's status as a highly-selective institution. Roughly one-third of the courses were taken in the humanities and languages, 40 percent in the social sciences, and 23 percent in sciences/math, with the remainder in other programs. Wellesley has distribution requirements such that all students must take 100- and 200-level courses in all divisions in the college

[2] For example, work by Goldin (2013) shows that women students are less likely to major in economics than their male counterparts if they receive a low grade in introductory courses. This point will be discussed further in the conclusion, but it is worth noting here that grades remain higher in the humanities and non-economics social sciences even after the policy, and so students who were sensitive to low grades might still prefer to take courses in these disciplines.

[3] Figure 1 indicates that some departments had more change to make in order for their grades to be compliant with the new policy. We experimented with an "intensity of treatment" variable rather than the simple dichotomous "treated" variable described above, and found that the results are qualitatively similar.

[4] Summary statistics for students and courses pre- and post-policy adoption are shown in online Appendix Tables 1 and 2. There are 116,374 student-course-semester observations covering outcomes for over 8,000 students. The average numeric value for grades before the policy was about 3.5. About 8 percent of students elected to take a given course "credit/non," which means that a student had to receive a C or better in order to pass the course but her GPA is not affected by receiving a "credit." In addition, 1.3 percent of students withdrew from classes. If a student withdraws after the "drop" period but before the end of classes, she receives a "withdrawal" on her transcript, but, again, this does not affect her GPA.

Table 1
Impact of Policy on Grades Awarded to Students

	All (1)	*All* (2)	*All* (3)	*Black* (4)	*Latina* (5)	*Lowest 5% SATVerbal* (6)	*Low 5% QR* (7)
Treated	0.341***	0.300***					
	(0.026)	(0.033)					
PostPolicy	−0.048***	−0.094***					
	(0.013)	(0.014)					
*PostPolicy*Treated*	−0.175***	−0.172***	−0.174***	−0.363***	−0.081**	−0.299***	−0.334***
	(0.020)	(0.020)	(0.021)	(0.043)	(0.038)	(0.039)	(0.037)
SATMath/100		0.0720***					
		(0.018)					
SATVerbal/100		0.036***					
		(0.009)					
Observations	104,454	104,454	104,454	5,476	5,497	4,217	4,814
R^2	0.074	0.136	0.469	0.501	0.494	0.497	0.507
Demographics	No	Yes	No	No	No	No	No
Other controls	No	Yes	Yes	Yes	Yes	Yes	Yes
Department fixed effects	No	No	Yes	Yes	Yes	Yes	Yes
Student fixed effects	No	No	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	No	No	Yes	Yes	Yes	Yes	Yes

Source: Authors using student-transcript-level data from Wellesley College.
Notes: In columns 3 through 7, department and semester fixed effects absorb the effects of *Treated* and *PostPolicy*, and student fixed effects absorb student demographics. *Black* is an indicator variable set equal to 1 for African American students and for students from other countries who self-identify as black. "Demographics" include indicator variables for race/ethnicity and for non-traditional-aged students, legacy students, and first-generation college students. "Other controls" include class size and an indicator variable for 200-level courses. Column 2 also includes indicator variables for humanities, social science, or science/math. Samples sizes for the last two columns differ because SAT Verbal scores are not available for all students and the Quantitative Reasoning test was not administered to the first set of graduating students in our sample. Robust standard errors are clustered by department.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

(humanities, science, and social sciences), meaning that all students will have experiences in both the treated and untreated departments.

Table 1 reports the estimated coefficients for regression models in which the dependent variable is the numeric value of the course grade. The first column reports the unadjusted difference-in-differences estimate (corresponding to the results in Figure 2). The second column controls for math and verbal SAT scores. It also includes dummy variables for black, Latina, international, Asian, non-traditional-aged student, first-generation college student, and legacy student as well as other controls for class size, whether a course is humanities, social science, or science/math and an indicator variable for whether it is a 200-level course. In column 3, student, department, and semester fixed effects are added. These control

for all fixed observable and unobservable differences across students, departments, and semesters; these fixed effects absorb the *Treated* and *PostPolicy* variables, as well as fixed student characteristics like race or SAT scores. Adding these controls has little effect on the estimate of β_3, the key parameter. Across columns 1–3, the coefficient on *PostPolicy*Treated* estimates the impact of the policy on grades to be about −0.17, or a relative drop of about a sixth of a letter grade in courses in treated departments.

Columns 4–7 report the estimated effect of the policy for various subgroups, using the same model as in column 3. The estimated drop in grades in treated departments is smaller for Latina students but much larger than average for black students (including African-Americans and foreign students who self-identify as black), those with low SAT verbal scores, and those with low Quantitative Reasoning scores. The results in columns 6 and 7 of Table 1 are very similar when black students are dropped from the sample, indicating that the results for those with low SAT verbal or low Quantitative Reasoning scores are not being driven by that group being disproportionately likely to be black. The estimated drops in grades for these groups are statistically different from the average drop at 1 percent level of significance, except for the Latina group, where the drop is statistically significant at the 5 percent level. If black students and those with low SAT verbal and Quantitative Reasoning scores have lower average outcomes than other students, and if, prior to the policy, those differences were being masked by grade compression in the treated departments, then the result is a natural outcome of the reduction in grade compression. If all students receive an A− or higher before the policy, and after the policy that is no longer possible, then the student who was receiving the "lowest A−" is likely to have her grades hit harder by the policy.[5] Whether that is a good or bad consequence of the policy will be taken up in the conclusion.

We now turn to examining in more detail how faculty complied with the new rule. Table 2 shows a series of linear probability estimates, where the outcomes are dummy variables equal to 1 if the student got a particular grade, and zero otherwise. So, in the first column, the outcome is 1 if the student got a straight A, and zero otherwise. In column 5, the outcome is equal to 1 if the student got a C− or below (including students who took the course "credit/non" and received no credit), and zero otherwise. Column 6 reports results for whether students withdrew or not, and column 7 for whether they elected to take the course credit/non. (The sample sizes are larger in columns 5–7 because students electing credit/non are included in the results.) The specifications control for department, semester, and student fixed effects and class size. The results are robust to the other specifications as well.

After the policy change, students were about 14 percentage points less likely to get a straight A in the treated departments. On average before the policy was

[5] Like black students, Latina students also tend to have lower grades on average than white students in the pre-policy period (even controlling for test scores), but the gaps are smaller, and Latina students and black students have a different distribution of courses. Since treatment "intensity" is different across departments (see Figure 1), the fact that Latina students occupied a different relative position in the grade distribution and have a different distribution of courses, may explain why their grades were less affected by the policy change.

Table 2
Impact of Policy on Faculty Grading Behavior

	Straight A (1)	*A or A−* (2)	*B+* (3)	*B+, B, or B−* (4)	*C− or below* (5)	*Withdraw* (6)	*Credit/non* (7)
*PostPolicy*Treated*	−0.141***	−0.184***	0.072***	0.167***	0.011	−0.007	0.069***
	(0.022)	(0.019)	(0.013)	(0.023)	(0.007)	(0.006)	(0.012)
Constant	0.077***	0.250***	0.243***	0.608***	0.049***	0.007	−0.039**
	(0.016)	(0.025)	(0.015)	(0.017)	(0.008)	(0.005)	(0.016)
Observations	104,454	104,454	104,454	104,454	116,374	116,374	116,374
R^2	0.317	0.370	0.132	0.250	0.227	0.114	0.136
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Department fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Source: Authors using student-transcript-level data from Wellesley College.
Notes: Other controls include class size as well as an indicator variable for 200-level courses. C− or below includes students who took a course "credit/non" but did not pass. Samples sizes for the last three columns differ because students who did not take the course for a letter grade are included in the subsamples. Robust standard errors (clustered by department) in parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

enacted, about 29 percent of the grades in these departments were straight As, so a 14 percentage point drop is substantial. Column 2 of Table 2 indicates that students were about 18 percentage points less likely to get an A or A−. Correspondingly, the probability of receiving a B+ increased by about 7 percentage points, and any type of B increased by about 17 percentage points. There was essentially no change in the incidence of very low grades, C− or below, and no change in the share of students withdrawing from courses. This evidence, then, suggests that the policy alleviated the compression of grades at the very top of the distribution. There is no evidence that faculty began using very low grades for some students in order to "preserve" A grades for other students while still meeting the policy goal of a 3.33 average.

There is evidence that the difference in the probability of taking a course credit/non changed between the treated and untreated departments. The change is driven by a precipitous drop in the credit/non elections in the departments that were *unaffected* by the grade cap. This effect, it turns out, can be traced to a policy adopted in Spring 2003—also designed to make outcomes in courses more informative—that moved the date by which students had to declare whether they were electing to take something credit/non much earlier in the semester, often before they would have received exam results. Before this policy change, students were more likely to take courses in the low-grading departments using the credit/non option. Once students had to decide earlier, their election of credit/non dropped in the lower-grading

departments, converging to the same rates as in the higher-grading departments. The overall effect of this change would likely be to push grades down in the departments that were "untreated" by the grade cap because low performance was less likely to be masked by taking a course for credit only. Thus, the relative decline in grades in the departments that were affected by the 3.33 grade cap may be understated.

One oft-cited concern about persistent differences in grade levels across departments is that students majoring in high-grading departments are disproportionately rewarded with Latin honors. When we investigated these patterns, we found that the probability of graduating *summa cum laude* was not (significantly) differentially changed across departments by the policy. In order to graduate *summa*, a student needs a grade point average of 3.9 or above. At Wellesley, with its substantial distribution requirements, this standard means that a student needs to get nearly straight As in many different types of courses. These students are top performers across disciplines, and thus the *summa* students were not differentially affected by grading policies across departments and so were not differentially affected by the imposition of the grading cap.

However, the probability that a student graduated *magna cum laude* was notably affected by the policy. For the treated departments, about 20 percent of students received this designation in the pre-policy era; this fell to 16 percent after the policy was enacted. There was no statistically significant differential change across treated and untreated departments in the probability that a student was graduated *cum laude*. Presumably, many of those who would have previously received a *magna* designation slipped into the *cum laude* category, off-setting any declines from that category. These patterns are consistent with less grade compression in the upper portion of the grade distribution.[6]

Students' Choices of Courses and Majors

Grade inflation is often blamed for distorting students' choices across fields by misinforming them about their relative strengths. Thus, we might expect that a change in grading policy would lead students to make different choices about which course to take or in which departments to major.[7] How did students alter their behavior in response to the grade cap policy?

[6] Details of these calculations are available from the authors. In addition, there was no relative change across treated and untreated departments in the probability of Phi Beta Kappa designation after the policy. Phi Beta Kappa is restricted to no more than 12 percent of a graduating class and is selected by a committee; the fact that there was no change suggests that the committee was implicitly taking into account that it was harder to reach a certain grade point average in some departments than others.

[7] One might also expect differential choices to lead to differential sorting across courses by student type. We investigated this by using initial test scores (for example, lowest 5 percent among Wellesley students on SAT verbal scores and lowest 5 percent on Quantitative Reasoning scores) as the outcome variable, and asking whether scores among students electing to take courses in treated and untreated departments changed after the policy. There was no statistically significant relative change in initial test scores. These results are available in Table 4 of the online Appendix available with this paper at http://e-jep.org.

Table 3
Impact of Policy on Enrollments and Majors

	Enrollment (1)	*ln(Enrollment)* (2)	*300-Level Enroll* (3)	*ln(300-Level Enroll)* (4)	*Majors* (5)	*ln(Majors)* (6)
*PostPolicy*Treated*	−51.26***	−0.186***	0.066	0.019	−8.406***	−0.307***
	(10.67)	(0.033)	(4.024)	(0.074)	(2.891)	(0.099)
Constant	348.8***	5.785***	66.16***	3.731***	35.80***	3.311***
	(11.93)	(0.041)	(2.819)	(0.105)	(1.784)	(0.077)
Observations	342	342	342	342	200	200
R^2	0.890	0.874	0.919	0.867	0.929	0.877
Department fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Source: Authors using student-transcript-level data from Wellesley College.
Notes: To avoid giving undue influence to enrollments in smaller departments, we combined smaller departments by area creating the following 19 departments (and 20 majors): Art History, English, French, Spanish, Studio Art, Other Languages, Other Humanities, Economics, History, Philosophy, Political Science, Psychology, Religion, Other Social Sciences, Biological Chemistry (major only), Biological Sciences, Chemistry, Computer Science, Mathematics, and Other Sciences. We count the number of majors using what the student had listed as her major(s) in her final semester. In columns 5 and 6, the post-treatment period begins with the class of 2006 to account for lags in the policy's effect on the number of majors. Robust standard errors are in parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Here we use department- and program-level data to examine the number of enrollments and majors. Similar to the earlier results, we use a difference-in-differences methodology to examine whether those departments that were "treated" by the policy change (because they had grades above the 3.33 grade cap) had a differential change in enrollments and majors after the policy change. Table 3 presents results for the number of students enrolled in courses and the number of majors. Wellesley has some very small departments and programs, so we combined these enrollments into groups (see note to Table 3 for details), resulting in 342 semester-by-department observations on total enrollments. In columns 1 and 2, the outcome is enrollments in 100- and 200-level courses with at least 10 students (the courses subject to the grading policy), in either levels or logs. The results suggest a substantial decline in total enrollments for departments affected by the policy: total enrollments fell by about 51 students (levels) or by about 18.6 percent (logs) in treated departments.

As a placebo test of these results, columns 3 and 4 report whether enrollments in upper-level classes at the 300-level differed contemporaneously across treated and untreated departments. These enrollments should not have changed at this time because upper-level classes were not bound by the policy of the average grade being no higher than a B+. Since students typically take lower-level classes before moving on to upper-level classes, the policy should only affect upper-level enrollments with

a lag. Thus, if we find a statistically significant contemporaneous *PostPolicy*Treated* coefficient for department enrollments in upper-level classes, that would be an indication that the result for 100- and 200-level enrollments is merely picking up differential trends in enrollments across departments. However, as columns 3 and 4 show, the *PostPolicy*Treated* coefficients are small and statistically insignificant for upper-level enrollments.

In columns 5 and 6, we examine whether the number of majors was affected. There are only 200 observations in these columns because we count the number of majors only among second-semester seniors. Only those graduating in 2006 or later are counted as being in the post-policy era because earlier cohorts would not have had time to change major in response to the policy. The results suggest that majors declined in the treated departments by about eight students, on average, representing a relatively large decline of about 30 percent. Which departments gained and which departments lost majors? The fraction of a graduating class majoring in economics (and to a lesser extent in the sciences) increased, and the fraction of a graduating class majoring in other social sciences fell, with the fraction remaining flat in the humanities.[8]

Students' Evaluations of their Professors

Despite the fact that students' evaluations of their professors are a contentious measure of teaching effectiveness, they are a nearly universal feature of the US higher education landscape.[9] Some observers contend that this system of evaluating professors has contributed to grade inflation because students' evaluations of their professors set up an implicit quid pro quo with professors offering higher student grades in exchange for higher evaluations—evaluations that are used in tenure, promotion, and merit reviews (Zangenehzadeh 1988; Pressman 2007). It is the case at Wellesley that students in courses with higher average grades also tend to have higher evaluations of the quality of their professors' instruction, but this correlation cannot be taken as evidence that higher grades yield higher evaluations since higher average grades may indicate the teacher was effective, students learned

[8] For additional evidence of this, see Figure 1 in the online Appendix available with this paper at http://e-jep.org.

[9] Recent research that relies on random assignment to classes tends to support how difficult it is to assess teaching effectiveness. Braga, Paccagnella, and Pellizzari (2011) and Carrell and West (2010) use the fact that students are randomly assigned to classes, and develop measures of teachers' effectiveness based on students' performance in *subsequent* classes. Carrell and West (2010) find that students whose professors' teaching efforts yielded higher grades on a common exam tended to do less well in subsequent courses, suggesting that "teaching to the test" resulted in worse learning outcomes later. Braga et al. (2011) find that students who evaluated their professors more highly in their randomly assigned compulsory courses tended to have worse outcomes in subsequent courses. Love and Kotchen (2010) develop a theoretical model that shows how grade inflation distorts behavior for both faculty and students, and how both more emphasis on research productivity and on student course evaluations in promotion decisions can reduce teaching effort.

more, and both students' grades and their evaluations of the professor reflect this.[10] Here we examine what happened to student evaluations of their professors when there is an exogenous reduction in grades created by the policy.

At Wellesley, students submit evaluations electronically and their ability to see their grades in a timely fashion is tied to submitting an evaluation during a specified period.[11] Thus, nearly 100 percent of students submit evaluations. Over the period studied, students gave their professors ratings on a four-point scale: "Strongly recommend"; "Recommend"; "Neutral"; and "Do not recommend." (There is also a qualitative component to the evaluation, but we only have access to the numeric component.) Students are generally well satisfied with their professors at Wellesley, with over 60 percent "strongly recommending" their professors (see online Appendix Table 3).

Table 4 shows the impact of the policy on student evaluations of their professors. The data are at the professor-by-course-by-semester level. Column 1 uses the average rating for the professor as dependent variable;[12] column 2 uses the percent of the students in the course that "strongly recommended" the professor as the outcome; outcomes are defined analogously for columns 3–5.

Students' evaluations of their professors fell after the policy in those departments that were affected by the grade cap. On average, students' ratings on the four-point scale fell by a statistically significant 0.11. The percent of students "strongly recommending" their professors fell by about 5 percentage points. There were statistically significant increases in the "neutral" and "do not recommend categories." In particular, the share of student evaluations in the "do not recommend" category rose from about 5 percent to slightly over 7 percent in the treated departments. In short, the results strongly indicate that students were less pleased with their instructors when the grading policy lowered average grades.

[10] In the online Appendix available with this paper at http://e-jep.org, online Appendix Table 5 presents more detailed findings on students' ratings of their professors at Wellesley in the pre-policy era. For example, holding various observable characteristics constant, visiting faculty receive lower student ratings; male and female faculty are rated without significant difference, professors with more experience at Wellesley receive higher teaching evaluations (which is not surprising since one does not get the opportunity to continue working at Wellesley if evaluations are poor); and students rate professors more highly in 200-level courses than in 100-level courses (again, unsurprising because generally speaking students have more choice over their 200-level courses). Students in general give higher scores to their professors in the humanities than they do in the sciences and the social sciences. Courses that have more withdrawals, and more students electing the credit/non option, have worse evaluations of the professors. Higher average course grade point average is associated with better professor evaluations.

[11] The privacy of these responses is taken very seriously, and a student's evaluation cannot be linked to the individual's academic record.

[12] In Wellesley's evaluation metric a "1" is the best score, but we have flipped the scoring system in column 1 because it is more intuitive that a negative coefficient indicates a lower rating.

Table 4
Impact of Policy on Student Evaluations

	Rating (1)	*% Strongly recommend* (2)	*% Recommend* (3)	*% Neutral* (4)	*% Do not recommend* (5)
*PostPolicy*Treated*	−0.111***	−0.050***	0.010	0.018***	0.022***
	(0.030)	(0.015)	(0.011)	(0.006)	(0.007)
Age (in years)	−0.001***	−0.005***	0.001	0.002***	0.002***
	(0.003)	(0.001)	(0.001)	(0.001)	(0.000)
Male Faculty	0.024	0.009	0.001	−0.008	−0.004
	(0.029)	(0.018)	(0.009)	(0.006)	(0.004)
Years Worked	0.009***	0.004**	−0.000	−0.001**	−0.002***
	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)
Non-Tenure-Track	−0.011	0.004	−0.010	−0.003	0.009
	(0.058)	(0.032)	(0.019)	(0.016)	(0.009)
Visiting Faculty	−0.160***	−0.075***	0.016	0.034***	0.025***
	(0.046)	(0.025)	(0.012)	(0.010)	(0.007)
Tenured Faculty	−0.056	−0.020	−0.001	0.006	0.015**
	(0.033)	(0.018)	(0.012)	(0.009)	(0.006)
200 level	0.069***	0.032***	−0.009	−0.010*	−0.014***
	(0.021)	(0.012)	(0.007)	(0.005)	(0.003)
Class size/10	−0.001	−0.003	0.003*	0.000	−0.001
	(0.001)	(0.000)	(0.000)	(0.002)	(0.000)
Pretreatment mean	3.410	0.608	0.245	0.097	0.051
Observations	5,378	5,378	5,378	5,378	5,378
R^2	0.145	0.133	0.132	0.092	0.108
Department fixed effects	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes	Yes

Source: Authors using anonymized professor-course-semester-level data from Wellesley College.
Notes: "Rating" is calculated on a scale of 1 through 4 using 4 for "Strongly Recommend," 3 for "Recommend" etc. Pretreatment means are shown in the table for comparison. Controls for racial/ethnic characteristics of faculty were included but not shown in the table. Column 2 uses the percent of the students in the course that "strongly recommended" the professor as the outcome; outcomes are defined analogously for columns 3–5.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Discussion and Conclusion

Grades provide information. Grade inflation, and the resulting compression of grades near the top, reduces the information content of that signal, especially when the degree of grade inflation and compression varies across departments. This distortion in information may lead to a misallocation of resources on scales small and large. On the smaller side, Wellesley (like other institutions) has tended to allocate academic support services where bigger gaps in grades are evident, because those gaps are taken to reflect gaps in learning. When the policy lowered

grades in the humanities and non-economics social sciences, gaps in grades by group appeared to be more similar across departments. If the increased gaps mean that administrators now have better information about student learning in different departments, academic support services can be targeted more efficiently to improve learning outcomes.[13]

On a larger scale, grades provide information to students about where their talents lie. If students interpret lower grades as an indication that they are not good at a given subject, then they may shy away from low-grading disciplines to their detriment. The results here indicate students' choices about courses and majors are sensitive to grades. Research has shown that students base decisions to take a subsequent course in a discipline more on absolute grades than relative rankings (and relative skill) in the class (Sabot and Wakeman-Linn 1991). Goldin (2013, or see discussion by Rampell 2014) points out that women's choice of major may be especially sensitive to grades; she also points out that choices about academic major can have profound consequences for future earnings. Thus, the costs of the distortion in information created by grade inflation may be disproportionately born by some groups.[14]

One hope of those who advocate addressing grade inflation is that such policies might encourage more students to enter the science, technology, engineering, and mathematics fields. Although we find an effect of grades on major choice, the switching of majors appears to be more of a function of switching between types of social sciences than from humanities and social sciences to the sciences fields. It is worth noting that economics is currently the biggest major at Wellesley, with nearly 20 percent of students declaring it as (at least one of) their major(s). Of course, Wellesley's policy did not force average grades to be equal across departments, and the high-grading departments continue to have substantially higher grades; bigger changes in grades might induce bigger changes in choice of major.

Given the pressures on individual faculty members to have high student evaluations, and the pressures on departments to maintain enrollments, reducing grade inflation and compression requires action at the institutional level. Even so, devising policies is fraught with potential unintended consequences. Institutions have usually followed one of two types of policies: implementing grade targets as at Wellesley,[15] or trying to give students, graduate schools, and employers more information about the content of grades. In 1996, Cornell University adopted the latter type of policy by making public the median grades in each course. As Bar, Kadiyali, and Zussman

[13] If the new bigger gaps between groups within treated departments reflect an adverse reaction to increased pressure over grades—as the literature on stereotype threat might suggest—then the interpretation of the bigger gaps post-policy is quite different, and less sanguine.

[14] Although this comment suggests that the responses to the grading policy that we find at Wellesley, a women's college, may be larger than they would be in other settings, the broader point holds: students who are sensitive to grades in their choice of academic major will distort their choices if there is differential grade inflation and compression across disciplines.

[15] Princeton University also has a targeting policy, which is that no more than 35 percent of grades in a department should be an A.

(2009) discussed in this journal, the policy accelerated grade inflation at Cornell as a higher fraction of students chose to take more leniently graded courses.

Any institution that attempts to deal with grade inflation on its own must consider the possibility of adverse consequences of this unilateral disarmament. At Wellesley College, for example, prospective students, current students, and recent alums all worry that systematically lower grades may disadvantage them relative to students at other institutions when they present their grades to those outside the college. They point to examples of web-based job application systems that will not let them proceed if their GPA is below at 3.5. The economist's answer that firms relying on poor information to hire are likely to fare poorly and to be poor employers in the long run proves remarkably uncomforting to undergraduates. These concerns lead to pressure to reverse the grade policy. If grade inflation is a systemic problem leading to inefficient allocation of resources, then colleges and universities may wish to consider acting together in response.

■ *We appreciate the support and feedback from administrative staff, faculty, and students at Wellesley College. We are grateful to the administration and institutional research staff for making the data available. We thank participants in seminars at Wellesley College, Rutgers University, Brandeis University, and participants in the NBER Education meetings for helpful comments. We thank YeJi Kee and Ashley Longseth for helpful research assistance. The research received support from Wellesley College, in the form of access to anonymized data on students and faculty. Administrators of Wellesley College had the right to review the manuscript before publication.*

References

Achen, Alexandra C., and Paul N. Courant. 2009. "What Are Grades Made Of?" *Journal of Economic Perspectives* 23(3): 77–92.

Babcock, Philip. 2010. "Real Costs of Nominal Grade Inflation? New Evidence from Student Course Evaluations." *Economic Inquiry* 48(4): 983–96.

Bar, Talia, Vrinda Kadiyali, and Asaf Zussman. 2009. "Grade Information and Grade Inflation: The Cornell Experiment." *Journal of Economic Perspectives* 23(3): 93–108.

Braga, Michela, Marco Paccagnella, and Michele Pellizzari. 2011. "Evaluating Students' Evaluations of Professors." Bank of Italy Temi di Discussione (Working Paper) 825. Available at SSRN: http://ssrn.com/abstract=2004361.

Carrell, Scott E., and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118(3): 409–32.

Chan, William, Li Hao, and Wing Suen. 2007. "A Signaling Theory of Grade Inflation." *International Economic Review* 48(3): 1065–90.

Franz, Wan-Ju Iris. 2010. "Grade Inflation Under the Threat of Students' Nuisance: Theory and Evidence." *Economics of Education Review* 29(3): 411–22.

Goldin, Claudia. 2013. "Can 'Yellen Effect'

Attract Young Women to Economics?" BloombergView, October 14. http://www.bloombergview.com/articles/2013-10-14/can-yellen-effect-attract-young-women-to-economics-.

Love, David A., and Matthew J. Kotchen. 2010. "Grades, Course Evaluations, and Academic Incentives." *Eastern Economic Journal* 36(2): 151–63.

Pressman, Steven. 2007. "The Economics of Grade Inflation." *Challenge* 50(5): 93–102.

Rampell, Catherine. 2014. "Women Should Embrace the B's in College to Make More Later." *Washington Post,* March 10. http://www.washingtonpost.com/opinions/catherine-rampell-women-should-embrace-the-bs-in-college-to-make-more-later/2014/03/10/1e15113a-a871-11e3-8d62-419db477a0e6_story.html?hpid=z3.

Rojstaczer, Stuart, and Christopher Healy. 2010. "Grading in American Colleges and Universities." *Teachers College Record,* March 4.

Sabot, Richard, and John Wakeman–Linn. 1991. "Grade Inflation and Course Choice." *Journal of Economic Perspectives* 5(1): 159–70.

Wongsurawat, Winai. 2009. "Does Grade Inflation Affect the Credibility of Grades? Evidence from US Law School Admissions." *Education Economics* 17(4): 523–34.

Zangenehzadeh, Hamid. 1988. "Grade Inflation: A Way Out." *Journal of Economic Education* 19(3): 217–26.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 205–216

The Research Productivity of New PhDs in Economics: The Surprisingly High Non-Success of the Successful[†]

John P. Conley and Ali Sina Önder

Economics PhD programs are primarily designed to produce research economists. There is little or no focus on training students to suit the needs of business or industry (Siegfried and Stock 1999). Our experience suggests that most students, especially at the better programs, enter graduate school planning to seek academic jobs, or at any rate, jobs that require research. We would also argue that students have a more-or-less common preference ordering over departments. In general, a student admitted to MIT or Princeton is unlikely to choose to go to Duke or Ohio State instead. Thus, the top programs have first pick over applicants to graduate programs in any given year, and this should concentrate both quantity and scholarly ambition in a strongly top-heavy way in programs specifically designed to train researchers.

To explore the effect of this dynamic, we construct a panel dataset consisting of two parts: a census of PhD recipients from academic institutions in the US and Canada who received their economics PhDs between 1986 and 2000, and a complete record of the journal publications of these individuals for the years 1985 to 2006 in the hundreds of journals listed in EconLit. This allows us to look at the distribution of research output of the PhDs from individual departments and also to compare research outcomes across programs of various ranks.

Our evidence shows that only the top 10–20 percent of a typical graduating class of economics PhD students are likely to accumulate a research record that

■ *John P. Conley is Professor of Economics, Vanderbilt University, Nashville, Tennessee. Ali Sina Önder is a Lecturer, University of Bayreuth, Bayreuth, Germany and an Affiliated Researcher, Uppsala Center for Fiscal Studies, Uppsala, Sweden. Their email addresses are j.p.conley@vanderbilt.edu and ali-sina.oender@uni-bayreuth.de.*

[†] To access the data Appendix, visit
http://dx.doi.org/10.1257/jep.28.3.205

doi=10.1257/jep.28.3.205

might lead to tenure at a medium-level research university. Perhaps the most striking finding from our data is that graduating from a top department is neither necessary nor sufficient for becoming a successful research economist. Top researchers come from across the ranks of PhD-granting institutions, and lower-ranked departments produce stars with some regularity, although with lower frequency than the higher-ranked departments. Most of the graduates of even the very highest-ranked departments produce little, if any, published research. Indeed, we find that PhD graduates of equal percentile rank from certain lower-ranked departments have stronger publication records than their counterparts at higher-ranked departments. In our data, for example, Carnegie Mellon's graduates at the 85th percentile of year-six research productivity outperform 85th percentile graduates of the University of Chicago, the University of Pennsylvania, Stanford, and Berkeley.

In this paper, we lay out and describe the patterns we find in publication in the first six years after a PhD. We conclude by discussing some implications for several groups: 1) undergraduate students considering the possibility of seeking a PhD degree in economics; 2) those who administer, teach, and advise in PhD programs in economics; 3) and the committees responsible for recruiting and hiring new assistant professors, who must make decisions about whether to go after a higher-ranked candidate from a lower-ranked PhD program or a lower-ranked candidate from a higher-ranked PhD program.

Publications Patterns of New PhD Economists

We start with a census of 14,299 economics PhD recipients from 154 academic institutions in the US and Canada who graduated between 1986 and 2000 compiled by the American Economic Association (AEA) and connect this to an EconLit database with 368,672 papers published between 1985 and 2006 in 1,113 peer-reviewed journals (including conference volumes to the extent that these are captured in EconLit). Pooling all years, 7,154 economics PhDs could be detected as authors of the 48,938 papers in EconLit. This study follows up on Conley, Crucini, Driskill, and Önder (2013), in which we examined recent trends in publication rates of young scholars in economics, and we refer readers to that paper for more details regarding the nature and origin of these data.

Next, we take each of the top-30 economics departments, combine all their graduates from 1986 to 2000 into a single sample, and look at total research productivity at the end of the sixth year after graduation. We did the same for graduates of non-top-30 departments as one combined group. We use a department ranking developed by Coupé (2003) based on faculty research productivity to choose the top-30 group. Of course, which departments are "top 30" is open to debate, and regardless of how the ranking is established, some departments are likely to have moved in and out of this group over the 15 year interval we study. Given this, it would be better to think of our "top 30" departments as representative of "top departments" in general.

Raw counts of publications are imperfect measures of the research productivity of individual scholars because of the variation in the quality of those publications. We therefore use journal quality indexes from Kalaitzidakis, Mamuenas, and Stengos (2003), which are appropriate for the time frame of our study, to convert each raw publication into a number of *American Economic Review*–equivalent papers which we refer to as "*AER* papers." To give some sense of the weights we used, the following publication lists (and it may be useful to think of these as entries on the curriculum vitae of tenure candidates) are all roughly equivalent to one *AER* paper: (a) one paper in the *American Economic Review* or *Econometrica;* (b) one and one half papers in the *Journal of Political Economy* or *Quarterly Journal of Economics*; (c) two papers in the *Review of Economic Studies, Journal of Econometrics*, *Econometric Theory*, or *Journal of Economic Theory*; (d) three papers in the *Journal of Monetary Economics* or *Games and Economic Behavior*; (e) four papers in the *European Economic Review*, *Review of Economics and Statistics*, *International Economic Review,* or *Economic Theory*; (f) five papers in the *Economic Journal*, *Journal of Public Economics*, or *Economics Letters*; or (g) six to ten papers in high-quality field journals. We also adjust for the number of coauthors on a given paper. Thus, if a PhD in our sample publishes a paper with C coauthors in a journal with a quality index of Q relative to the *AER*, then the graduate is credited with Q/C *AER* papers. We should note that in Conley, Crucini, Driskill, and Önder (2013) we looked at alternative journal rankings and also dispensed with discounting for coauthorship. The results are qualitatively robust to such variations.

Table 1 shows the number of *AER* papers that appear on the (constructed) CVs of graduates of each department at the end of their sixth year after graduation by productivity percentile. For example, Harvard graduates in the 95th percentile of research productivity relative to their classmates published the equivalent of 2.36 *AER* papers in this period. We order the table using the Coupé (2003) ranking because this gives a kind of "prior" about how students ought to perform, while the rest of the columns give a sort of "posterior" of actual performance.

Table 1 reveals a rapid drop-off in research productivity of PhD graduates regardless of department as class rank decreases. At Harvard, for example, a student has to be in the 85th percentile or above to be likely to publish even a single *AER* paper in six years. The median Harvard graduate publishes only .04 *AER* papers. On the other hand, the 90th percentile of graduates of Carnegie Mellon or the University of California, San Diego, and the 80th percentile of Rochester graduates can also be expected to have one *AER* paper or more by year six. Going farther down this table, we see that a 95th percentile graduate of a typical non-top-30 department has a stronger publication record than the 70th percentile graduate of Harvard, Chicago, U Penn, Stanford, or Yale, or an 80th percentile graduate of Berkeley, Michigan, NYU, UCLA, or Columbia.

Research productivity in economics is generally highly concentrated. In Conley, Crucini, Driskill, and Önder (2013), we find that the top 1 percent of publishing research economists across the whole sample produce 13 percent of all (quality-adjusted) research output, and the top 20 percent of publishing economists produce 80 percent of it. What is most surprising in this present analysis is that this

Table 1
Number of *AER*-Equivalent Publications of Graduating Cohorts from 1986 to 2000

	Percentiles of graduates' AER*-equivalent publications 6 years after PhD*									*Average cohort size*	*Publishing grads (%)*
	99th	*95th*	*90th*	*85th*	*80th*	*75th*	*70th*	*60th*	*50th*		
Harvard	4.31	2.36	1.47	1.04	0.71	0.41	0.30	0.12	0.04	30.5	66.3
Chicago	2.88	1.71	1.04	0.72	0.51	0.33	0.19	0.06	0.01	27.3	59.4
U Penn	3.17	1.52	1.01	0.60	0.40	0.27	0.22	0.06	0.02	19.3	59.5
Stanford	3.43	1.58	1.02	0.67	0.50	0.33	0.23	0.08	0.03	24.7	67.9
MIT	4.73	2.87	1.66	1.24	0.83	0.64	0.48	0.20	0.07	25.5	70.0
UC Berkeley	2.37	1.08	0.55	0.35	0.20	0.13	0.08	0.04	0.02	28.0	62.4
Northwestern	2.96	1.92	1.15	0.93	0.61	0.47	0.30	0.14	0.06	10.1	65.8
Yale	3.78	2.15	1.22	0.83	0.57	0.39	0.19	0.08	0.03	15.7	64.8
U MI, Ann Arbor	1.85	0.77	0.48	0.29	0.17	0.09	0.05	0.02	0.01	19.1	54.0
Columbia	2.90	1.15	0.62	0.34	0.17	0.10	0.06	0.01	0.01	17.4	54.8
Princeton	4.10	2.17	1.79	1.23	1.01	0.82	0.60	0.36	0.19	16.2	76.1
UCLA	2.59	0.89	0.49	0.26	0.14	0.06	0.04	0.02	0	17.9	48.5
NYU	2.05	0.89	0.34	0.20	0.07	0.03	0.02	0.01	0	11.7	46.0
Cornell	1.74	0.65	0.40	0.23	0.12	0.07	0.05	0.02	0.01	17.3	57.9
U WI, Madison	2.39	0.89	0.51	0.31	0.20	0.11	0.06	0.03	0.01	25.0	60.3
Duke	1.37	1.03	0.59	0.49	0.23	0.19	0.11	0.05	0.02	7.8	59.8
Ohio State U	0.69	0.41	0.13	0.07	0.04	0.02	0.02	0.01	0	15.9	47.9
U Maryland	1.12	0.37	0.23	0.10	0.07	0.05	0.03	0.01	0.01	13.5	56.2
Rochester	2.93	1.94	1.56	1.21	1.14	0.98	0.70	0.34	0.17	8.7	78.5
U TX, Austin	0.92	0.53	0.21	0.06	0.05	0.02	0.01	0	0	10.3	38.3
Minnesota	2.76	1.20	0.68	0.46	0.29	0.21	0.12	0.04	0.01	22.2	59.5
U IL, Urbana-Ch	1.00	0.38	0.21	0.10	0.06	0.04	0.03	0.01	0.01	26.4	54.8
UC Davis	1.90	0.66	0.42	0.27	0.12	0.08	0.05	0.02	0.01	6.2	53.8
Toronto	3.13	1.85	0.80	0.61	0.29	0.19	0.15	0.07	0.03	6.4	64.6
British Columbia	1.51	1.05	0.71	0.60	0.52	0.45	0.26	0.22	0.11	4.5	73.1
UC San Diego	2.29	1.69	1.17	0.88	0.74	0.60	0.46	0.30	0.18	6.1	78.3
U Southern CA	3.44	0.34	0.14	0.09	0.03	0.02	0.02	0.01	0	4.9	43.8
Boston U	1.59	0.49	0.21	0.08	0.05	0.02	0.02	0	0	12.5	41.0
Penn State U	0.93	0.59	0.25	0.12	0.08	0.06	0.02	0.01	0.01	7.1	51.4
Carnegie Mellon	2.50	1.27	1.00	0.86	0.71	0.57	0.52	0.21	0.09	2.0	66.7
Non-Top-30	1.05	0.31	0.12	0.06	0.04	0.02	0.01	0	0	16.8	40.1

Source: Based on the authors own calculations using the data described in the paper.
Note: We order the table using the Coupé (2003) ranking of economics departments.

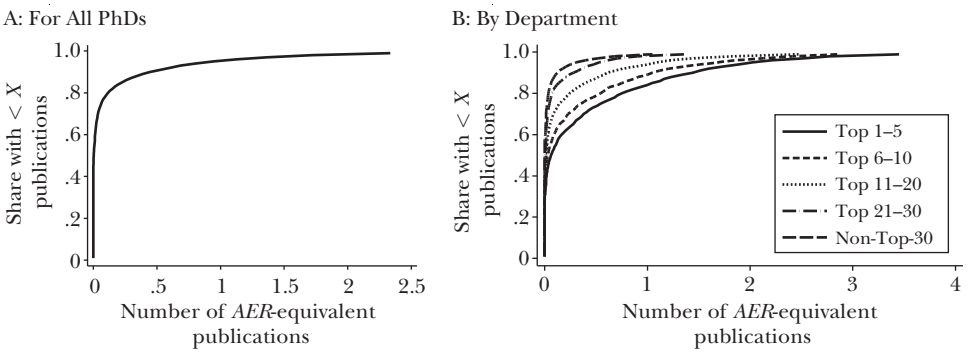
pattern is mirrored at each individual department. Thus, even though the top five or ten have their pick of applicants each year, they still produce only a few winners in the research game.

Figures 1A and 1B can help to visualize the quick drop-off in productivity documented in Table 1. The cumulative distribution of publications for all PhDs is shown in Figure 1A: when we rank all PhDs based on their productivity (independent of their alma mater), 80 percent of all PhDs accumulate about 0.2 *AER*-equivalent papers or less within six years after graduation, and about 90 percent of PhDs do not reach 0.5 *AER*-equivalent papers within that time.

Figure 1

Cumulative Distribution Functions of Number of PhDs' Publications

(share with less than X*-AER-equivalent publications six years after graduation)*

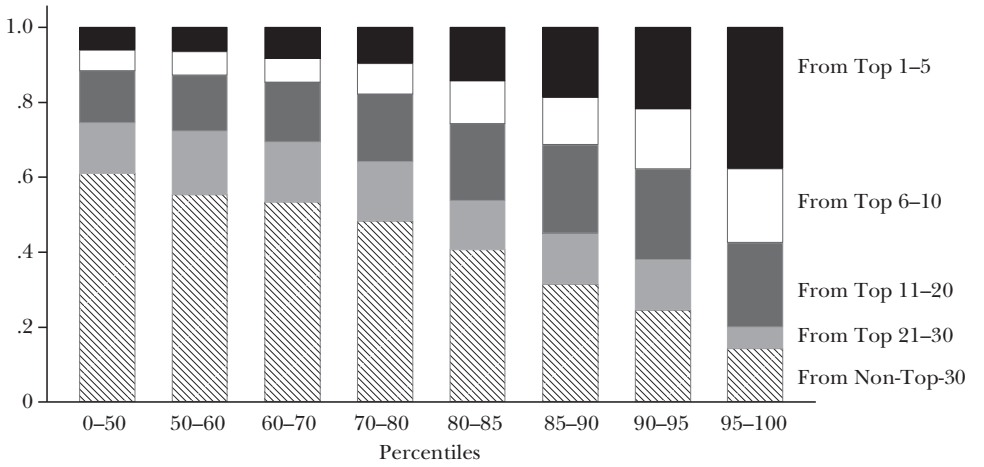


Source: Based on the authors own calculations using the data described in the paper.
Note: Figure 1B shows separate cumulative distributions by department tiers where tiers are defined by the productivity of graduates.

Figure 1B shows separate cumulative distributions by department tiers where tiers are defined by the productivity of graduates. We re-rank departments in Table 1 based on the productivity of their 95th percentile graduate, which enables us to compare departments based on how much more (or less) productive their PhDs are compared to others. When ranked in this way, the top five departments in the first tier are MIT, Harvard, Princeton, Yale, Rochester, and the five departments in the next tier are Northwestern; University of Toronto; Chicago; University of California, San Diego; and Stanford. The next ten departments and the ten after that are the third and fourth tiers, respectively. Non-top-30 departments constitute the fifth tier. Comparing cumulative distribution functions of these tiers portrays a clear pattern: 20 percent of PhDs from the top tier departments have at least one *AER*-equivalent paper, while the ratio for third tier (top 11–20) departments is about 10 percent, and it drops to about 1–2 percent for departments outside top-20 departments. An even more striking observation is that about 40 to 60 percent of PhDs in each tier do not have any publications.

At the very top end, consider "superstar" graduates who manage to publish 2.5 or more *AER*-equivalent papers at year six (remember, this could easily mean one paper in *AER*, then a number of other papers in highly ranked journals). By our measure, the top one or two graduates from Harvard or MIT will typically meet this standard, along with the top graduate from Stanford, Yale, or Princeton, if these departments are having a good year. Once every other year, Chicago, U Penn, and Minnesota should produce a superstar. Other departments will do so with less frequency. Of course, others may become publication superstars later in their careers, but only seven or eight in a given graduation cohort are likely to reveal themselves as such by the sixth year after receiving their PhD.

Figure 2
Department Tiers' Share in Productivity Percentiles



Source: Based on the authors own calculations using the data described in the paper.

Consider instead a standard of .6 *AER*-equivalent papers—which is, remember, equal to roughly 4–6 papers in well-regarded field journals in the six years after graduation. This level of research productivity is on average reached by the 80th percentile and above of PhD graduates from Harvard, MIT, Northwestern, Yale, Princeton, Rochester, University of California, San Diego, and Carnegie Mellon. To put it another way, 80 percent or more of the graduates of the following universities will not have .6 *AER*-equivalent papers at the end of six years: Chicago; U Penn; Stanford; UC Berkeley; University of Michigan, Ann Arbor; Columbia; UCLA; NYU; Cornell; University of Wisconsin, Madison; Duke; Ohio State; University of Maryland; University of Texas, Austin; Minnesota; University of Illinois, Urbana–Champaign; University of Toronto; University of British Columbia; University of Southern California; Boston University; and Penn State.

Figure 2 offers another way of visualizing how research productively is distributed over departmental quality tiers. For example, the figure shows that about 40 percent of students who end up above the 95th percentile of research productivity at year six will come from the top five departments, while about 10 percent will come from non-top-30 departments.

Although regularities are captured when departments are aggregated into various tiers, some of the most interesting results arise from individual comparisons. One such finding is that a few departments perform relatively better at producing successful students who are not in the top percentiles and so have a longer tail of relatively productive students than more highly ranked departments. Put another way, it is not the case that better-ranked schools always outperform all lesser-ranked

Table 2

Department Rankings based on Graduating Cohort's Publication Performance at Different Percentiles of the Graduating Class (1986–2000)

	Coupé ranking	*Ranking at percentile of the class*								
		99th	*95th*	*90th*	*85th*	*80th*	*75th*	*70th*	*60th*	*50th*
Harvard	1	2	2	4	4	5	8	6	8	8
Chicago	2	12	8	8	9	10	10	12	12	17
U Penn	3	7	11	10	13	12	12	10	13	14
Stanford	4	6	10	9	10	11	11	9	9	10
MIT	5	1	1	2	1	3	3	4	6	6
UC Berkeley	6	17	15	17	16	17	16	16	15	13
Northwestern	7	9	6	7	5	7	6	7	7	7
Yale	8	4	4	5	8	8	9	11	10	11
U MI, Ann Arbor	9	21	21	20	19	18	19	21	20	23
Columbia	10	11	14	15	17	19	18	18	22	20
Princeton	11	3	3	1	2	2	2	2	1	1
UCLA	12	14	19	19	21	20	22	22	21	26
NYU	13	19	20	23	23	24	26	26	27	30
Cornell	14	22	23	22	22	21	21	19	19	15
U WI, Madison	15	16	18	18	18	16	17	17	17	19
Duke	16	25	17	16	14	15	15	15	14	12
Ohio State U	17	31	27	30	29	29	27	27	24	28
U Maryland	18	26	29	25	25	25	24	23	25	21
Rochester	19	10	5	3	3	1	1	1	2	3
U TX, Austin	20	30	25	27	31	27	29	31	31	27
Minnesota	21	13	13	14	15	14	13	14	16	18
U IL, Urbana-Ch	22	28	28	26	26	26	25	24	26	24
UC Davis	23	20	22	21	20	22	20	20	18	16
Toronto	24	8	7	12	11	13	14	13	11	9
British Columbia	25	24	16	13	12	9	7	8	4	4
UC San Diego	26	18	9	6	6	4	4	5	3	2
U Southern CA	27	5	30	29	27	31	28	28	28	25
Boston U	28	23	26	28	28	28	30	29	30	29
Penn State U	29	29	24	24	24	23	23	25	23	22
Carnegie Mellon	30	15	12	11	7	6	5	3	5	5
Non-Top-30		27	31	31	30	30	31	30	29	31

Source: Based on the authors own calculations using the data described in the paper.
Note: The first column shows the Coupé (2003) ranking of economics departments.

schools in the sense of first order stochastic dominance. Table 2 gives a set of departmental rankings based on the productivity of different percentiles of the graduating class. Thus, at the 95th percentile of students, MIT graduates are more productive at year six than those of any other department. If we look at students in the 70th percentile, however, MIT's ranking drops to fourth.

Table 2 shows that some departments like Harvard, MIT, Yale, and to a lesser extent Chicago and U Penn follow a downward trend in these rankings from left to right across the percentiles. That is, they do better at training top students than middle- or lower-level students in a relative sense. Other departments, such as

Rochester, University of British Columbia, University of California, San Diego, and Carnegie Mellon, do not compete with the top departments in producing the very top research scholars, but are able to turn out lower-ranked students who dominate the similarly ranked graduates at better-ranked departments. For example, Rochester is third-best at producing students at the 90th and 85th percentile, and as we look across to still lower percentiles, it mostly trades the one and two spots with Princeton.

Discussion

If the objective of graduate training in top-ranked departments is to produce successful research economists, then these graduate programs are largely failing. Only a small percentage of economics PhDs manage to produce a creditable number of publications by their sixth year after graduation. Even at the top five departments, it would be hard to argue that the bottom half of their students are successful in terms of academic research. The number of *AER*-equivalent papers of the median at year six is below 0.1 in all cases and is in fact zero in most. At the majority of the departments ranked in the top ten in conventional rankings (such as Coupé 2003), 60 percent of their students fail to meet this 0.1 *AER*-equivalent standard, and for the majority of the PhD graduates of the top 30 departments, 70 percent fail. A tenure standard of 0.1 *AER*-equivalent papers is roughly equal to publishing one paper in a second-tier field journal over six years. This record would not be enough to count as "research-active" in most departments, much less to result in tenure. Even from the highest-ranked departments, very few graduates prove to be stars. Lower-ranked departments, on the other hand, produce stars with some regularity, although not as often as top departments.

For graduate students in economics (and also potential graduate students), the message is that becoming a successful research economist is difficult. The good news is that one does not have to go to a top department in order to become a successful research economist. The bad news is that wherever one goes, only the very best of each class is likely to find academic success as defined by research publications.

Indeed, to become a tenured professor of economics one must cross many hurdles. Admission to an economics PhD program is difficult: most well-ranked departments receive several hundred highly competitive applications for entering classes that generally number between 10 and 30. Many of those admitted to a graduate program will ultimately fail to complete their degree. For example, Stock, Siegfried, and Finegan (2011) find that graduation rates from economics PhD programs are on the order of 30 percent by the fifth year after admission, rising to around 60 percent by the eighth year. (There is wide variability, but the higher-ranked programs seem to have higher graduation rates in general.) Even for those who do complete the PhD, the likelihood of ultimately accumulating a research record that might gain tenure at a top-100 department (much less a top-30 or top-10 department) is not very great. Thus, students thinking about applying to

PhD programs in economics would be well advised to have "Plan B's" for every stage of the journey—including the possibility of not being accepted into a PhD program, the possibility of not completing the program, the possibility of not finding a suitable academic job, and the possibility of not receiving tenure. We hasten to add that there are many rewarding and worthwhile nonresearch and nonacademic career paths open to those who obtain masters or doctorate degrees in economics, and many students discover, either while in graduate school or during their untenured years, that they actually prefer these sorts of jobs to the academic life.

These results also raise some concerns for those of us who sit on admission committees and teach in graduate programs. To be admitted to a top PhD program in economics, an applicant has to have great grades, near-perfect test scores, strong and credible recommendations, and package these credentials in a way that stands out to the admission committee. Thus, successful candidates must be hardworking, intelligent, well-trained, savvy, and ambitious. Why is it that the majority of these successful applicants, who apparently did all the right things up to the time they arrived at graduate school and even managed to complete their PhDs, have such unimpressive careers as researchers? Are we failing the students or are the students failing us?

Three possible answers suggest themselves. First, perhaps what makes a successful research economist is not well-measured by tests and grades. For example, along with being hardworking, well-trained, and intelligent, a successful career might also require attributes like being creative, self-motivated, thick-skinned, or having an aptitude for academic networking. Of course, such attributes are quite difficult to discern in the application process.

A second possible answer is that there might be a virtuous circle in professional success. If a new graduate (given an underlying level of fundamental quality) gets a good first job, is well mentored and fostered by new colleagues, and has early success in publishing, that new graduate may be more likely to have more papers accepted by good journals in the future. Oyer (2006) discusses learning-on-the-job aspects in academic careers and establishes a causal relationship between landing a research-oriented first job after the PhD, and life-cycle publication productivity. Luck may also play a role, as some new PhD economists find that their subfield or topic offers more fertile ground for additional research than others.

A third possibility is that both students and professors in certain departments may find themselves playing a positional game. The faculty will attempt to identify the top students in an entering class, give them more time and attention, and suggest better projects to them. In turn, the students identified in this way may work harder to preserve their position. The pattern of only a few high research performers followed by a very quick drop-off would be consistent with this hypothesis (for some interesting speculations about this dynamic in many occupations, see Gladwell 2013).

When we started this project, one issue in the forefront of our thinking was a common problem faced by many hiring committees. The most highly-ranked departments in economics are able to choose their new assistant professors from among the top graduates of other top departments. However, at lesser departments, there is always a debate about whether it is better to hire lower-ranked graduates

from top-ranked departments of economics, or the best graduates from lower-ranked departments. Our conclusion is that it is indeed worthwhile for lower-ranked schools to look outside the top-ranked departments for new hires, though only at the top students of such programs in general.

Our evidence is based on the accumulated record after six years, which unfortunately is not the information available to the hiring committee at the time the hire is made. Some evidence suggests that hiring committees may not be very accurate at forming expectations of quality when a new PhD hits the job market. Smeets, Warzynski, and Coupé (2006) explore the efficiency of the academic job market in matching students to positions. They study the 1992 and 1993 PhD cohorts from the 26 best graduate schools and discover that the matching of quality students to quality first jobs is not as tight as one might hope. They further show there is substantial, mostly downward, movement from the first to the final job hold, and overall, the research productivity of students who get first jobs of various qualities does not differ as starkly as we see in Table 1. This finding suggests that the students who are identified as top graduates in a given year (and who get top jobs as result) might not line up with the students who end up being the most productive six years later. For example, our data show that publishing a paper before graduation is uncorrelated with the productivity over the six-year probationary period before a tenure decision.

Students put tremendous efforts into acquiring the credentials that allow them to gain admission to graduate school. Graduate schools, in turn, put tremendous effort into figuring which of these applicants are worthy of admission and then spend countless hours in their training and supervision. The hiring process takes weeks of thought and attention as recommendations are written, papers are read, candidates to be interviewed are identified, fly-outs are scheduled, and seemingly endless job talks are attended. These data suggest, but by no means prove, that our long-standing and expensive process may not be very effective. It may in fact be that many students and graduates have the potential for success, but realizing it is a matter of luck, position, or a having random but hard-to-measure endowment of something special.

An explanation might be that all parties are investing in lottery tickets, hoping for the prize. Students may overestimate their abilities and the overall odds that anyone succeeds, while admission and hiring committees think that they have a better than average insight into who will turn into a winner. Of course, some students and committees will be right in their estimations, but our data show that most of them will be wrong. We have no wish to depress all of these happy optimists. However, it does seem that there is substantial room to improve either our profession's mechanism for selecting who enters PhD programs in economics, or our method of training economics PhDs, or both.

■ *We thank Mario Crucini, Robert Driskill, and John Siegfried for their contributions to this work. We benefited also greatly from comments by JEP editors David Autor and Timothy Taylor.*

References

Conley, John P., Mario J. Crucini, Robert A. Driskill, and Ali Sina Önder. 2013. "The Effects of Publication Lags on Life-Cycle Research Productivity in Economics." *Economic Inquiry* 51(2): 1251–76.

Coupé, Tom. 2003. "Revealed Performances: Worldwide Rankings of Economists and Economics Departments, 1990–2000." *Journal of the European Economic Association* 1(6): 1309–45.

Gladwell, Malcolm. 2013. *David and Goliath: Underdogs, Misfits, and the Art of Battling Giants.* Little, Brown and Company.

Kalaitzidakis, Pantelis, Theofanis P. Mamuneas, and Thanasis Stengos. 2003. "Rankings of Academic Journals and Institutions in Economics." *Journal of the European Economic Association* 1(6): 1346–66.

Oyer, Paul. 2006. "Initial Labor Market Conditions and Long-Term Outcomes for Economists." *Journal of Economic Perspectives* 20(3): 143–60.

Siegfried, John J., and Wendy A. Stock. 1999. "The Labor Market for New Ph.D. Economists." *Journal of Economic Perspectives* 13(3): 115–34.

Smeets, Valérie, Frédéric Warzynski, and Tom Coupé. 2006. "Does the Academic Labor Market Initially Allocate New Graduates Efficiently?" *Journal of Economic Perspectives* 20(3): 161–72.

Stock, Wendy A., John J. Siegfried, and T. Aldrich Finegan. 2011. "Completion Rates and Time-to-Degree in Economics PhD Programs" (with comments by David Colander, N. Gregory Mankiw, Melissa P. McInerney, and James M. Poterba). *American Economic Review* 101(3): 176–93.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 217–236

The Economics of Fair Trade

Raluca Dragusanu, Daniele Giovannucci, and Nathan Nunn

Fair Trade is a labeling initiative aimed at improving the lives of the poor in developing countries by offering better terms to producers and helping them to organize. Although Fair Trade–certified products still comprise a small share of the market—for example, Fair Trade–certified coffee exports were 1.8 percent of global coffee exports in 2009—growth has been very rapid over the past decade.[1] Fair Trade coffee sales have increased from 12,000 tonnes in 2000 (Fairtrade International, 2012b, p. 24) to 123,200 tonnes in 2011 (Fairtrade International, 2012a, p. 41).

Whether Fair Trade can achieve its intended goals has been hotly debated in academic and policy circles. In particular, debates have been waged about whether Fair Trade makes "economic sense" and is sustainable in the long run. Development economist Paul Collier (2007, p. 163), in his book *The Bottom Billion*, writes: "They [Fair Trade–certified farmers] get charity as long as they stay producing the crops that have locked them into poverty." *The Economist* (2006) writes: "perhaps the most cogent objection to Fairtrade is that it is an inefficient way to get money to poor producers." Those on the other side of the debate argue that Fair Trade benefits farmers by providing higher incomes and greater economic stability. For example, Laura Raynolds (2009, p. 1083) writes that Fair Trade "offers farmers

[1] The statistic is for coffee sold as Fairtrade by Fair Trade International. Statistics on Fairtrade exports are from Fair Trade International, and those on total exports are from International Coffee Organization.

■ *Raluca Dragusanu is a PhD candidate in Business Economics, Harvard University, Cambridge, Massachusetts. Daniele Giovannucci is the cofounder and President of the Committee on Sustainability Assessment (COSA). Nathan Nunn is Professor of Economics, Harvard University, Cambridge, Massachusetts. The authors' email addresses are Raluca .dragusanu@gmail.com, DG@thecosa.org, and nnunn@fas.harvard.edu.*

http://dx.doi.org/10.1257/jep.28.3.217 doi=10.1257/jep.28.3.217

and agricultural workers in the global South better prices, stable market links and resources for social and environmental projects" and that it "provides consumers with product options that uphold high social and environmental standards."

The emergence of modern Fair Trade labels can be traced back to 1988, when a faith-based nongovernment organization from the Netherlands began an initiative that aimed to ensure that growers of crops in low-income countries were provided "sufficient wages." The organization created a fair trade label for their products. It was called Max Havelaar, after a fictional Dutch character who opposed the exploitation of coffee pickers in Dutch colonies. Over the next few years, the concept was replicated in other countries across Europe and North America, with a number of organizations emerging, such as TransFair and Global Exchange. In 1997, the various national labeling initiatives formed an umbrella association called Fairtrade International. A common Fair Trade Certification mark was launched in 2002 and there are several Fair Trade bodies operating today.

In 2012, Fairtrade International's largest adherent, Transfair USA, split from the organization to launch a parallel label, Fair Trade USA. One of the primary reasons for the division was the difference in beliefs about whether the Fair Trade label should only be available to small-scale producers. While Fairtrade International believes that certification should generally be restricted to small producers, Fair Trade USA feels that that large producers and plantations should also be certified.

"Fairtrade," the one-word form, is used by Fairtrade International for their certification mark and for references to their specific market. We use "Fair Trade" to refer to the general initiative and movement without reference to a particular certification.

Fair Trade attempts to achieve several goals; the primary and best-known is to provide prices that deliver a basic livelihood for producers. In addition, Fair Trade has a number of other goals, including longer-term buyer–seller relationships that facilitate greater access to financing for producers; improved working conditions; the creation and/or maintenance of effective producer or worker organizations; and the use of environmentally friendly production processes. A third-party certification process regularly checks that producers and suppliers adhere to a set of requirements whose purpose is to achieve these objectives. The Fair Trade label that is displayed on certified products is a signal to consumers that the product was produced and traded in accordance with these requirements.

Fairtrade is one of the many voluntary sustainability standards that have emerged. These standards share some common overlapping goals but each has its own focus and priorities. In addition to Fair Trade, other certification standards include Organic, Rainforest Alliance, and UTZ Certified, and there are similarly prominent labels for different products such as those of the Forest Stewardship Council, Marine Stewardship Council, Roundtable on Sustainable Palm Oil, and Global G.A.P.[2]

[2] For information on Organic, see the website of the International Federation of Organic Agricultural Movements at http://www.ifoam.org. For information on UTZ Certified, see https://www.utzcertified.org. For Rainforest Alliance, see http://www.rainforest-alliance.org. For Forest Stewardship Council, see https://us.fsc.org. For Marine Stewardship Council, see http://www.msc.org. For Global G.A.P., see http://www.globalgap.org. Also see the summary in Raynolds, Murray, and Heller (2007) and Potts et al. (2014).

Table 1
Number of Fairtrade Producers and Workers by Product

Product	*Number of producers and workers*
Coffee	580,200
Tea	258,100
Cocoa	141,800
Seed cotton	66,500
Flowers and plants	37,500
Cane sugar	37,200
Bananas	20,300
Fresh fruit	18,700
Nuts	14,300

Note: Data are from *Monitoring the Scope and Benefits of Fairtrade* (2012), fourth edition, Fairtrade International.

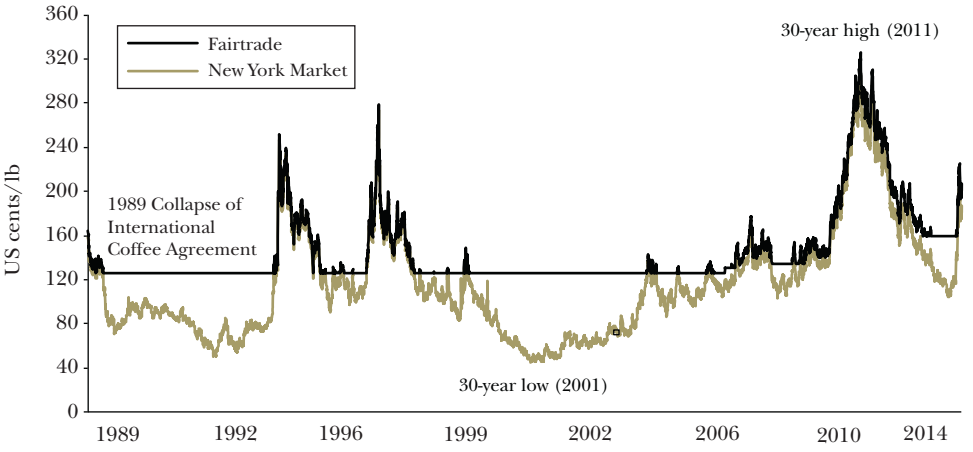
The aim of this article is to provide a critical overview of the economic theory behind Fair Trade, describing the potential benefits and potential pitfalls. We also provide an assessment of the empirical evidence of the impacts of Fair Trade to date. Because coffee is the largest single product in the Fair Trade market (see Table 1), our discussion here focuses on the specifics of this industry, although we will also point out some important differences with other commodities as they arise.

The Mechanisms of Fair Trade Standards

The stated goal of Fair Trade is to improve the living conditions of farmers and workers in developing countries. The specific mechanisms for achieving this goal are a combination of guidelines for price negotiation and requirements for certification, which we summarize here.

1) Price floor. The central characteristic of Fair Trade is the minimum price for which a Fair Trade–certified product can be sold to a Fair Trade buyer, which is intended to cover the average costs of sustainable production and meet a broadly determined living wage in the sector (originally set in accordance with the data of the International Coffee Organization). A Fair Trade buyer agrees to pay certified producers at least the minimum price when the world price is below this price. In all situations, producers and traders remain free to negotiate higher prices on the basis of quality and other attributes. By providing a guaranteed minimum price for products sold as Fair Trade, the price floor is intended to reduce the risk faced by growers. As we discuss in more detail below, there is no guarantee that all coffee that meets the certification requirements and is eligible to be sold as Fair Trade is indeed sold as such. Just producing and certifying a product does not guarantee that a buyer will purchase it as Fair Trade and provide the associated benefits and price. The relationship between the guaranteed minimum price

Figure 1
Comparison of Fairtrade and Market Prices for Coffee, 1989–2014



Source: © Fairtrade Foundation, adapted and used with permission.
Notes: NB Fairtrade Price = Fairtrade Minimum Price* of 140 cents/lb + 20 cents/lb Fairtrade Premium.** When the New York prices is 140 cents or above, the Fairtrade Price = New York price + 20 cents. The New York Price is the daily settlement price of the 2nd position Coffee C Futures contract at ICE Futures US.
* Fairtrade Minimum Price was increased on June 1, 2008, and April 1, 2011.
** Fairtrade Premium was increased on June 1, 2007, and April 1, 2011.

and the market price between 1989 and 2014 is shown in Figure 1. Although in recent years, the market price of coffee has usually been higher than the Fairtrade minimum price, data from the price crashes of the late 1990s and early 2000s indicate that the price floor can provide significant risk protection to farmers who sell their coffee as Fair Trade certified.

2) Fair Trade premium. Another important characteristic is a price premium, often termed the community development or social premium. This is paid by the buyer to the cooperative organization in addition to the sales price. Prior to 2008, for coffee, this premium was set at 10 cents per pound but is now 20 cents per pound with 5 cents earmarked for productivity improvement. The premium is designed to foster the associativity and democratic process that are tenets of the Fair Trade philosophy. The specifics of how the premium is to be used must to be decided in a democratic manner by the producers themselves. Projects that are typically funded with the Fair Trade premium include investments made to increase farmer productivity; investments in community infrastructure such as the building of schools, health clinics, and crop storage facilities; offering training for members of the community; the provision of educational scholarships; improvements in water treatment systems; conversion to organic production techniques; and so on.

3) Stability and access to credit. Fair Trade buyers agree to long-term contracts (at least one year and often several years) and to provide some advance crop financing to producer groups (up to 60 percent) if requested.

4) Working conditions. Where workers are present, they must have freedom of association, safe working conditions, and wages at least equal to the legal minimum or regional averages. Some forms of child labor are prohibited.

5) Institutional structure. Farmers are encouraged to organize as associations or cooperatives, where decisions are made democratically and with a transparent administration that can facilitate sales and administer the premium paid to the organization in an accountable manner. For some products, such as tea, bananas, pineapples, and flowers, larger enterprises can become Fair Trade certified.[3] In such larger enterprises, joint committees of workers and managers must be formed and democratically structured.

6) Environmental protection. Certain harmful chemicals are prohibited for Fair Trade production. The environmental criteria are meant to ensure that the members work towards good environmental practices as an integral part of farm management by minimizing or eliminating the use of less-desirable agrochemicals and replacing them, where possible, with natural biological methods, as well as adopting practices that ensure the health and safety of farm families, workers, and the community. Producers must provide basic environmental reports summarizing their impacts on the environment. The production of genetically modified crops by farmers is not allowed. (In practice, this is only relevant for a few crops for which genetically modified varieties are available to these farmers, namely cotton and rice.)

For a product to be sold under the Fair Trade mark, all actors in the supply chain—including importers and exporters—must also be Fair Trade certified. The standards are tailored for each crop and for the different actors involved in the chain. The dominant entities in the global Fair Trade system are Fairtrade International, which is responsible for setting and maintaining standards for all commodities, and FLO-CERT, an independent certification company that is in charge of inspecting and certifying producers and traders.[4]

To obtain the Fair Trade certification, producer organizations, firms or qualified farms submit an application with FLO-CERT. If the application is accepted, the organization goes through an initial inspection process carried out by a FLO-CERT representative in the region. If the minimum requirements are met, the organization is issued a certificate that is usually valid for a year and can be renewed following re-inspection. During the early years of Fair Trade, inspection and certification were free of charge. However, since 2004 producer organizations must pay application, initial certification, and renewal certification fees.

[3] For several key crops such as cocoa and coffee that are smallholder dominated, Fairtrade International believes that certification should generally be restricted to small producers. However, Fair Trade USA (a former member of Fairtrade International under the name Transfair USA) believes that unorganized producers and plantations should also be certified, and this difference is a fundamental reason for its recent split from Fairtrade International and its subsequent certification of some larger producers under its own label.

[4] As noted earlier, there are number of other Fair Trade organizations. For example, Fair Trade USA uses another independent certifier, but generally accepts FLO-CERT as equivalent.

Certification and Credible Information for Consumers

One important rationale for the Fair Trade initiative is that it provides credible information to the consumer. A number of consumers may derive utility from the manner in which a good is produced rather than simply the physical characteristics of the final product. Although many consumers prefer to purchase goods produced in a socially and environmentally responsible manner (and would be willing to pay more for these goods) and many producers would be willing to produce in this manner (particularly for a higher price), without a credible way to differentiate between more-responsible and less-responsible production processes, a market for responsibly produced products may not exist. The Fair Trade label, as well as other third-party certifications, provides the consumer with information about the nature of the production process. It also provides producers a way to credibly signal the nature of the production process. In this way, certifications, such as Fair Trade, can provide information that facilitates mutually beneficial transactions that otherwise would not occur.

A number of studies have formally modeled the logic of Fair Trade, finding that if consumers value the nature of the production process, then voluntary certifications unambiguously improve aggregate welfare. For example, Podhorsky (2010), focusing on environmental standards, shows that in an environment with heterogeneous firms, a voluntary certification program never decreases consumer welfare. Similarly, Podhorsky (2013b) shows that in a two-country model of North–South trade with differentiated products, voluntary certifications improve aggregate welfare.

However, all of this relies on consumers caring about whether goods are produced in a socially or environmentally responsible manner. Do consumers in fact care about this? A number of studies have tackled this question, attempting to quantify the extent to which consumers are willing to pay for responsible production.

Hertel, Scruggs, and Heidkamp (2009) survey 258 individuals and find that 75 percent of coffee buyers report that they would be willing to pay 50 cents extra for a pound of coffee (approximately 15 percent of the sales price) if it was Fair Trade certified. Over half would be willing to pay $1 more per pound.

Complementing the evidence from survey questions asking about hypothetical scenarios is evidence from field experiments that observe real-life behavior. Hainmueller, Hiscox, and Sequeira (2011) conduct a number of experiments in 26 stores belonging to a major US grocery chain. The authors randomly placed Fair Trade labels on bulk bins of coffee that were Fair Trade certified. In a second experiment, the authors also randomly varied the prices of the coffee. Each treatment lasted four weeks. The authors found that sales were 10 percent greater when the coffee was labeled as Fair Trade. They also found that demand for more expensive (and arguably higher quality) Fair Trade coffee was insensitive to price, which is consistent with an earlier finding by Arnot, Boxall, and Cash (2006) for brewed coffee sold at a Canadian university. Interestingly, demand for a cheaper and lower-quality Fair Trade coffee was sensitive to the price: a 9 percent increase in price resulted in a 30 percent decline in demand. In a follow-up experiment using coffee sold on eBay, Hiscox, Broukhim, and Litwin (2011) find that on average, consumers are willing to pay a 23 percent premium for coffee labeled as Fair Trade.

In a series of auxiliary experiments looking at Fair Trade labeling for nonfood consumer items, Michael Hiscox and various coauthors have accumulated a large amount of additional evidence that confirms the findings from Hainmueller, Hiscox, and Sequeira (2011). Examining fair labor standards for candles and towels sold in a large retail store in New York City, Hiscox and Smyth (2011) find that the label increased sales by 10 percent, and when combined by a price markup of 10–20 percent, sales rose even more, in the range of 16–33 percent. Examining consumers' willingness to pay for goods using an auction environment on eBay, Hiscox, Broukhim, Litwin, and Wolowski (2011) find that consumers paid a 45 percent premium for polo shirts labeled as being certified for fair labor standards.

Overall, the evidence from these experiments indicates that consumers value production that occurs according to Fair Trade standards and they believe that certification conveys credible information.

Does Fair Trade Work?

In side-by-side comparisons, Fair Trade–certified producers do receive higher prices, follow specified work standards, and use environmentally friendly methods. We review this evidence, but also explore the more difficult questions of interpretation. Are the changes that are *correlated* with Fair Trade production also *caused* by certification or is some other factor like the entrepreneurial capacity of the producer affecting both outcomes? What factors make producers more likely to join Fair Trade? What may happen to the advantages of receiving a higher price from being a Fair Trade producer as more producers seek to join? After taking these factors into account, the balance of the evidence does suggest that Fair Trade works—but the evidence is admittedly both mixed and incomplete.

Fair Trade and Higher Prices: Direct Comparisons

There is overwhelming evidence that Fair Trade–certified producers do receive higher prices than conventional farmers for their products. For example, Méndez et al. (2010) surveyed 469 households for 18 different cooperatives in four countries—El Salvador, Guatemala, Mexico, and Nicaragua—during the 2003/2004 coffee harvest. In all four countries, they find a significant positive relationship between average sales price for coffee and both Fair Trade and Organic certification. In a study of 845 coffee farmers from southern Mexico during the 2004–2005 season, Weber (2011) finds that farmers who were Fair Trade and Organic certified received an average of 12 cents more per pound of coffee sold.

Bacon (2005) examines the sales price of coffee during the coffee price crisis of 2000/2001 for a sample of 228 coffee farmers from Nicaragua and finds that Fair Trade–certified farmers obtained significantly higher prices for their coffee. Farmers selling coffee as Fair Trade received an average price of $0.84 per pound (net of costs paid to the cooperative for transport, processing, certification, debt service, and export), farmers selling coffee as Organic received $0.63 per pound, while farmers selling conventional coffee to a cooperative received $0.41 per pound. Because Fair

Trade and/or Organic farmers are not able to sell all of their coffee as certified, the average price received by certified and conventional farmers for their full harvest is lower than the figures above. Fair Trade and/or Organic farmers received an average price of $0.56 per pound, while conventional farmers received an average price of $0.40 per pound.

In a follow-up study, Bacon, Méndez, Gomez, Stuart, and Flores (2008) attempt to get a better sense of the causal mechanisms behind these differences. Examining the same set of Fair Trade–certified farmers as in Bacon (2005), they find that 100 percent of these farmers felt that the cooperative they certified with helped them obtain higher prices. This figure can be contrasted to the response of farmers in conventional cooperatives. Among this comparison group, only 50 percent of farmers felt that the cooperative helped them obtain higher prices.

Given the price premium and price floor associated with Fair Trade, it is unsurprising that Fair Trade–certified farmers receive higher prices. However, what is less obvious before looking at the evidence is whether production volumes and, as a consequence, total incomes would be affected by certification. Overall, the evidence does suggest that Fair Trade is often also associated with higher output and higher incomes. Arnould, Plastina, and Ball (2009) examine 1,269 farmers from Nicaragua, Peru, and Guatemala in 2004–2005 and find that in addition to higher prices, Fair Trade–certified farmers also have greater sales and higher incomes. Jaffee (2009) also finds the same pattern for 51 coffee producers (26 Fair Trade–certified and 25 conventional) from Oaxaca, Mexico, surveyed between 2001 and 2005. He also finds that Fair Trade–certified producers were less likely to experience food shortages and had diets that contained more meat, milk, and cheese.

Interpreting the Evidence: Causal Inference

Of course, simple comparisons of certified and noncertified farmers raise obvious questions. Perhaps the characteristics that cause farmers to become Fair Trade certified also cause farmers to produce and sell more, to produce better-quality coffee that sells for a higher price, and to earn more income as a result. Therefore, such comparisons may not capture a causal effect.

Well aware of these empirical difficulties, a number of studies have attempted to reduce the bias in their estimates through the use of matching methods. Intuitively, rather than compare Fair Trade farmers to non–Fair Trade farmers, matching estimates instead compare each certified farmer to conventional farmers that are similar based on observable characteristics that may affect their propensity to certify and be successful producers, such as educational attainment, age, family size, farm size, specialization of production, farm tenure, value of assets, and so on. The hope is that by matching on these characteristics, one is then comparing a Fair Trade–certified farmer to an otherwise similar conventional farmer.

Using this method, Beuchelt and Zeller (2011) examine 327 members of coffee cooperatives in Nicaragua and find that farmers associated with Fair Trade cooperatives are able to obtain higher prices for their coffee (as are Organic producers). In contrast, Fort and Ruben (2009) and Ruben and Fort (2012) examine 360 coffee farmers from three Fair Trade–certified cooperatives and

three noncertified cooperatives in Peru. They find no statistically significant evidence that Fair Trade–certified farmers receive higher prices, using either ordinary least squares or matching estimates.

Although arguably a methodological improvement, matching estimates also have their own shortcomings. The choice of which variables should be used to match farmers is not clear. Certain variables like the age of the household head, along with experience or educational attainment, are arguably exogenous, but other variables like farm size, specialization or diversification of production, the legal status of farm tenure, and value of assets may be endogenous to the certification process itself. In addition, matching requires that the omitted factors be observable. If the important omitted factors are unobservable, like the entrepreneurial zeal of farmer, the bias arising from this factor cannot easily be eliminated.

Yet another strategy, though less-commonly employed, is to examine a panel of producers over time rather than just a cross-section in one time period. In this way, one can examine whether a producer begins to obtain higher prices (for example) just after they become Fair Trade certified. Using such a strategy, Dragusanu and Nunn (2014) examine an annual panel of 262 coffee mills from Costa Rica between 1999 and 2010. They find that Fair Trade–certified mills receive 5 cents more per pound for exports than conventional mills. They find no difference between Fair Trade–certified and conventional mills in terms of the total quantity sold or exported.

Selection into Certification

Another way to tackle the question of causality is to develop a deeper understanding of what determines which cooperatives choose to become Fair Trade certified (and which farmers choose to join Fair Trade–certified cooperatives). Again, the primary research concern is that the "best" farmers or cooperatives in some difficult-to-observe but real way become certified and also produce more and obtain higher prices—that is, that there is positive selection into Fair Trade.

At a theoretical level, it is unclear whether the selection into Fair Trade should be positive or negative. On one hand, Fair Trade intentionally targets producers who are small and economically disadvantaged, with limited capital, market access, and bargaining power, which suggests that they may be negatively selected. In addition, because the price premium is a fixed amount, it is relatively more appealing (that is, the premium is a larger share of the final price) for producers selling lower-quality coffee. This too suggests negative selection. On the other hand, farmers and cooperatives who join Fair Trade tend to have some measure of organizational ability, social cohesion, and governance, which suggests the possibility of positive selection.

Seeking to better understand the nature of selection into Fair Trade, Dragusanu and Nunn (2014) interviewed members of Fair Trade–certified cooperatives and conventional mills in Costa Rica. They found four important determinants of certification. First, it turns out that many mills in Costa Rica often also operate stores that sell agricultural products, including certain chemicals (primarily pesticides) that are banned under Fair Trade requirements. The mills that obtain greater revenues from selling banned chemicals find Fair Trade more

costly and are less likely to certify. Second, mills that forecast lower prices in the future perceived a greater benefit from Fair Trade's price floor, and thus were more likely to join. Third, individual farmers who believed in environmental or socially responsible farming practices were more likely to join Fair Trade. Finally, access to information about the logistics of becoming certified and managerial ability were also important. While positive selection likely arises from the last determinant, the nature of selection from the first three is ambiguous.

Some empirical studies have estimated how various factors affect the probability of certification, usually by estimating a propensity score to match farmers belonging to certified mills with conventional producers. These studies tend to find evidence that point towards negative selection. For example, Sáenz-Segura and Zúñiga-Arias (2009) estimate a very strong negative relationship between Fair Trade certification and experience, education, and income within a sample of 103 Costa Rican coffee producers. This finding is echoed in Ruben and Fort's (2012) study of 360 Peruvian coffee farmers (also see Fort and Ruben (2009)). In their sample, farmers that are less educated and own smaller farms are more likely to become certified.

This question of how farmers via cooperatives select into Fair Trade is important and understudied. We view the evidence as incomplete but suggestive of negative selection. If this is the case, then the correlational evidence may actually understate the true causal impacts of Fair Trade.

Fair Trade in the Long Run: Dynamics and the Role of Free Entry in Production

We now turn to the question of the dynamics of Fair Trade. Consider the case in which a small number of producers in a country are Fair Trade certified. Thus, for the same yield and quality of coffee, certified farmers earn more than the other producers in the region. Other producers observe this outcome, and, if they qualify, will likely apply to become Fair Trade certified. In other words, entry will occur. Over time, as more producers become Fair Trade certified, holding constant the total demand for Fair Trade, the proportion of each Fair Trade farmer's output that can be sold as Fair Trade declines. Many economic models have the property that entry dissipates rents. In this case, entry could continue until the expected benefits of Fair Trade certification just equals the cost to producers. The rents that originate from the greater utility consumers obtain from consuming Fair Trade–certified products end up all going to paying the costs of certification. This process of dissipation is the centerpiece of the model developed in de Janvry, McIntosh, and Sadoulet (2012). It is also a feature, though less central, in a number of other models of Free Trade (for example, Podhorsky 2010).

Based on the predictions of their model, de Janvry, McIntosh, and Sadoulet (2012) argue that free entry represents the death knell for the notion that Fair Trade can actually help farmers in the long run. However, one needs to consider a number of other aspects before accepting this conclusion. First, a number of barriers to certification limit the extent of entry. An important barrier is limits on farm sizes. For example, when it comes to coffee, Fairtrade International targets small family farms that do not hire permanent labor year round.

Second, and most importantly, Fair Trade and other certification standards include many nonmonetary goals: creating better conditions for hired workers, creating democratic and transparent cooperatives, encouraging environmentally sustainable production, improving access to credit, and establishing stable long-term buyer–seller relationships. If a high level of entry means that higher-than-normal economic rents are fully dissipated, it also means that these other outcomes are spreading. Indeed, other certifications like UTZ, Organic, Bird Friendly, and Rainforest Alliance have an even greater focus than Fair Trade on goals other than increased incomes for farmers.

This aspect of certification is illustrated in the model developed by Podhorsky (2010). Although the certification in her model is for environmentally responsible production, the logic is identical to that for Fair Trade certification. In her model, although excess profits are competed away by free entry as firms choose more environmentally responsible production processes, consumers are unambiguously better off because they value environmentally sustainable production. Here certification and free entry work together to increase the prevalence of sustainable production.

This link between free entry and rents provides an interesting dilemma for certification agencies. On the one hand, they wish to induce the spread of socially and environmentally responsible production as much as possible. On the other hand, they may also wish to structure certain limits to entry so that they can continue to maintain higher-than-average rents for certified producers.

Free Entry into Other Certifications

Another important issue that is not yet fully understood is the consequence of entry into certification. Fairtrade International (2012a, p. 47) reports that in 2011, 80 percent of Fair Trade–certified producers organizations reported holding at least one additional certification: 61 percent also held Organic, 8 percent had Rainforest Alliance, and 7 percent also had UTZ.

An important role of certifications is to provide credible information to the consumer about the nature of the production process. A potential concern is that if many different standards have distinct yet overlapping requirements, then certification may introduce a measure of confusion and may therefore be less effective.

A second issue is related to the incentives and potential agency issues that can arise. In general, it may not always be in the interest of the certifying agency to enforce certification requirements fully. Although this concern potentially arises with third-party nongovernment organizations, it is particularly a concern with private certifications. For example, one critique is that some of the more recent private certifications may be little more than smart marketing and attempts to cash in on consumers' willingness to pay for sustainably produced products. The existence of these additional certifications may affect consumers' views about the validity and reliability of third-party certifications generally.

A final issue is that from the producer's perspective, multiple certifications mean multiple reports, multiple audits, greater administrative costs, and a greater tax on scarce managerial capital. Further, it is possible that the existence of multiple standards may decrease the extent to which farmers can fully understand and

benefit from each certification. For example, Valkila and Nygren (2009) found that Nicaraguan farmers belonging to Fair Trade–certified cooperatives had a poor understanding of Fair Trade, including its requirements, and potential benefits. According to the authors, one reason was the multiplicity of certification schemes, quality standards, and rural development projects faced by farmers. They simply were not able to keep track of them all and to distinguish one program from another. Méndez et al. (2010) also found that farmers were often unclear or confused about certifications, particularly about Fair Trade, although farmers did have a better understanding of Organic certification.

Overall, the consequences of the rapid growth of certifications are something we know little about, although it is potentially very important.

Does Fair Trade Provide Greater Financial Stability to Farmers?

Fair Trade seeks to increase financial stability for certified farmers through a number of mechanisms, including a price floor, financing from purchasers and co-ops, and longer-term ties between producers and buyers. The evidence seems to indicate that in many environments these benefits are observed. However, important exceptions do occur. For example, Raynolds (2009) collects information from interviews and focus groups with members and leaders of four cooperatives in Peru and Mexico. She reports that corporate buyers of coffee, what she calls "market driven" buyers (for example, importers that sell to Starbucks, Nestle, and Costco), in practice often refuse to buy from cooperatives that request credit. She also finds that these market-driven mainstream buyers, unlike other Fair Trade buyers, are less willing to enter into longer-term stable contracts. They often sign one-year contracts as a minimum Fair Trade requirement, but do little else to create longer-term partnerships with suppliers. However, despite the behavior of these corporate buyers, she still finds that the producer cooperatives view financing as the second-most beneficial aspect of Fair Trade—after the price floor.

Other studies confirm that Fair Trade has succeeded in increasing the credit available to farmers. Bacon et al. (2008) finds that, within a sample of 177 Nicaraguan coffee farmers, 77 percent of Fair Trade–certified farmers reported that their cooperative provided pre-harvest credit, while this figure was only 33 percent for farmers belonging to conventional cooperatives. Méndez et al. (2010), examining data from 469 households from four Latin American countries, find that Fair Trade–certified farmers are more likely to report having access to credit than conventional farmers. Interestingly, they find no relationship between access to credit and Organic certification. Since Organic buyers are not required to provide access to credit, this result suggests that the greater access to credit for Fair Trade–certified farmers may arise due to a causal effect of certification rather than due to selection. If "better" farmers are more likely to certify with Fair Trade and Organic, then one might also have expected a similar relationship between farms and credit to exist with Organic certification too. Absence of such a relationship is thus evidence against positive selection.

Perhaps the most important tool through which Fair Trade aims to provide greater stability to farmers is the price floor. However, not all coffee produced by Fair Trade–certified farmers can be sold for the Fair Trade price. In the Méndez

et al. (2010) sample of Fair Trade–certified farmers from four Latin American countries, 60 percent of certified coffee was sold as Fair Trade. Among the four Fair Trade cooperatives interviewed by Dragusanu and Nunn (2014), the proportion of coffee sold in the previous year as Fair Trade was 10, 40, 53, and 80 percent. These figures are in line with official statistics of the Fairtrade Labor Organization, which report that on average 45 percent of coffee sold by Fair Trade–certified producers is sold on Fair Trade terms. This figure is slightly higher for bananas (72 percent), cane sugar (54 percent), cocoa (61 percent), and cotton (60 percent) (Fairtrade International 2012a, p. 44).

Although empirical evidence remains limited, existing studies often find that Fair Trade–certified farmers perceive and experience greater economic stability than conventional farmers. For example, Bacon (2005) examines a sample of 228 coffee farmers from Nicaragua and finds that Fair Trade farmers report being less concerned about losing their farm in the coming year than conventional farmers.

The Impacts of Certifications on the Environment

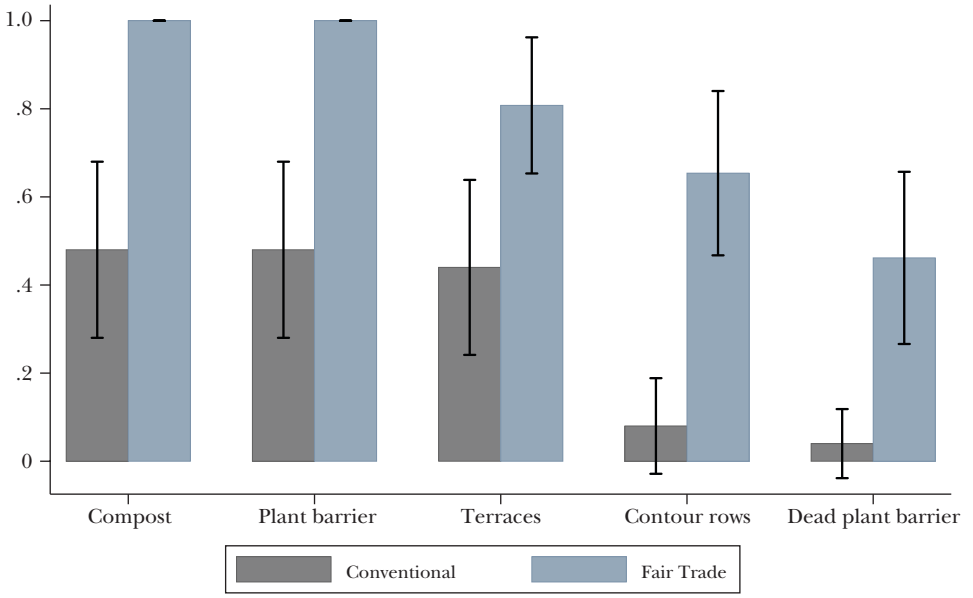
Fair Trade and other environmental labels have been successful in promoting more environmentally friendly farming practices among certified farmers. For example, Jaffee (2009) finds that among a sample of 51 Mexican coffee farmers—26 Fair Trade and 25 conventional—there is a strong association between Fair Trade certification and environmentally friendly farming practices. These include producing compost and applying this to coffee plants; building terraces and contour rows (to reduce soil erosion on sloped land): and building live and dead plant barriers (also to reduce soil erosion). Figure 2 shows the differences in these practices by reporting the proportion of Fair Trade producers and proportion of conventional producers in his sample that were engaged in each practice. In all five cases, the differences between the two groups are statistically significant. Similarly, based on a sample of 177 coffee farmers from Nicaragua, surveyed in 2006, Bacon et al. (2008) find that 68 percent of Fair Trade farmers had implemented ecological water purification systems, compared to 40 percent for conventional farmers. Moreover, 43 percent of Fair Trade farmers had implemented soil and water conservation practices, while only 10 percent of conventional farmers had done so.

Other certifications that target the environment also seem to increase environmentally friendly farming practices. For example, Blackman and Naranjo (2012) examine the impacts of Organic certification among 2,603 coffee farmers in Costa Rica (36 of them certified organic and 2,567 conventional). Using propensity score matching, they find strong evidence that organic farmers were less likely to engage in the use of pesticides, herbicides, and chemical fertilizers and they were more likely to use organic fertilizers, shade trees, and windbreaks and to undertake a variety of soil conservation measures.

How Certifications Affect Local Governance and Institutions

The empirical evidence on whether Fair Trade has been able to strengthen local institutions remains limited. In their study of 360 randomly sampled coffee farmers in Peru, Ruben and Fort (2012) show that when comparing Fair Trade–certified

Figure 2
Differences in Environmental Practices between Fair Trade and Conventional Coffee Producers in Oaxaca, Mexico
(proportion of producers engaging in environmental conservation practices)



Source: Jaffee (2009, chap. 9, table 6).
Note: Averages (and 95 percent confidence intervals) are reported for 25 Fair Trade farms and 26 conventional farms.

farmers against matched conventional farmers, certified farmers are more likely to strongly identify with their cooperative and are more likely to believe the cooperative is important and helpful in the sales process. Interestingly, these differences are only robustly statistically significant when comparing Fair Trade and conventional farmers that are both also certified Organic.

There is also evidence that suggests that Fair Trade does not foster stronger institutions. Elder, Zerrifi, and Le Billon (2012), examining 107 coffee farmers from Rwanda in 2009, find that membership in a Fair Trade–certified cooperative is associated with less trust in the leaders of the cooperative. By contrast, they found no difference on trust in the members of one's own community. Along similar lines, Sáenz-Segura and Zúñiga-Arias (2009), using propensity score matching among a sample of 103 farmers from Costa Rica, find that, relative to conventional farmers, Fair Trade–certified farmers identify less with their cooperative and perceive their cooperative to function more poorly.

These quantitative findings suggests a pattern of tension between farmers belonging to Fair Trade–certified cooperatives and the cooperative itself that is consistent with the qualitative findings from a number of studies. Prevezer (2013) interviewed farmers belonging to Fair Trade–certified coffee cooperatives in

Tanzania. He found that farmers commonly complained of a lack of communication about the use of the premiums, the reasoning behind decisions, and the decision-making process itself. Prevezer also found evidence of elected farmers on the boards misusing the funds—for example, by paying themselves for attending meetings. In their case study of a Fair Trade coffee cooperative in Costa Rica (*Coopermontes de Oro R.L.*), Sáenz-Segura and Zúñiga-Arias (2009) found a significant amount of distrust in the cooperative arising because of deficient management in the past. Mendez et al. (2010) report evidence, from surveys undertaken in Mexico, Nicaragua, Guatemala, and El Salvador, of dissatisfaction and concern over a lack of transparency, accountability, and communication on the part of cooperative members (that is, farmers) directed towards cooperative leaders.

An important factor is likely a lack of knowledge about Fair Trade on the part of farmers and particularly about the existence and intended nature of the Fair Trade price premium. In their case study of a Fair Trade coffee cooperative in Costa Rica, Sáenz-Segura and Zúñiga-Arias (2009) found that one-third of producers did not know about the existence of a premium and about half felt that they had not received any benefits from Fair Trade certification. As another example, Fort and Ruben (2009) find that among 180 Fair Trade–certified coffee farmers from Peru, 12 percent did not know about the existence of the Fair Trade premium and 77 percent felt that they did not receive any benefits from the premium.

Although much more research is needed, one can imagine a number of theoretical reasons why Fair Trade certification could erode trust and increase tensions within a cooperative. With certifications come increased rents. With more at stake, it is natural that tensions may escalate. In addition, because the specifics of Fair Trade remain opaque to members, this may directly cause increased suspicion on their part.

Distributional Considerations of Fair Trade

According to the World Fair Trade Organization and Fairtrade Labelling Organizations International (2009, p. 4), Fair Trade seeks to contribute "to sustainable development by offering better trading conditions to, and securing the rights of, marginalized producers and workers." In this quotation, as in many discussions, producers and farm owners are often lumped together with workers as being potential beneficiaries of certification. But to what extent are the benefits of certification felt by the farm owners and to what extent do they also reach the hired workers?

In the coffee industry, the majority of farms are small and family-run with few hired workers, who tend to be seasonal harvesters. In addition, the cooperative itself may hire workers. While coffee and cacao are primarily smallholder crops, commodities like bananas, citrus fruits, and tea are more commonly produced on large plantations. For these products, certification has been expanded to also include larger plantations with the same general principles of Fair Trade being followed, although details of the certification standards vary by product and by organizational form (for example, plantation vs. cooperative).

The evidence on the distribution of the benefits of Fair Trade remains limited, but the available studies suggest that, within the coffee industry, Fair Trade certification benefits workers little or not at all. Valkila and Nygren (2009) interviewed 94 producers and 64 hired workers from 11 Nicaraguan coffee cooperatives over a six-month period in 2005 and 2006. They found that although the records of Fair Trade farmers indicated that they received higher prices for their coffee, the workers (like the rest of rural Nicaragua) were paid the minimum wage and were not given benefits like social security, medical care, vacation, pension, and paid sick leave. Dragusanu and Nunn (2014) come to a similar conclusion in their empirical study of the impacts of Fair Trade coffee in Costa Rica. Examining a panel of over 110,000 individuals annually between 2003 and 2010, they find that while Fair Trade certification is associated with significantly greater incomes for farmers, it is not associated with increased incomes for hired workers.

Jaffee's (2009) direct comparison of 26 Fair Trade coffee farmers and 25 conventional coffee farmers from Oaxaca, Mexico in 2002–2003, also suggests that few of the benefits of Fair Trade are passed on to workers. Although he finds that the average price obtained by Fair Trade–certified farmers is 130 percent higher than for conventional farmers (13.22 versus 5.74 pesos per kilogram), the wages paid to hired workers are only 7 percent higher (47 versus 44 pesos per day). As a result, labor costs (measured as a share of coffee sales) were lower for Fair Trade–certified farms than for conventional farms (57.2 versus 68.3 percent).

Looking outside of coffee, the effects of Fair Trade on workers are more noticeable. When large plantations are certified, Fair Trade regulations pay particular attention to workers (who comprise a larger group relative to farm owners in these products). In the case of plantations, certification requires workers' freedom of association, safe and equitable working conditions, the absence of forced or child labor, and salaries that are at least as high as the established minimum wage. In addition, certification requires that workers form a democratically elected workers' organization if one does not already exist.

Ruben and van Schendel (2009) survey workers from two banana plantations in Ghana, one Fair Trade certified, the other not. Comparing 50 workers from the conventional plantation with 50 matched workers from the Fair Trade plantation, they find that workers in the conventional plantation received a higher base salary, but they worked longer hours and received fewer fringe benefits. Fair Trade workers felt a greater sense of identity and co-ownership with the plantation. Granville and Telford (2013) survey 381 workers (273 Fair Trade workers and 108 conventional workers) in South Africa's wine industry. Directly comparing the two groups, they find that Fair Trade workers are more likely to make more than minimum wage and as a result they are also able to save more of their income. Consistent with this, 91 percent of Fair Trade workers said in survey results that Fair Trade (and their membership in the joint body) was responsible for improving their living standards. In particular, 95 percent of workers reported that Fair Trade provided help with education and/or health (and 51 percent reported being helped with both).

Another distributional concern is whether Fair Trade results in greater inequality between farmers that are certified and those that are not. Kadow (2011)

examines this possibility theoretically in a Ricardian model of North–South trade. In the model, Fair Trade and conventional coffee are differentiated products. The price of Fair Trade coffee is higher and the price gap is due to preferences of Northern consumers for consumption of coffee produced in a socially responsible manner. In the model, Fair Trade increases global welfare. This arises primarily by reducing the income differences between the North and South, since there is diminishing marginal utility of income. In addition, because Fair Trade increases the incomes of certified producers in the South and does not benefit conventional producers, Fair Trade would theoretically increase inequalities within the South. While it is clearly true that Fair Trade–certified farmers benefit more from Fair Trade, in reality there are reasons why noncertified farmers may also gain (or even those outside of the coffee industry). The primary reason that noncertified farmers may benefit from Fair Trade is due to the use of the price premium for projects that benefit the broader community, such as education, health care services, water, electricity, road infrastructure, and so on. Dragusanu and Nunn (2014) describe a scholarship program, Children of the Field Foundation, initially implemented in Costa Rica by COOCAFE using Fair Trade premiums. Since its implementation in 1996, the program has provided scholarships to 2,598 students and financial support to 240 schools. COOCAFE estimates that in total, over 5,800 students have been helped by the foundation.

Dragusanu and Nunn's (2014) empirical study of Fair Trade coffee producers in Costa Rica is one of the few studies that directly estimates potential spillovers from Fair Trade certification. In their panel analysis, they estimate the effects of more Fair Trade certification on other households living in the same area. They find that incomes of households (that are not coffee producers) increase as the extent of Fair Trade certification increases in a district. Although not conclusive, this finding provides preliminary evidence that spillovers may exist and may be positive.

The model developed by Podhorsky (2013a) illustrates another mechanism through which certification can benefit conventional farmers. In the model, a small number of oligopolistic intermediaries buy from local producers and sell to foreign buyers. (This is the typical structure of the industry.) In her model, the existence of a Fair Trade certification decreases the market power of intermediaries, and as a result, even producers that do not choose to become Fair Trade–certified benefit. It is interesting to note that the original impetus for Max Havelaar was to correct the exploitation of coffee farmers by middlemen due to a severe imbalance of bargaining power. Therefore, Podhorsky's model (2013a) seems particularly appropriate for thinking about Fair Trade and its spillovers.

Conclusions

Many consumers value goods produced in a socially and environmentally responsible manner. As a result, efficiency and welfare gains are possible from credible third-party certifications, like Fair Trade, that provide consumers with information about the production process.

The existing empirical evidence, based primarily on conditional correlations, suggests that Fair Trade does achieve many of its intended goals, although on a comparatively modest scale relative to the size of national economies. Fair Trade farmers do on average receive higher prices, have greater access to credit, perceive their economic environment as being more stable, and are more likely to engage in environmentally friendly farming practices. However, some aspects of Fair Trade and its consequences are not yet well understood. There is evidence that farmers in Fair Trade cooperatives may not be fully aware of the details of Fair Trade and can sometimes mistrust those who run the cooperative. Another issue is the trade-off between limiting certification to small-scale disadvantaged producers and allowing larger plantation-style producers to also become certified. By scaling-up Fair Trade and increasing entry into certification, the increased entry may dissipate some of the monetary benefits of certification.

Some scholars have argued that consumers may be better off using institutions that directly transfer benefits to producers in developing countries rather than using market-based mechanisms like Fair Trade. We are skeptical that anything resembling direct transfers is preferable. It has long been recognized that direct transfers of money distort incentives, diverting effort away from productive activities and towards rent-seeking and corruption. For example, a number of recent empirical studies show that foreign aid (whether it is economic, military or food aid) increases conflict (Crost, Felter, and Johnston forthcoming; Dube and Naidu 2012; Nunn and Qian forthcoming). In our view, the largest potential benefit of market-based systems like Fair Trade is that they do not distort incentives in as deleterious a way as foreign aid. Instead, they work within the marketplace and reward productive activities and production processes that are valued by consumers and that are good for the local environment and economy.

■ *We have received valuable comments from David Atkin, David Autor, Michael Hiscox, Ulrike Malmendier, Marc Muendler, Nina Pavcnik, Andrea Podhorsky, and Sandra Sequeira, and in particular Timothy Taylor. We also thank Stephanie Cappa and Matthew Summers for valuable research assistance. Giovannucci received financial support for this project from the Committee on Sustainability Assessment. Nunn and Dragusanu received financial support for this project from Harvard's Weatherhead Center for International Affairs. Dragusanu also received financial support for this project from Harvard University's International Economics Grant.*

References

Arnot, Chris, Peter C. Boxall, and Sean B. Cash. 2006. "Do Ethical Consumers Care About Price? A Revealed Preference Analysis of Fair Trade Coffee Purchases." *Canadian Journal of Agricultural Economics* 54(4): 555–65.

Arnould, Eric J., Alejandro Plastina, and Dwayne Ball. 2009. "Does Fair Trade Deliver on Its Core Value Proposition? Effects on Income, Educational Attainment, and Health in Three Countries." *Journal of Public Policy and Marketing* 28(2): 186–201.

Bacon, Christopher. 2005. "Confronting the Coffee Crisis: Can Fair Trade, Organic, and Specialty Coffee Reduce Small-Scale Farmer Vulnerability in Northern Nicaragua?" *World Development* 33(3): 497–511.

Bacon, Christopher M., V. Ernesto Méndez, María Eugenia Flores Gómez, Douglas Stuart, and Sandro Raúl Díaz Flores. 2008. "Are Sustainable Coffee Certifications Enough to Secure Farmer Livelihoods? The Millennium Development Goals and Nicaragua's Fair Trade Cooperatives." *Globalizations* 5(2): 259–74.

Beuchelt, Tina D., and Manfred Zeller. 2011. "Profits and Poverty: Certification's Troubled Link for Nicaragua's Organic and Fairtrade Coffee Producers." *Ecological Economics* 70(7): 1316–24.

Blackman, Allen, and Maria A. Naranjo. 2012. "Does Eco-Certification Have Environmental Benefits? Organic Coffee in Costa Rica." *Ecological Economics* 83: 58–66.

de Janvry, Alain, Craig McIntosh, and Elisabeth Sadoulet. 2012. "Fair Trade and Free Entry: Can a Disequilibrium Market Serve as a Development Tool?" Unpublished paper, University of California, Berkeley.

Collier, Paul. 2007. *The Bottom Billion*. New York: Oxford University Press.

Crost, Benjamin, Joseph Felter, and Patrick Johnston. Forthcoming. "Aid Under Fire: Development Projects and Civil Conflict." *American Economic Review*.

Dragusanu, Raluca, and Nathan Nunn. 2014. "The Impacts of Fair Trade Certification: Evidence from Coffee Producers in Costa Rica." http://scholar.harvard.edu/nunn/publications/impacts-fair-trade-certification-evidence-coffee-producers-costa-rica.

Dube, Oeindrila, and Suresh Naidu. 2012. "Bases, Bullets, and Ballots: The Effect of U.S. Military Aid on Political Conflict in Colombia." Center for Global Development Working Paper 197. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1542699.

Economist, The. 2006. "Voting with Your Trolley." December 7. http://www.economist.com/node/8380592.

Elder, Sara D., Hisham Zerriffi, and Philippe Le Billon. 2012. "Effects of Fair Trade Certification on Social Capital: The Case of Rwandan Coffee Producers." *World Development* 40(11): 2355–67.

Fairtrade International. 2012a. "Monitoring the Scope and Benefits of Fair Trade." Fourth Edition.

Fairtrade International. 2012b. "Fair Trade and Coffee." Commodity Briefing.

Fairtrade International. 2013. "Unlocking the Power: Annual Report, 2012–13." September.

FLO-CERT. "Fee System Small Producer Organization: Explanatory Document." Effective as of: 01.01.2013.

Fort, Ricardo, and Ruerd Ruben. 2009. "The Impact of Fair Trade Certification on Coffee Producers in Peru." In *The Impact of Fair Trade*, edited by Ruerd Ruben, 75–98. Wageningen: Wageningen Academic Publishers.

Granville, Brigitte, and Steven Telford. 2013. "Empirical Evidence from South Africa." In *The Process and Practices of Fair Trade*, edited by Brigitte Granville and Janet Dine, 286–340. London: Routledge.

Hainmueller, Jens, Michael J. Hiscox, and Sandra Sequeira. 2011. "Consumer Demand for the Fair Trade Label: Evidence from a Field Experiment." Unpublished paper.

Hertel, Shareen, Lyle Scruggs, and C. Patrick Heidkamp. 2009. "Human Rights and Public Opinion: From Attitudes to Action." *Political Science Quarterly* 124(3): 443–59.

Hiscox, Michael J., Michael Broukhim, and Claire S. Litwin. 2011. "Consumer Demand for Fair Trade: New Evidence from a Field Experiment using eBay Auctions of Fresh Roasted Coffee." Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1811783.

Hiscox, Michael J., Michael Broukhim, Claire S. Litwin, and Andrea Woloski. 2011. "Consumer Demand for Fair Labor Standards: Evidence from a Field Experiment on eBay." Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1811788.

Hiscox, Michael J., and Nicholas F. B. Smyth. 2011. "Is There Consumer Demand for Fair Labor Standards? Evidence from a Field Experiment." Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1820642.

Jaffee, Daniel. 2009. "'Better, But Not Great': The Social and Environmental Benefits and Limitations of Fair Trade for Indigenous Coffee Producers in Oaxaca, Mexico." In *The Impact*

of Fair Trade, edited by Ruerd Ruben, 195–222. Wageningen: Wageningen Academic Publishers.

Kadow, Alexander. 2011. "The Fair Trade Movement: An Economic Perspective." Working Paper 2001_05. University of Glasgow.

Méndez, V. Ernesto, Christopher M. Bacon, Meryl Olson, Seth Petchers, Doribel Herrador, Cecilia Carranza, Laura Trujillo, Carlos Guadarrama-Zugasti, Antonio Cordón, and Angel Mendoza. 2010. "Effects of Fair Trade and Organic Certifications on Small-Scale Coffee Farmer Households in Central America and Mexico." *Renewable Agriculture and Food Systems* 25(3): 236–51.

Nunn, Nathan, and Nancy Qian. Forthcoming. "US Food Aid and Civil Conflict." *American Economic Review*.

Podhorsky, Andrea. 2010. "Environmental Labeling." Available at the author's website: http://dept.econ.yorku.ca/~andrea/research.html.

Podhorsky, Andrea. 2013a. "A Positive Analysis of Fair Trade Certification." Available at the author's website: http://dept.econ.yorku.ca/~andrea/research.html.

Podhorsky, Andrea. 2013b. "Certification Programs and North–South Trade." *Journal of Public Economics* 108: 90–104.

Potts, Jason, Matthew Lynch, Ann Wilkings, Gabriel A. Huppé, Maxine Cunningham, and Vivek Voora. 2014. *The State of Sustainability Initiatives Review 2014: Standards and the Green Economy*. International Institute for Sustainable Development and London and the International Institute for Environment and Development.

Prevezer, Martha. 2013. "Fair Trade Governance and Its Impact on Local Development: A Framework." In *The Process and Practices of Fair Trade*, edited by Brigitte Granville and Janet Dine, pp. 19–42. London: Routledge.

Raynolds, Laura T. 2009. "Mainstreaming Fair Trade Coffee: From Partnership to Traceability." *World Development* 37(6): 1083–93.

Raynolds, Laura T., Douglas Murray, and Andrew Heller. 2007. "Regulating Sustainability in the Coffee Sector: A Comparative Analysis of Third-Party Environmental and Social Certification Initiatives." *Agriculture and Human Values* 24(2): 147–63.

Ruben, Ruerd, and Ricardo Fort. 2012. "The Impact of Fair Trade Certification for Coffee Farmers in Peru." *World Development* 40(3): 570–82.

Ruben, Ruerd, Ricardo Fort, Guillermo Zúñiga-Arias. 2009. "Measuring the Impact of Fair Trade on Development." *Development in Practice* 19(6): 777–88.

Ruben, Ruerd, and Lucie van Schendel. 2009. "The Impact of Fair Trade in Banana Plantations in Ghana: Income, Ownership and Livelihoods of Banana Workers." *The Impact of Fair Trade*, edited by Ruerd Ruben, 137–53. Wageningen: Wageningen Academic Publishers.

Sáenz-Segura, Fernando, and Guillermo Zúñiga-Arias. 2009. "Assessment of the Effect of Fair Trade on Smallholder Producers in Costa Rica: A Comparative Study in the Coffee Sector." *The Impact of Fair Trade*, edited by Ruerd Ruben, 117–35. Wageningen: Wageningen Academic Publishers.

Sidwell, Marc. 2008. *Unfair Trade*. London: ASI (Research) Ltd.

Valkila, Joni, and Anja Nygren. 2009. "Impacts of Fair Trade Certification on Coffee Farmers, Cooperatives, and Laborers in Nicaragua." *Agriculture and Human Values* 27(3): 321–33.

Weber, Jeremy. 2011. "How Much More Do Growers Receive for Fair Trade-Organic Coffee?" *Food Policy* 36(5): 678–85.

World Fair Trade Organization and Fairtrade Labelling Organizations International. 2009. "A Charter of Fair Trade Principles." January 2009. http://www.fairtrade.net/fileadmin/user_upload/content/2009/about_us/documents/Fair_Trade_Charter.pdf.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 237–248

Evaluating Counterterrorism Spending

John Mueller and Mark G. Stewart

In 2010, a committee of the National Research Council of the National Academies of Science issued a report on how effectively the Department of Homeland Security (DHS) was spending its funds. The committee praised the approaches used in the department for decisions about natural hazards, saying they were "near state of the art . . . based on extensive data, have been validated empirically, and appear well suited to near-term decision needs." In contrast, with regard to the department's spending on counterterrorism, the committee could "not find any DHS risk analysis capabilities and methods that are yet adequate for supporting DHS decision making" and observed that "little effective attention was paid to the features of the risk problem that are fundamental" (National Research Council 2010). As far as we can tell, this seemingly newsworthy report inspired no media coverage whatever.

In general, it seems, counterterrorism agencies simply identify a potential source of harm and try to do something about it, rather than systematically thinking about the likely magnitude of harm caused by a successful terrorist attack, the probability of that attack occurring, and the amount of risk reduction that can be expected from counterterrorism efforts. Without considering such factors, it is impossible to evaluate whether security measures reduce risk sufficiently to justify their costs.

■ *John Mueller is Senior Research Scientist, Mershon Center for International Security Studies and Adjunct Professor of Political Science, both at Ohio State University, Columbus, Ohio. He is also a Senior Fellow, Cato Institute, Washington, DC. Mark G. Stewart is Australian Research Council Professorial Fellow and Professor and Director, Centre for Infrastructure Performance and Reliability, both at the University of Newcastle, New South Wales, Australia. Their email addresses are bbbb@osu.edu and mark.stewart@newcastle.edu.au, respectively.*

http://dx.doi.org/10.1257/jep.28.3.237 doi=10.1257/jep.28.3.237

In this article, we lay out a simple, back-of-the-envelope approach for evaluating counterterrorism spending that uses only four variables: the consequences of a successful attack, the likelihood of a successful attack, the degree to which the security measure reduces risk, and the cost of the security measure. After measuring the cost of a counterterrorism measure, we explore a range of outcomes for the costs of terrorist attacks and a range of possible estimates for how much risk might be reduced by the measure. Then working from this mix of information and assumptions, we can calculate how many terrorist attacks (and of what size) would need to be averted to justify the cost of the counterterrorism measure in narrow cost–benefit terms.

To illustrate this approach, we first apply it to the overall increases in domestic counterterrorism expenditures that have taken place since the terrorist attacks of September 11, 2001, and alternatively we apply it to just the FBI's counterterrorism efforts. We then evaluate evidence on the number and size of terrorist attacks that have actually been averted or might have been averted since 9/11. We also consider the degree to which our conclusions would be altered if we believe deaths from terrorist attacks should be given greater weight than deaths caused by hazards like natural disasters or traffic accidents.

Assessing Increases in Domestic Counterterrorism Expenditures since 9/11

Our evaluation of the rise in domestic counterterrorism expenditures relies on three main ingredients: how much such expenditures have in fact increased; the range of possible costs of terrorist actions; and the risk reduction from the increase in counterterrorism expenditures. Based on these three factors, we can calculate how many terrorist attacks would have had to be averted for the rise in counterterrorism spending since 9/11 to be justified. From this calculation, it appears likely that the rise in counterterrorism spending was too large.

By our calculation, domestic counterterrorism expenditures per year were about $25 billion in 2010 dollars before the terrorist attacks of September 11, 2001. These increased by about $75 billion in the subsequent decade or so. Spending on homeland security by the Department of Homeland Security, the Department of Defense, the Department of Justice, the Department of Health and Human Services, the Department of Energy, and 26 other such federal agencies was $50 billion more in 2010 dollars than in 2001, adjusting for inflation. To this we add the costs of US intelligence focused on the homeland. Overall intelligence operations were $80 billion in 2010. A core function is "protecting against the threat of international terrorism in the United States," and we conservatively estimate increased intelligence expenditures since 9/11 devoted to domestic homeland security to be $15 billion in 2010. Finally, enhanced outlays for state and local homeland security spending are approximately $10 billion per year.

Although we will use this figure of $75 billion per year for the annual increase in spending on domestic counterterrorism, it should be viewed as a very conservative

Table 1

How Many Terrorist Attacks Would Need to Occur Each Year in the Absence of All Counterterrorism Measures in Order to Begin to Justify a Counterterrorism Expenditure of $75 Billion

	Type of terrorist attack					
	Ft. Hood Shooting	*Boston, Times Square bombing*	*London bombing*	*9/11*	*Nuclear bomb, port*	*Nuclear bomb, Grand Central Station*
Losses per incident	*$100 million*	*$500 million*	*$5 billion*	*$200 billion*	*$1 trillion*	*$5 trillion*
Level of risk reduction assumed						
10 percent	7,500	1500	150	4	.75	.15
25 percent	3,000	600	60	2	.30	.06
50 percent	1,500	300	30	.75	.15	.03
75 percent	1,000	200	20	.50	.10	.02
90 percent	833	167	17	.42	.08	.02
100 percent	750	150	15	.38	.08	.02

Notes: If the $75 billion expenditure is expected to reduce the risk (the likelihood of, and/or the damage caused by, a successful terrorist attack) by 50 percent, those expenditures would need to deter, disrupt, or protect against at least half of the attacks in each entry in the 50 percent line. For the boxed entries, that would be 150 Boston-type attacks per year, 15 London-type attacks each year, or one 9/11-type attack about every three years.

measure of the degree to which homeland security expenditures have risen since 9/11. It leaves out nearly $50 billion in various opportunity costs, like time spent in airports in security lines, as well as privacy issues and $10 billion in costs incurred by the private sector (Mueller and Stewart 2011, pp. 81–82).

For thinking about the possible costs—both direct and indirect—inflicted by a terrorist attack, we lay out a range of possibilities across the top of Table 1. Deaths at the hands of terrorists are very often taken to be far more significant than those from other causes. Indeed, a study commissioned by the Department of Homeland Security suggests that, although human life is often taken to have a value of some $6–$7 million, lives lost to terrorism should be valued at twice that amount (Robinson, Hammitt, Aldy, Krupnick, and Baxter 2010). Others might suggest even higher multiples. In estimating the costs inflicted by a terrorist event, however, we prefer to value life at the more common figure, and then, on a case-by-case basis, we can discuss and add back in the indirect costs from economic, social, and psychological side-effects.

This approach can be illustrated by examining the attack costs arrayed in Table 1. Terrorism mostly inflicts losses that are quite low—in general, terrorism

is not only a low probability event, but also a low-consequence one, and the vast majority of terrorist attacks do not kill anyone (Mueller and Stewart 2011, chap. 3). However, at the low end of the scale in the table we start with events that impose a substantial loss of $100 million. An example would be the shootings at Fort Hood in Texas in 2009 in which 13 people were killed. Although this has been by far the greatest loss of life inflicted in a terrorist act in the United States since 2001, almost all of the damage came in direct costs in the form of death and injury. It did not seem to cause additional substantial economic losses or widespread fear or anguish.

The direct and indirect losses inflicted in the Boston Marathon bombings of 2013 were probably much higher—in the vicinity of $500 million—even though the death toll was lower. In addition to three deaths, hundreds of injuries, and property damage, the Boston terrorists inflicted considerable indirect losses on the region through the costs of pursuing them over the nearly four days they were on the loose. Travel to Boston was cancelled or deferred, and the large crime scene forced the closure of many businesses. The daily GDP for Boston is close to $1 billion (Bureau of Economic Analysis 2013), so a reduction in economic activity of just 5 or 10 percent for several days represents a large sum. This amount might also be roughly the costs of the damage that would be inflicted at Times Square by a car bomb similar to the one a rather inept terrorist tried unsuccessfully to detonate there in 2010.

The losses sustained at the 2005 London and 2004 Madrid bombings that killed 52 and 191 commuters respectively are sometimes estimated to amount to $5 billion in direct and indirect losses, with most estimates around $2 or $3 billion (for additional discussion of these estimates, see Mueller and Stewart 2011, chap. 3).

A number of studies have sought to assess the direct and indirect costs of the 9/11 terrorist attacks—far and away the most destructive single terrorist act in history and one in which the indirect costs considerably outweigh the (obviously horrific) direct ones. They generally conclude that a fair, if somewhat high, estimate for the full losses sustained in the attack—lives lost, property damaged or destroyed, psychological trauma, and indirect losses from travel and tourism reductions, business interruption, and economic shocks—would be some $200 billion with loss of life valued at $20 billion, direct physical damage at $30 billion, and loss of GDP at $70 to $140 billion (equivalent to 0.5 to 1 percent of GDP) (as reviewed in Mueller and Stewart 2011, chap. 3). The potential losses if terrorists were able to set off an atomic bomb or device at an important port might reach $1 trillion (Meade and Molander 2006). The losses for an atomic explosion at Grand Central Station in New York could be $5 trillion (Bunn, Weir, and Holdren 2003), although the likelihood that terrorists could do this seems very small (Mueller 2010).

An additional consideration concerns what might be called extended (as opposed to indirect) costs. Thus, 9/11 not only led to considerable indirect costs as people avoided flying and traveling for a time, but the attacks also propelled the United States into expensive overseas wars (Stiglitz and Bilmes 2008). Few terrorist events trigger such extreme reactions, which can be considered either as contributors to the losses sustained in the terrorist attack or to the costs of counterterrorism.

We do not include such extended costs in the basic analysis here, but will return to this subject later in the discussion.

The rows of Table 1 show various degrees of risk reduction. In assessing risk reduction for increased security expenditures since 9/11, it is important to remember that a number of homeland security measures were already in place before that attack. In addition, the attacks of 9/11 massively heightened public awareness regarding the threat of terrorism, resulting in extra vigilance that has often resulted in the arrest of terrorists or the foiling of terrorist attempts. Most dramatically, airplane passengers rather than government counterterrorism efforts foiled the "shoe bomber" on a flight in December 2001 and the "underwear bomber" on a flight in December 2009.

In our analysis, we will assume that the security measures in place before 9/11 along with the extra vigilance of the public after that event combine to reduce risk by 45 percent. This estimate of risk reduction *not* associated with increased spending after 9/11 should be viewed as conservative. After all, the most cost-effective security measures are generally the first to be implemented. For example, a study of security measures in shopping centers found that the least costly measures, suspicious package reporting, reduced risk by 60 percent, but the costly and inconvenient searching of bags at entrances achieved only 15 percent risk reduction (LaTourrette, Howell, Mosher, and MacDonald 2006).

We will also assume that the increase in US expenditures on homeland security since 2001 has been dramatically effective at closing the gap. If the preexisting measures and the extra public vigilance reduce the risk by 45 percent, we will assume that the additional security expenditures put in place after 9/11 reduce the risk by another 50 percent. Thus the total risk reduction supplied by all the security measures is assumed to be 95 percent.

In Table 1, we evaluate the contribution of a security measure or set of measures which cost $75 billion per year. The cells show the number of successful attacks per year that would be required to take place in the absence of all counterterrorism measures in order to begin to justify that expense (the breakeven point). This is shown for various attack scenarios and for various levels of risk reduction.

The boxed entries are for the assumption that $75 billion in security expenditures reduces the risk of a terrorist attack (its consequences and/or its likelihood) by 50 percent. Under that assumption, in order for the costs of a $75 billion security measure or set of security measures to begin to be justified, there would have to have been 300 attacks like the Boston bombing each year—or about one a day—in the absence of all security measures. Or 30 London-size attacks per year—more than one a week. Or about three 9/11 attacks every four years. To justify its expense, a security measure that reduces risk by 50 percent would be expected to deter, disrupt, or protect against half of these—the 300 Boston-type attacks per year would be reduced to 150, for example. In our case, that would be the task of the set of security measures added to those already in place in September 2001, while the existing security measures, combined with the added vigilance inspired by 9/11, would separately deter, disrupt, or protect against almost all of the rest.

Of course, one can also put together more complex mixtures of larger and smaller terrorist events, or tinker with the calculations of how much additional counterterrorism spending might reduce the risks. The approach applied here is designed to provide decision makers with a coherent perspective on the relevant parameters for counterterrorism expenditures and how they interact, but it does not of itself make the decision. The assumptions given here suggest that the increase of $75 billion per year in counterterrorism spending has not been justified on a cost-effectiveness basis because such a justification requires that the risk has been reduced for an implausibly high number of terrorist operations. In a later section, we will consider various possible objections and qualifications for this conclusion.

Tables similar to Table 1 can easily be calculated for specific counterterrorism measures, some of which, such as hardened cockpit doors and the federal flight deck program (which allows pilots, flight engineers, and navigators to volunteer for training so that they can carry a firearm on flights), certainly appear cost-effective (Stewart and Mueller 2013b). Let's apply the same analytical framework to evaluate just the FBI counterterrorism spending on its own. The Bureau's highest priority since 9/11 has been to "prevent terrorism and promote the nation's security consistent with the rule of law," and it currently is involved in following up more than 5,000 terrorism tips—or, as they are known internally, "threats"—each day (Graff 2011, pp. 579, 398–99). Counterterrorism now accounts for close to $3 billion of FBI expenditures while the budget for its criminal division is $2.5 billion (Federal Bureau of Investigation 2013).

A table for a $3 billion counterterrorism expenditure would be the same as Table 1 except that each entry would be divided by 25. We generously assume that by its efforts the FBI has succeeded in reducing the risks of a terrorist attack—that is, the consequences and/or the probability of an otherwise successful attack—by a full 90 percent. To justify its $3 billion counterterrorism budget under that condition, the FBI spending alone would have to deter, disrupt, or protect against some six attacks of the size of the Boston Marathon attacks each year—one every two months. Alternatively, the FBI's efforts would need to reduce by 90 percent the effect of one or two London-type bombings every two years—some six or seven over the course of a decade. Or again alternatively, the FBI budget would justify itself by reducing by 90 percent a huge attack with direct and indirect damage equivalent to that inflicted by 9/11 once every 60 years. The assumption about risk reduction is quite significant: if the FBI's counterterrorism efforts only reduce the total risk of losses in a terrorist attack by 50 percent rather than 90 percent, the number of terrorist events that it would need to prevent or mitigate would nearly double.

Evaluating Prospective Terrorist Attacks in the United States

Up to this point, the discussion has not sought to discuss the actual likelihood or size of potential terrorist attacks in the United States. Instead, the analysis has worked backward from the size of counterterrorism budgets (building on assumptions about risk reduction) to infer how much damage from terrorist attacks would

need to be mitigated to justify the budget. But here, we tackle the question of the likelihood and size of prospective terrorist attacks in the United States. The underlying question is: are such attacks likely to be numerous and large enough to justify the size of the counterterrorism budget?

How Many Disclosed Terrorists?

Since September 11, 2001, 54 cases have come to light that involve Islamist terrorists who were either planning to commit, or actually did commit, violence within the United States (Mueller 2014). In the twelve years since the September 11 attacks, these terrorists have managed to kill 19 Americans, 16 of them with gunfire, and three with primitive homemade bombs. It is likely that, without counterterrorism spending, more of the plots would have reached fruition and caused damage, but it seems implausible that many of them would have resulted (for example) in $500 million in damage. Some of the plots being hatched were relatively small-scale: for example, setting off a grenade in a trash bin in a mall or taking some potshots at a military recruitment center. Plotters in other cases did sometimes harbor visions of toppling large buildings, destroying airports, setting off dirty bombs, or bringing down the Brooklyn Bridge, but such visions seem to have been well beyond their actual capacities (Mueller and Stewart 2011, pp. 83–89; Mueller and Stewart 2012). Most other terrorists do not have as destructive an agenda as the Islamist ones (a notable exception from the years before 9/11 would be the Oklahoma City bombing in 1995), and so their inclusion would probably not change the general considerations of our analysis very much.

How Many Undisclosed Terrorists?

It is sometimes argued that there are many terrorist plots out there in addition to the ones that have entered the public record, and that information about these plots cannot be disclosed for various reasons. One possible source of this ominous feeling is the "Threat Matrix," a spreadsheet in which thousands of leads are paraded daily before top government decision makers. As Graff (2011, pp. 19, 489, 345) vividly describes the process, it comes off as "a catalogue of horrors," as the "daily looming prognoses of Armageddon." According to former CIA Director George Tenet (2007, p. 232), "You could drive yourself crazy believing all or even half of what was in it." This perspective is stressed as well by another insider who notes that the constant stream of scary information, combined with a "want of actionable intelligence," led not to the conclusion that there was nothing to find, but rather to "an aggressive, panicked attitude that assumed the worst about threats" (Goldsmith 2007).

But given this mentality and given the record of the terrorist events that have actually occurred since 9/11, the claims of many averted but undisclosed terrorist attacks should at this point be taken with a grain of salt. Few, if any, of the disclosed terrorist plots remotely justify panic, and it is difficult to believe that it is only the big ones we haven't heard about. Moreover, when terrorist plots have been blocked, it certainly appears that policing agencies generally have been

anything but tight-lipped about their accomplishment (Johnston and Shane 2009; Graff 2011, pp. 368; Aaronson 2013, pp. 202–6, 215–16). The *Washington Post*'s Dana Priest says she frequently heard claims about averted but undisclosed plots from counterterrorism agencies and government officials. In response, she says she "asked them to share with us anything they could, plots that were foiled that we could put in the paper because we didn't have many examples. We said give us things, just in generalities." But "we didn't receive anything back" (National Public Radio 2010).

Underestimating Terrorism because of Alternative Charges?

Terrorism arrests and indictments are made, of course, only when prosecutors think they have evidence to obtain a conviction. However, it could be that authorities have encountered aspirational terrorists, but lacking enough evidence to convict on terrorism charges, instead have levied lesser charges, such as immigration violations. One FBI estimate, in fact, is that only one terrorism case in four leads to terrorism charges, while simpler criminal charges are used with other cases (Graff 2011, pp. 420–21, 557; but see also Nakashima 2013).

A number of people assumed to be potential terrorists have thus been picked up and then convicted on minor charges. Some of these were deported and so have presumably vanished from the picture. Others, however, were given short sentences and then released. None in this latter group appear, upon release, to have ever done or plotted terrorism later: they do not come up in the 54 cases of terrorism since 2011 (Mueller 2014). This means either 1) that they were never potential terrorists in the first place, or 2) that all their terrorism leanings evaporated when they were picked up on minor charges. If the latter, they don't seem to have been terribly dedicated to the cause.

What about Deterred Terrorists?

One can hypothesize that a number of potential terrorists pulled back from actually committing violence because they were intimidated by security measures. For example, insofar as military installations have been targeted in the United States, these have typically been recruiting offices within cities, not military bases, which are far more secure.

However, although security measures may have complicated terrorism planning in some cases, no dedicated would-be terrorist would have much difficulty finding other potential targets if the goal is to make a statement by killing people or destroying property. If security measures deter terrorism, they must primarily do so not because they are so effective, but rather because the would-be terrorists are not very dedicated in the first place and are rather easily dissuaded. Moreover, the notion that many terrorists are deterred needs to grapple with why were there so few plots in the months and years following the September 11 attacks, before enhanced security measures were effectively deployed. It may be that counterterrorism efforts are more likely to waylay impotent scheming than to prevent consummated violence.

What about the Smart Terrorists?

There is an argument that we only catch the incompetent terrorists, while the smart ones get away. But more than a decade after 9/11, one needs to explain why all these lurking would-be smart terrorists have not yet actually *done* something. It may be that smart people are dissuaded from committing terrorism not because of counterterrorism measures, but rather from the realization that no matter how deeply-felt their grievances and outrage, expressing them in random or semi-random civilian destruction is likely to be counterproductive to their cause (Abrahms 2011).

Conclusion

It is inevitably difficult to evaluate the appropriate level of counterterrorism spending, and even harder to evaluate the extent to which marginal increases in that spending are justified. But given the great expansion in resources devoted to counterterrorism spending as well as the opportunity costs of time and privacy, grappling with the question of how much counterterrorism spending is appropriate is not a question to be sidestepped.

Our approach supplies a framework for dealing with this question. Our calculations do suggest that for much counterterrorism spending to be justified, it would need to avert an implausibly high number of attacks of very substantial size every year. However, our analysis also suggests that FBI counterterrorism spending would be justified if, on its own, it prevents one attack of 9/11 magnitude every 60 years. Some might find that plausible enough to justify the FBI expenditures. Others might be inclined to consider 9/11 to be an aberration, stressing the fact that the human damage it inflicted was seven times higher than any other terrorist attack in history whether in war zones or not, while the property destruction it inflicted was even more exceptional.

Similarly, returning to what we have called "extended costs," one might argue that if counterterrorism spending undercuts the political impetus for a costly protracted war, then such spending would be worthwhile. But although counterterrorism spending could in theory be a substitute for war, after 9/11 the two were complementary public policies as the United States did both.

Or one might argue that a heightened anxiety about terrorism, as opposed to other potential dangers, generates fear and engages the emotions much more than other hazards to human life, and therefore it justifies a relatively high level of spending. But that conclusion may be too broad: for example, the terrorist tragedy at Fort Hood in 2009 did not seem to cause widespread fear and anxiety, nor did it have much of an economic effect. It could also be maintained that officials in charge of public safety—the foundational reason for government—need to be roughly, if not necessarily completely, risk neutral (Sunstein 2006). Indeed, the Office for Management and Budget (1992) and most international regulators recommend a risk-neutral approach. However, homeland security decisions are often exceedingly risk averse: few, if any, government agencies, including

the Nuclear Regulatory Commission and the Environmental Protection Agency, exhibit anywhere near this level of risk aversion in their public decision making (for an expected utility analysis of United States homeland security spending, see Stewart, Ellingwood, and Mueller 2011). From this perspective, it is irresponsible for public officials and regulators to give in to political and emotional pressures and spend public and social resources on measures that save few lives when the same resources, used otherwise, might save many (Mueller and Stewart 2014).

The case for or against the increased levels of counterterrorism spending necessarily rests on such arguments. Another important reason for considering costs and benefits of counterterrorism spending is that it puts other types of security expenditures into contrast. For example, whatever one's conclusion about the benefit–cost ratio of the FBI's counterterrorism efforts, they are certainly superior to some other security measures. For example, the Transport Security Administration's Federal Air Marshal Service and its full body scanner technology together are nearly as costly as the entire FBI counterterrorism budget, but their risk reduction over the alternatives appears to be negligible (Mueller and Stewart 2011; Stewart and Mueller 2011, 2013a, 2013b). Moreover, the body scanner technology only deals with specific threats associated with hijacking and body-borne bombs on aircraft. In comparison, enhanced FBI expenditures would seem a preferable option: they deal with all terrorism threats, almost certainly do reduce the terrorism threat, and can be rapidly deployed or redeployed as threats emerge or evolve.

References

Aaronson, Trevor. 2013. *The Terror Factory: Inside the FBI's Manufactured War on Terrorism*. Brooklyn, NY: Ig Publishing.

Abrahms, Max. 2011. "Does Terrorism Really Work? Evolution in the Conventional Wisdom since 9/11." *Defence and Peace Economics* 22(6): 583–94.

Bunn, Matthew, Anthony Weir, and John P. Holdren. 2003. *Controlling Nuclear Warheads and Materials: A Report Card and Action Plan*. Washington, DC: Nuclear Threat Initiative and the Project on Managing the Atom, Harvard University.

Bureau of Economic Analysis, US Department of Commerce. 2013. "Economic Growth Widespread across Metropolitan Areas in 2012." News Release, September 17.

Federal Bureau of Investigation, US Department of Justice. 2013. *FY 2014: Authorization and Budget Request to Congress*. April. http://www.justice.gov/jmd/2014justification/pdf/fbi-justification.pdf.

Goldsmith, Jack. 2007. *The Terror Presidency: Law and Judgment Inside the Bush Administration*. New York: W. W. Norton & Co.

Graff, Garrett M. 2011. *The Threat Matrix: The FBI at War in the Age of Global Terror*. New York: Little, Brown, and Co.

Johnston, David, and Scott Shane. 2009. "Terror Case Is Called One of the Most Serious in Years." *New York Times*, September 24.

LaTourrette, Tom, David R. Howell, David E. Mosher, and John MacDonald. 2006. *Reducing Terrorism Risk at Shopping Centers: An Analysis of Potential Security Options*. Technical Report TR-41. Santa Monica, CA: RAND.

Meade, Charles, and Roger C. Molander. 2006. *Considering the Effects of a Catastrophic Terrorist Attack*. Technical Report TR-391-CTRMP. Santa Monica, CA: RAND.

Mueller, John. 2010. *Atomic Obsession: Nuclear Alarmism from Hiroshima to Al Qaeda*. New York: Oxford University Press.

Mueller, John, ed. 2014. *Terrorism since 9/11: The American Cases*. Columbus, OH: Mershon Center, Ohio State University. http://politicalscience.osu.edu/faculty/jmueller//since.html.

Mueller, John, and Mark G. Stewart. 2011. *Terror, Security, and Money: Balancing the Risks, Benefits, and Costs of Homeland Security*. New York: Oxford University Press.

Mueller, John, and Mark G. Stewart. 2012. "The Terrorism Delusion: America's Overwrought Response to September 11." *International Security* 37(1): 81–110.

Mueller, John, and Mark G. Stewart. 2014. "Terrorism and Counterterrorism in the US: The Question of Responsible Policy-Making." *International Journal of Human Rights* 18(2): 228–40.

Nakashima, Ellen. 2013. "Audit: Justice Department Office Overstated Terrorism Conviction Statistics." *Washington Post*, September 17.

National Public Radio. 2010. "Intelligence Failures in 'Top Secret America.'" *Talk of the Nation,* transcript, July 19.

National Research Council of the National Academies. 2010. *Review of the Department of Homeland Security's Approach to Risk Analysis*. Washington, DC: National Academies Press.

Office of Management and Budget. 1992. *Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programs (Revised)*, October 29. Circular No. A-94. Washington, DC.

Robinson, Lisa A., James K. Hammitt, Joseph E. Aldy, Alan Krupnick, and Jennifer Baxter. 2010. "Valuing the Risk of Death from Terrorist Attacks." *Journal of Homeland Security and Emergency Management* 7(1): Article 14.

Stewart, Mark G., Bruce R. Ellingwood, and John Mueller. 2011. "Homeland Security: A Case Study in Risk Aversion for Public Decision-Making." *International Journal of Risk Assessment and Management* 15(5/6): 367–86.

Stewart, Mark G., and John Mueller. 2011. "Cost-Benefit Analysis of Advanced Imaging Technology Full Body Scanners for Airline Passenger Security Screening." *Journal of Homeland Security and Emergency Management* 8(1): Article 30.

Stewart, Mark G., and John Mueller. 2013a. "Terrorism Risks and Cost-Benefit Analysis of Aviation Security." *Risk Analysis* 33(5): 893–908.

Stewart, Mark G., and John Mueller. 2013b. "Aviation Security, Risk Assessment, and Risk Aversion for Public Decisionmaking." *Journal of Policy Analysis and Management* 32(3): 615–33.

Stiglitz, Joseph E., and Linda J. Bilmes. 2008. *The Three Trillion Dollar War*. New York: Norton.

Sunstein, Cass R. 2006. "Misfearing: A Reply." University of Chicago, John M. Olin Program in Law and Economics Working Paper No. 274.

Tenet, George. 2007. *At the Center of the Storm: My Years at the CIA*. With Bill Harlow. New York: HarperCollins.

Journal of Economic Perspectives—Volume 28, Number 3—Summer 2014—Pages 249–256

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

Gary Clyde Hufbauer, Cathleen Cimino, and Tyler Moran evaluate "NAFTA at 20: Misleading Charges and Positive Achievements." "In truth the claims on both sides of the NAFTA issue 20 years ago were overblown. Since the Mexican economy is less than one-tenth the size of the US economy, it is not plausible that trade integration could dramatically shape the giant US economy, even though integration could exert a substantial impact on the relatively small Mexican economy. But exaggeration and sound bites are the weapons of political battle, and trade agreements have been on the front line for two decades." "Ample econometric evidence documents the substantial payoff from expanded two-way trade in goods and services. Through multiple channels, benefits flow both from larger exports and larger imports. . . . The

■ *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.28.3.249. doi=10.1257/jep.28.3.249

channels include more efficient use of resources through the workings of comparative advantage, higher average productivity of surviving firms through 'sifting and sorting,' and greater variety of industrial inputs and household goods. . . . As a rough rule of thumb, for advanced nations, like Canada and the United States, an agreement that promotes an additional $1 billion of two-way trade increases GDP by $200 million. For an emerging country, like Mexico, the payoff ratio is higher: An additional $1 billion of two-way trade probably increases GDP by $500 million. Based on these rules of thumb, the United States is $127 billion richer each year thanks to 'extra' trade growth, Canada is $50 billion richer, and Mexico is $170 billion richer. For the United States, with a population of 320 million, the pure economic payoff is almost $400 per person." Peterson Institute for International Economics, May 2014, Number PB14-13. http://www.piie.com/publications/pb/pb14-13.pdf.

In Chapter 3 of its *World Economic Outlook*, the IMF considers "Perspectives on Global Real Interest Rates." "Real interest rates worldwide have declined substantially since the 1980s and are now in slightly negative territory. Common factors account for much of these movements, highlighting the relevance of global patterns in saving and investment. Since the late 1990s, three factors appear to account for most of the decline. First, a steady increase in income growth in emerging market economies during 2000–07 led to substantially higher saving rates in these economies. Second, the demand for safe assets increased, largely reflecting the rapid reserve accumulation in some emerging market economies and increases in the riskiness of equity relative to bonds. Third, there has been a sharp and persistent decline in investment rates in advanced economies since the global financial crisis. This chapter argues that global real interest rates can be expected to rise in the medium term, but only moderately, since these three factors are unlikely to reverse substantially. . . . In summary, real [interest] rates are expected to rise. However, there are no compelling reasons to believe in a quick return to the average level observed during the mid-2000s (that is, about 2 percent)." April 2014. http://www.imf.org/external/Pubs/ft/weo/2014/01/pdf/c3.pdf.

Melissa S. Kearney and Benjamin H. Harris have edited an e-book, *Policies to Address Poverty in America,* with 14 short essays on specific policies. As one example, Robert I. Lerman advocates "Expanding Apprenticeship Opportunities in the United States." "Today apprentices make up only 0.2 percent of the U.S. labor force, far less than in Canada (2.2 percent), Britain (2.7 percent), and Australia and Germany (3.7 percent). . . . While total annual government funding for apprenticeship in the United States is only about $100 to $400 per apprentice, federal, state, and local annual government spending per participant for two-year public colleges is approximately $11,400. Not only are government outlays sharply higher, but the cost differentials are even greater after accounting for the higher earnings (and associated taxes) of apprentices compared to college students." "Stimulating a sufficient increase in apprenticeship slots is the most important challenge. Although it is easy to cite examples of employer reluctance to train, the evidence from South Carolina and Britain suggests that a sustained, business-oriented marketing effort can persuade a large number of employers to participate in apprenticeship training. Both programs

were able to more than quadruple apprenticeship offers over about five to six years." Hamilton Project, Brookings Institution. 2014, http://www.brookings.edu/~/media/research/files/papers/2014/06/19_hamilton_policies_addressing_poverty/policies_address_poverty_in_america_full_book.

Aaron Edlin and Rebecca Haw discuss "Cartels by Another Name: Should Licensed Occupations Face Antitrust Scrutiny?" "Once limited to a few learned professions, licensing is now required for over 800 occupations. And once limited to minimum educational requirements and entry exams, licensing board restrictions are now a vast, complex web of anticompetitive rules and regulations. . . . State-level occupational licensing is on the rise. In fact, it has eclipsed unionization as the dominant organizing force of the U.S. labor market. While unions once claimed 30% of the country's working population, that figure has since shrunk to below 15%. Over the same period of time, the number of workers subject to state-level licensing requirements has doubled; today, 29% of the U.S. workforce is licensed and 6% is certified by the government. The trend has important ramifications. Conservative estimates suggest that licensing raises consumer prices by 15%. There is also evidence that professional licensing increases the wealth gap; it tends to raise the wages of those already in high-income occupations while harming low-income consumers who cannot afford the inflated prices." "We contend that the state action doctrine should not prevent antitrust suits against state licensing boards that are comprised of private competitors deputized to regulate and to outright exclude their own competition, often with the threat of criminal sanction." *University of Pennsylvania Law Review,* April 2014, pp. 1093–1164. http://www.pennlawreview.com/print/162-U-Pa-L-Rev-1093.pdf.

The US Bureau of Labor Statistics has published "The First Hundred Years of the Consumer Price Index: A Methodological and Political History." "Of all the economic statistics produced by the U.S. federal government, none has a direct impact on the lives of everyday Americans quite like the Consumer Price Index (CPI). Numerous government programs, such as Social Security benefits, are adjusted each year on the basis of changes to the CPI. Countless contracts—whether business agreements, government obligations, leases, or court orders—also utilize the CPI, to adjust the dollar amounts associated with these settlements. For some, the CPI seems to be a rather difficult and abstract thing to understand. Others view the index with suspicion, a statistic produced by the recondite, esoteric labors of government economists and statisticians. . . . This article presents a history of the creation and evolution of the CPI: a history of both how the Bureau of Labor Statistics (BLS, the Bureau) has gone about measuring the change in the cost of purchasing some mix of consumer goods and services and how the CPI has been used over approximately the previous 100 years. . . . The article tells the history of the CPI in seven short, self-contained minihistories. The story begins in the late 19th century, proceeds through World War I, the New Deal, World War II, the postwar era, and the 1960s and 1970s, and closes with events that took place from the 1980s through 2004." *Monthly Labor Review*, April 2014. http://www.bls.gov/opub/mlr/2014/article/the-first-hundred-years-of-the-consumer-price-index.htm. The same issue includes "One Hundred Years of

Price Change: The Consumer Price Index and the American Inflation Experience" at http://www.bls.gov/opub/mlr/2014/article/one-hundred-years-of-price-change-the-consumer-price-index-and-the-american-inflation-experience.htm.

Complements of JEP

Chiara Criscuolo, Peter N. Gal, and Carlo Menon compile empirical evidence concerning "The Dynamics of Employment Growth: New Evidence from 18 Countries." "[N]ot all small businesses are net job creators, showing that only young businesses—predominantly small—create a disproportionate number of jobs, confirming recent evidence for the United States. When disentangling the role of entry from the role of expansion of incumbent young firms, the data clearly shows that entry explains most of the contribution to job creation, followed by start-ups (i.e., firms that are less than three year old). While this remains true even during the recent great recession, the data shows a sharp decline in the contribution of entry and young firms to aggregate employment growth during the recession. More generally, the findings point to a decline in start-up rates over the past decade across all countries considered, which gives cause for concern, given their strong contribution to job creation." OECD Science, Technology and Industry Policy Papers No. 14, May 21, 2014. http://www.oecd-ilibrary.org/science-and-technology/the-dynamics-of-employment-growth_5jz417hj6hg6-en. This evidence and discussion is a useful complement to the paper in this issue "The Role of Entrepreneurship in US Job Creation and Economic Dynamism," by Ryan Decker, John Haltiwanger, Ron Jarmin, and Javier Miranda.

The *African Economic Outlook 2014*, published by the African Development Bank Group, OECD, and the UN Development Programme, devotes a special section and several chapters to how global supply chains could boost economic growth in Africa. "In the past, for a country to industrialise it had to develop the domestic capacity to perform all major steps in the value chains of complex manufactured products. Today, through linking into an international production network, countries can establish a specific section of a product's value chain without having all the upstream capabilities in place. These remain elsewhere and are linked through shipments of intermediate products and communication of the know-how necessary for the specific step in the value chain present in the country. . . . Through participation in a value chain, countries and firms can acquire new capabilities that make it possible to upgrade, i.e. to capture a higher share of the value added in a global value chain. . . . Despite their name, global value chains exhibit high regional concentration, which is shrinking slowly. Africa does not play a significant role yet. When measuring the linkages between major supply-chain traders, the strongest relationships can be found within the regional blocks of East Asia, Europe and North America. About 85% of global value chain (GVC) trade in value added takes place in and around these three hubs. While other regions remain marginal, their share has increased from only 10% in 1995 to 15% in 2011. Africa's share in GVC participation increased from 1.4% to 2.2%

during the same time. . . . The global business processing outsourcing market was forecast to grow 5.1% in 2013 and reach USD 304 billion. The race is on among countries such as Egypt, Kenya, Ghana, Mauritius, South Africa, Tunisia and Uganda to become the new 'India' in Africa using incentives and special economic zones to develop their outsourcing sectors." http://www.africaneconomicoutlook.org/fileadmin/uploads/aeo/2014/PDF/E-Book_African_Economic_Outlook_2014.pdf. This study complements the two-paper symposium on "Global Supply Chains" in the Spring 2014 issue of this journal.

I contributed "Economics and Morality" to *Finance & Development*. "Economists prefer to sidestep moral issues. They like to say they study trade-offs and incentives and interactions, leaving value judgments to the political process and society. But moral judgments aren't willing to sidestep economics. Critiques of the relationship between economics and moral virtue can be grouped under three main headings: To what extent does ordinary economic life hold a capacity for virtue? Is economic analysis overstepping its bounds into zones of behavior that should be preserved from economics? Does the study of economics itself discourage moral behavior? . . ." "I have become wary over the years of questions framed in a way that seeks to pit economics against moral virtue in a winner-takes-all brawl. No economist would recommend consulting an economics textbook as a practical source of transcendent moral wisdom. As the recent global economic crisis reminded anyone who needed reminding, economics doesn't have answers for all of the world's economic problems. But to be fair, moral philosophers don't have answers for all the world's spiritual and ethical problems. . . . Economists cannot banish the importance of moral issues in their field of study and should not seek to do so. But when moral philosophers consider topics that touch on the ordinary business of life, they cannot wish away or banish the importance of economics either." June 2014, vol. 51, no. 2, pp. 34–88. http://www.imf.org/external/pubs/ft/fandd/2014/06/pdf/taylor.pdf.

Some of the themes in this essay follow-up on past JEP articles, including the two-paper symposium on "Economics and Moral Virtues" in the Fall 2013 issue.

Interviews and Speeches

Douglas Clement interviews Glenn Loury, with some emphasis on the economics of discrimination. Here's Loury: "[S]uppose I have a regression equation with wages on the left-hand side and a number of explanatory variables—like schooling, work experience, mental ability, family structure, region, occupation and so forth—on the right-hand side. These variables might account for variation among individuals in wages, and thus one should control for them if the earnings of different racial or ethnic groups are to be compared. One could put many different variables on the right-hand side of such a wage regression. Well, many of those right-hand-side variables are determined within the very system of social interactions that one wants to understand if one is to effectively explain large and persistent earnings differences between groups. That is, on the average, schooling, work experience, family

structure or ability (as measured by paper and pencil tests) may differ between racial groups, and those differences may help to explain a group disparity in earnings. But those differences may to some extent be a consequence of the same structure of social relations that led to employers having the discriminatory attitudes they may have in the work place toward the members of different groups. So, the question arises: Should an analyst who is trying to measure the extent of 'economic discrimination' hold the group accountable for the fact that they have bad family structure? Is a failure to complete high school, or a history of involvement in a drug-selling gang that led to a criminal record, part of what the analyst should control for when explaining the racial wage gap—so that the uncontrolled gap is no longer taken as an indication of the extent of unfair treatment of the group?" *The Region,* The Federal Reserve Bank of Minneapolis, June 2014, pp. 12–25. http://www.minneapolisfed.org/pubs/region/14-06/region_june_2014_interview_glenn_loury.pdf.

David A. Price interviews Mark Gertler, with some focus on the dynamics of financial crisis and the Great Recession. On the concept of "financial accelerators": "That's what we wanted to capture with the financial accelerator, that is, the mutual feedback between the real sector and the financial sector. We also wanted to capture the primary importance of balance sheets—when balance sheets weaken, that causes credit to tighten, leading to downward pressure on the real economy, which further weakens balance sheets. I think that's what helped to develop the concept of financial accelerators one saw in the financial crisis. . . . I didn't speak to Bernanke a lot during the height of the crisis. But one moment I caught him, asked him how things were going, and he said, 'Well, on the bright side, we may have some evidence for the financial accelerator.'" On the high level of excess reserves at the Fed: "The Fed now is acting as an investment bank, and it's taking over those activities. Instead of Lehman Brothers holding these mortgage-backed securities, the Fed is. And the Fed is issuing deposits, if you will, against these securities, the same way these private financial institutions did. It's easier for the Fed, because it can issue essentially risk-free government debt, and these other institutions couldn't. . . . It's possible, as interest rates go up, that the Fed could take some capital losses, as private financial institutions do. But the beauty of the Fed is it doesn't have to mark to market; it can hold these assets until maturity, and let them run off. So I'm in a camp that thinks there's been probably a little too much preoccupation with the size of the balance sheet." *Econ Focus*, Federal Reserve Bank of Richmond, Fourth Quarter 2013, pp. 32–36. Available at: https://www.richmondfed.org/publications/research/econ_focus/2013/q4/q4.cfm.

Jason Furman asks "Whatever Happened to the Great Moderation?" "Disaggregating the GDP data, the reduced volatility of consumption is one of the major sources of the Great Moderation—and this reduced volatility has continued to hold up during and after the Great Recession, especially in consumer durables. The continued stability in consumption stands in contrast to other components of GDP like business fixed investment, which became less volatile during the initial Great Moderation but has since at least partially reverted to its earlier volatility. . . . From 1960 to 1984, inventories were quite volatile, and were also procyclical,

meaning that when sales increased, inventories also increased, further contributing to the volatility of production. During the post-1984 Great Moderation period, inventory investment itself became much less volatile, and the previous relationship between inventories and sales reversed, so that the two became negatively correlated. Focusing specifically on durable goods, the change in the covariance between inventories and sales accounts for nearly half of the decline in the variance in durable goods output. However, including the Great Recession, it appears that the relationship between output, sales and inventories partially reverted to the pre-Great Moderation pattern. The covariance of inventories and sales turned positive again, suggesting that improved inventory management was not enough to cushion the massive blow of the Great Recession, and in fact exacerbated it." Address to the 23rd Annual Hyman P. Minsky Conference on the State of the US and World Economies. April 10, 2014. http://www.whitehouse.gov/sites/default/files/docs/2014-04-10-minsky-conference_speech.pdf.

Lawrence Summers delivered the keynote address at the National Association of Business Economists Policy Conference on the subject of "U.S. Economic Prospects: Secular Stagnation, Hysteresis, and the Zero Lower Bound." "I think it is fair to say that six years ago, macroeconomics was primarily about the use of monetary policy to reduce the already small amplitude of fluctuations about a given trend, while maintaining price stability. That was the preoccupation. It was supported by historical analysis emphasizing that we were in a great moderation, by policy and theoretical analysis suggesting the importance of feedback rules, and by a vast empirical program directed at optimizing those feedback rules. Today, we wish for the problem of minimizing fluctuations around a satisfactory trend. . . . I shall argue three propositions. First, as the United States and other industrial economies are currently configured, simultaneous achievement of adequate growth, capacity utilization, and financial stability appears increasingly difficult. Second, this is likely to be related to a substantial decline in the equilibrium or natural real rate of interest. Third, addressing these challenges requires different policy approaches than are represented by the current conventional wisdom." *Business Economics*, 2014, vol. 49, no. 2, pp. 65–73, http://larrysummers.com/wp-content/uploads/2014/06/NABE-speech-Lawrence-H.-Summers1.pdf.

Discussion Starters

The Global Environmental Alert Service of the United Nations Environment Programme reports on "Sand, Rarer than One Thinks." "Globally, between 47 and 59 billion tonnes of material is mined every year, of which sand and gravel, hereafter known as aggregates, account for both the largest share (from 68% to 85%) and the fastest extraction increase . . ." "A conservative estimate for the world consumption of aggregates exceeds 40 billion tonnes a year. This is twice the yearly amount of sediment carried by all of the rivers of the world, making humankind the largest of the planet's transforming agent with respect to aggregates . . ." "Thus, the world's

use of aggregates for concrete can be estimated at 25.9 billion to 29.6 billion tonnes a year for 2012 alone. This represents enough concrete to build a wall 27 metres high by 27 metres wide around the equator." March 2014. http://www.unep.org/pdf/UNEP_GEAS_March_2014.pdf.

Kenneth Button makes the case for "Really Opening Up the American Skies." "The deregulation of the 1970s, by removing entry quantitative controls, led to a considerable increase in services. It also increased the capability of individuals to access a wider range of destinations from their homes via the hub-and-spoke system of routings that emerged. This pattern has been reversed since 2007. The largest 29 airports in the United States lost 8.8 percent of their scheduled flights between 2007 and 2012, but medium-sized airports lost 26 percent and small airports lost 21.3 percent. . . . In sum, the 1978 Airline Deregulation Act only partially liberalized the U.S. domestic airline market. One important restriction that remains is the lack of domestic competition from foreign carriers. The U.S. air traveler benefited from the country being the first mover in deregulation, and this provided lower fares and consumer-driven service attributes some 15–20 years before they were enjoyed in other markets; the analogous reforms in Europe only fully materialized after 1997. But the world has changed, and so have the demands of consumers and the business models adopted by the airlines. . . . But remaining regulations still limit the amount of competition in the market and, with this, the ability of travelers to enjoy even lower fares and a wider range of services." *Regulation*, Spring 2014, pp. 40–45 http://object.cato.org/sites/cato.org/files/serials/files/regulation/2014/4/regulation-v37n1-8.pdf.

What's the most efficient order for passengers to board an airplane? R. John Milne and Alexander R. Kelly explain how the theory has developed and offer their own proposal in "A New Method for Boarding Passengers onto an Airplane." "Steffen (2008) presents an optimum boarding method that assigns passengers to a specific numerical position in line that depends upon their ticketed seat location. . . . For ease of explanation, this paper assumes a 20 row airplane with six seats per row. To get to their seats, passengers walk down an aisle which separates the right side of the plane from the left. . . . Steffen's model places the first passenger to board the plane in a window seat in the last (20th) row in the last (6th) column in the window seat . . . The next passenger is also seated along the window, two rows in front of the first passenger . . . This process continues for one side of the plane and then repeats on the other side. . . . The key aspect of our proposed method is that it assigns airplane passengers to seats so that their carry-on luggage is spread roughly evenly throughout the plane. This reduces the time passengers take to find available storage in the overhead bins when storing their luggage. . . . Given that airlines incur a cost of about $30 per ground minute and Delta Airlines conducts 5800 flights per day, a reduction in boarding time of 0.16 min per flight would translate to a cost savings of about $10 million annually for a large airline such as Delta." *Journal of Air Transport Management*, January 2014, vol. 34, pp. 93–100, http://www.sciencedirect.com/science/article/pii/S0969699713001166. The Steffen (2008) article, "Optimal Boarding Method for Airline Passengers," also appeared in the *Journal of Air Transport Management* (14, pp. 146–50), and is available at http://arxiv.org/pdf/0802.0733.pdf.