*The Journal of*

# *Economic Perspectives*

*A journal of the*
*American Economic Association*

*Spring 2012*

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*
*e-mail: <jep@jepjournal.org>*

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

Composed by American Economic Association Publications, Pittsburgh, Pennsylvania, U.S.A.
Printed by R. R. Donnelley Company, Jefferson City, Missouri, 65109, U.S.A.

*THE JOURNAL OF ECONOMIC PERSPECTIVES* (ISSN 0895-3309), Spring 2012, Vol. 26, No. 2. The *JEP* is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $20.00, $30.00, or $40.00 depending on income; for an additional $15.00, you can receive this journal in print. E-reader versions are free. For details and further information on the AEA go to <http://www.vanderbilt.edu/AEA/>. Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives,* 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

*The Journal of*
**_Economic Perspectives_**

# Contents

*Volume 26 • Number 2 • Spring 2012*

**Symposia**

***100 Issues of JEP***

***International Trade***

**Articles**

**Features**

## Statement of Purpose

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

## Policy on Data Availability

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

## Policy on Disclosure

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

# Journal of Economic Perspectives
### Advisory Board

*Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 3–18*

# *The Journal of Economic Perspectives* at 100 (Issues)

David Autor

When I was a graduate student, I discovered that the *Journal of Economic Perspectives* embodied much of what I love about the field of economics: the clarity that pierces rhetoric to seek the core of a question; the rigor to identify the causal relationships, tradeoffs, and indeterminancies inherent in a problem; the self-assurance to apply the disciplinary toolkit to problems both sacred and profane; and the force of logic to reach conclusions that might be unexpected, controversial, or refreshingly bland. While the fire-hose of theory, method, and data that drenched me daily in graduate classes was variously exhilarating and dispiriting, the *JEP* symposia that I read in my spare time nurtured my enthusiasm for economics, reassuring me that if I could survive the deluge of graduate study, I'd love one day being on the other end of the fire-hose.

It never occurred to me in those years that one day I would edit the journal. While doing so is a privilege and a pleasure, I equally confess that it's no small weight to be the custodial parent of one of our profession's most beloved offspring. No less intimidating is the task of stipulating what this upstart youth has accomplished in its first 25 years and 100 issues in print. Like any empiricist, I recognize that the counterfactual world that would exist without the *JEP* is unknowable, but my strong hunch is that our profession would be worse off in that counterfactual world. In this essay, I reflect on the journal's accomplishments and articulate some of my own goals for the *JEP* going forward.

■ *David Autor is Professor of Economics, Massachusetts Institute of Technology, and Faculty Research Associate, National Bureau of Economic Research, both in Cambridge, Massachusetts. He has been Editor of the* Journal of Economic Perspectives *since January 2009. His e-mail address is ⟨dautor@mit.edu⟩.*

http://dx.doi.org/10.1257/jep.26.2.3. doi=10.1257/jep.26.2.3

## Measuring Some Effects

In his 1987 "Report of the Editor," published in the *AER Papers and Proceedings* some months prior to the first issue of the journal, Joseph Stiglitz wrote, "We have chosen the name *Economic Perspectives* to capture the journal's twin missions of providing perspective on current economic research, and explaining how economics provides perspective on questions of general interest." This statement presaged two roles that the *JEP* has come to serve. One role is to supply a vital *intra*-disciplinary conduit for elevating ideas from the depths of sub-disciplinary trenches, opening topics ripe for economic analysis, staging debate on findings and methods, and channeling the Zeitgeist of our prolific but sometimes methodologically abstruse discipline. The other is to provide policymakers, educators, students, and practitioners with a trove of well-reasoned, well-written, and well-chosen analytical essays that focus the lens of economic reasoning on topics across the social sciences.

How well has *JEP* succeeded in these goals? For academic journals, citations are always one plausible starting point, and although *JEP* is not a conventional journal, I'll begin there as well. Of course, citation counts have numerous limitations as a measure of impact: counts vary by field size; bad papers may be cited as counterexamples; and papers are often dutifully cited without being read. In the case of *JEP*, one could also note that researchers might use a *JEP* essay as a starting point to learning more about a topic but then not feel obliged to cite the journal as a primary source of scholarship. Moreover, academic citations largely fail to capture the effect of *JEP* on policymakers, undergraduates, and other interested readers. I was a regular reader of *JEP* for years before I started citing it, and I more often read the articles because I was intrigued by the topic or the author rather than because I was tracking a cited reference.

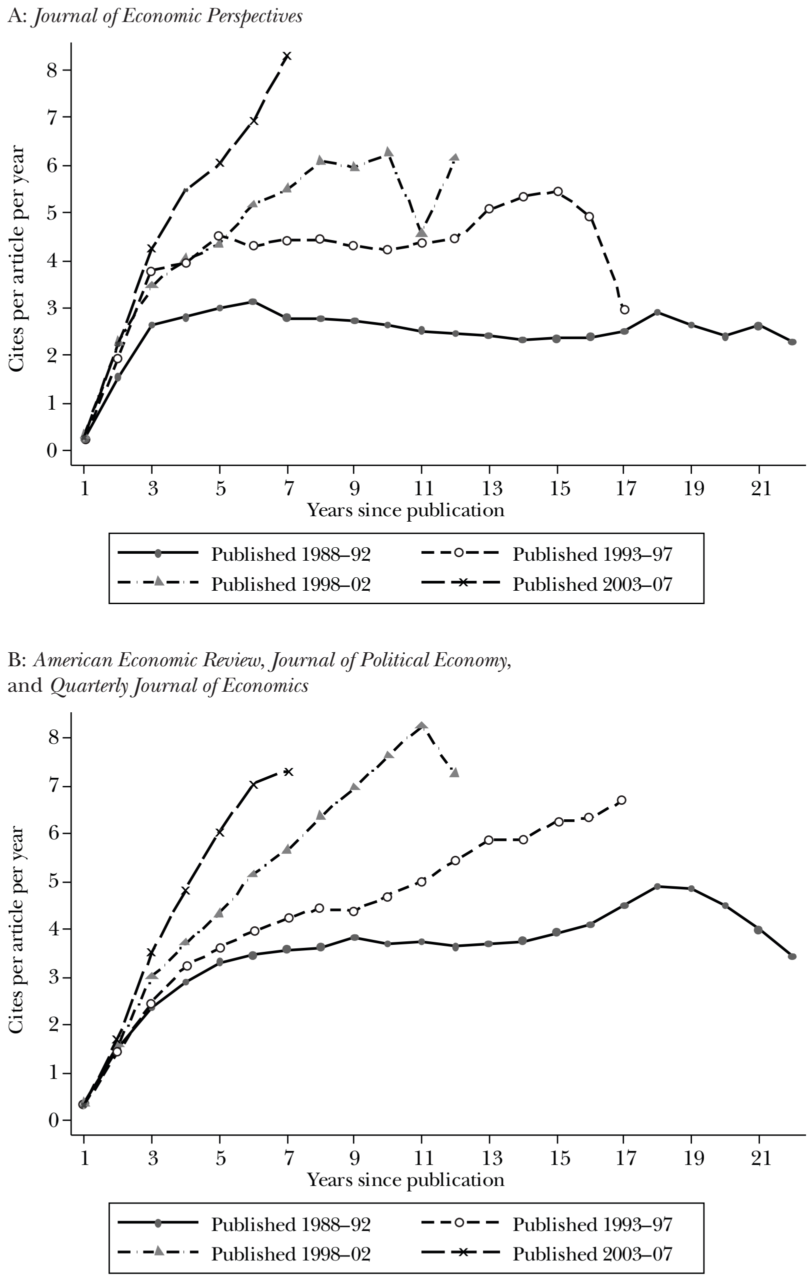
Despite these cautions, citations tell a reasonably encouraging tale of *JEP*'s impact. Figure 1 draws on Thompson-Reuters Web of Science to track average annual citations per article per year for *JEP* articles published from the journal's inception through 2007 in Panel A. For comparison, Panel B summarizes average annual citation rates for articles published in the *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics* during the same years.[1] As the figure shows, while *JEP* is not as highly cited as these three top journals, it's not far off. Indeed, *JEP* articles published between 2003 and 2007 have received

[1] Citation counts are from the Thompson-Reuters Web of Knowledge: Social Sciences Citation Index database for years 1988–2010. Included items from the four journals in Figure 1 are limited to articles. Letters, editorial matter, proceedings papers, and reviews are filtered out by Web of Knowledge. In addition, the *JEP* features "Classroom Games" and "Recommendations for Further Reading" are also excluded from the analysis. In Panel A of the figure, articles from the first two *JEP* issues, Summer and Fall 1987, are averaged in with articles published during 1988 through 1992. I exclude *Econometrica* and *Review of Economic Studies* because these journals are even less-applied relative to *JEP* than the comparison set.

*Figure 1*

**Average Annual Citations by Year since Publication**

*(Social Science Citations Index Citations)*



A: *Journal of Economic Perspectives*

B: *American Economic Review, Journal of Political Economy,* and *Quarterly Journal of Economics* slightly *more* citations per article than articles published in the same years in *AER*, *JPE,* and *QJE*.[2]

One contrast also emerges that reflects the distinct orientation of *JEP* vis-à-vis other journals: *JEP* articles appear to have a shorter half-life. While *JEP* citations appear to peak approximately seven years following publication, Figure 1 suggests that the comparison set of three journals sees little reduction in citations even in the "out years." This contrast strikes me as a healthy intellectual division of labor. *JEP*'s primary goal is to illuminate the research frontier rather than to push it outward (though it sometimes succeeds on the former without really trying), and hence we might expect *JEP* articles to have their biggest impact in their first five to seven years after publication.

Which articles have contributed the most to the *JEP*'s high citation rate? Table 1 provides a list of the top 40 most-cited *JEP* articles. Since citations take time to accumulate, this list naturally favors articles published some years ago.[3] To supplement, Table 2 provides a list of "young upstarts" published in the most recent ten *JEP* volumes, excluding articles published in 2011 and those included in Table 1. I'll return to this list below.

To glimpse *JEP*'s effect beyond the world of scholarly journals, we must resort to other measures. Since economic education is central to *JEP*'s mission, Table 3 provides an estimate of the number of citations to *JEP* articles on class syllabi at the top 100 research universities based upon Google searches of their websites.[4] Notably, 94 of the top 100 research universities use *JEP* in the classroom. And this is probably a substantial undercount: 26 of the schools had five or fewer *JEP* articles on their syllabi (including six with none at all), which we strongly suspect says more about the syllabi that aren't freely available online than it does about *JEP* usage. Nevertheless, we find an average of 43 references per school (including the zeros) to *JEP* articles on course syllabi at these institutions and, happily, no obvious (to my eye) relationship between the methodological orientation of the economics department and its intensity of *JEP* usage.[5]

[2] The rising citation rate for all journals likely reflects in large part the growth of the field of economics.

[3] To wit, only one article in Table 1 was published after 2001. One might think that regression adjusting citations for time since publication would identify additional newly published articles that are likely to accumulate high citation counts over time. But this strategy offers no improvement over a simple citation count. While outlier articles receive an order of magnitude more cites than the average paper, the distance of these outliers from the regression lines bears little relationship to their publication date. For example, the correlation between citations and publication year in the Top 40 list in Table 1 is 0.016.

[4] Top 100 Research Universities as ranked by *U.S. News and World Report* in 2010. Counts are based on a Google search of these universities' websites using the terms "Journal of Economic Perspectives" and "syllabus."

[5] We also tried to search the online syllabi of the top 100 U.S. liberal arts colleges. This search produced a total of 452 citations to *JEP* articles. But here the Internet accessibility issue appears much more severe—only half of these searches produced any hits. Some subsequent hand-checking confirmed, however, that many of these syllabi are gated or nonsearchable through standard means. For example, our search of Barnard College's website found no *JEP* mentions. But a minute of poking around the public but nonsearchable section of Barnard's course content site, immediately yielded two *JEP* articles in the reading list of a Spring 2012 class, "ECON X2010.001 The Economics of Gender."

*Table 1*

**Top 40 Most Cited *JEP* Articles of All Time**

| *Rank* | *Title* | *Authors* | *Year* | *Volume, issue number* | *Citations* |
|---|---|---|---|---|---|
| 1 | Toward a New Conception of the Environment-Competitiveness Relationship | Porter, Michael E.; van der Linde, Claas | 1995 | 9(4) | 657 |
| 2 | Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias | Kahneman, Daniel; Knetsch, Jack L.; Thaler, Richard H. | 1991 | 5(1) | 572 |
| 3 | Contingent Valuation: Is Some Number Better than No Number? | Diamond, Peter A.; Hausman, Jerry A. | 1994 | 8(4) | 524 |
| 4 | Fairness and Retaliation: The Economics of Reciprocity | Fehr, Ernst; Gächter, Simon | 2000 | 14(3) | 490 |
| 5 | Systems Competition and Network Effects | Katz, Michael L.; Shapiro, Carl | 1994 | 8(2) | 448 |
| 6 | Institutions | North, Douglass C. | 1991 | 5(1) | 395 |
| 7 | Quantile Regression | Koenker, Roger; Hallock, Kevin F. | 2001 | 15(4) | 375 |
| 8 | The Boundaries of Multinational Enterprises and the Theory of International Trade | Markusen, James R. | 1995 | 9(2) | 375 |
| 9 | Inside the Black Box: The Credit Channel of Monetary Policy Transmission | Bernanke, Ben S.; Gertler, Mark | 1995 | 9(4) | 365 |
| 10 | The Origins of Endogenous Growth | Romer, Paul M. | 1994 | 8(1) | 365 |
| 11 | Beyond Computation: Information Technology, Organizational Transformation and Business Performance | Brynjolfsson, Erik; Hitt, Lorin M. | 2000 | 14(4) | 350 |
| 12 | Unemployment and Labor Market Rigidities: Europe versus North America | Nickell, Stephen | 1997 | 11(3) | 344 |
| 13 | Choice under Uncertainty: Problems Solved and Unsolved | Machina, Mark J. | 1987 | 1(1) | 338 |
| 14 | Valuing the Environment through Contingent Valuation | Hanemann, W. Michael | 1994 | 8(4) | 332 |
| 15 | Anomalies: Ultimatums, Dictators, and Manners | Camerer, Colin; Thaler, Richard H. | 1995 | 9(2) | 316 |
| 16 | Collective Action and the Evolution of Social Norms | Ostrom, Elinor | 2000 | 14(3) | 313 |
| 17 | Healthy Bodies and Thick Wallets: The Dual Relation between Health and Economic Status | Smith, James P. | 1999 | 13(2) | 311 |
| 18 | The Market for Corporate Control: The Empirical Evidence since 1980 | Jarrell, Gregg A.; Brickley, James A.; Netter, Jeffry M. | 1988 | 2(1) | 295 |
| 19 | New Evidence and Perspectives on Mergers | Andrade, Gregor; Mitchell, Mark; Stafford, Erik | 2001 | 15(2) | 290 |
| 20 | Standing on the Shoulders of Giants: Cumulative Research and the Patent Law | Scotchmer, Suzanne | 1991 | 5(1) | 280 |
| 21 | Organizations and Markets | Simon, Herbert A. | 1991 | 5(2) | 278 |
| 22 | Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades | Bikhchandani, Sushil; Hirshleifer, David; and Welch, Ivo | 1998 | 12(3) | 273 |

*Table 1—continued*

| Rank | Title | Authors | Year | Volume, issue number | Citations |
|---|---|---|---|---|---|
| =23 | Social Norms and Economic Theory | Elster, Jon | 1989 | 3(4) | 272 |
| =23 | Integration of Trade and Disintegration of Production in the Global Economy | Feenstra, Robert C. | 1998 | 12(4) | 272 |
| =23 | Does Studying Economics Inhibit Cooperation? | Frank, Robert H.; Gilovich, Thomas; Regan, Dennis T. | 1993 | 7(2) | 272 |
| =23 | Whom or What Does the Representative Individual Represent? | Kirman, Alan P. | 1992 | 6(2) | 272 |
| =27 | Takeovers: Their Causes and Consequences | Jensen, Michael C. | 1988 | 2(1) | 268 |
| =27 | Political Regimes and Economic Growth | Przeworski, Adam; Limongi, Fernando | 1993 | 7(3) | 268 |
| 29 | Medical Care Costs: How Much Welfare Loss? | Newhouse, Joseph P. | 1992 | 6(3) | 265 |
| 30 | Investment and Hysteresis | Dixit, Avinash | 1992 | 6(1) | 259 |
| 31 | The Resurgence of Growth in the Late 1990s: Is Information Technology the Story? | Oliner, Stephen D.; Sichel, Daniel E. | 2000 | 14(4) | 257 |
| 32 | Why Have Americans Become More Obsese? | Cutler, David M; Glaeser, Edward L.; Shapiro, Jesse M. | 2003 | 17(3) | 250 |
| 33 | Auctions and Bidding: A Primer | Milgrom, Paul | 1989 | 3(3) | 242 |
| 34 | The Contingent Valuation Debate: Why Economists Should Care | Portney, Paul R. | 1994 | 8(4) | 239 |
| 35 | Explaining Bargaining Impasse: The Role of Self-Serving Biases | Babcock, Linda; Loewenstein, George | 1997 | 11(1) | 231 |
| 36 | Endogenous Innovation in the Theory of Growth | Grossman, Gene M.; Helpman, Elhanan | 1994 | 8(1) | 225 |
| 37 | Tightening Environmental Standards: The Benefit-Cost or the No-Cost Paradigm | Palmer, Karen; Oates, Wallace E.; Portney, Paul R. | 1995 | 9(4) | 222 |
| 38 | Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments | Angrist, Joshua D.; Krueger, Alan B. | 2001 | 15(4) | 221 |
| 39 | Divergence, Big Time | Pritchett, Lant | 1997 | 11(3) | 209 |
| =40 | Anomalies: Cooperation | Dawes, Robyn M.; Thaler, Richard H. | 1988 | 2(3) | 206 |
| =40 | Bargaining and Distribution in Marriage | Lundberg, Shelly; Pollak, Robert A. | 1996 | 10(4) | 206 |

While these metrics are intriguing, they miss two important silent readerships of *JEP*. The first is the set of policymakers, practitioners, and economic reporters who consult *JEP* to inform their thinking, even when *JEP* is not referenced explicitly in their later comments or articles. A second audience is the set of readers who regard the *JEP* as kind of economist's *New Yorker* — bedside, beach, and bathroom reading for the social scientist. Many such readers will rarely have occasion to cite

*Table 2*
**Some Highly Cited *JEP* Articles Published since 2002**

| *Title* | *Authors* | *Year* | *Volume, issue number* | *Citations* |
|---|---|---|---|---|
| What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? | Levitt, Steven D.; List, John A. | 2007 | 21(2) | 174 |
| Executive Compensation as an Agency Problem | Bebchuk, Lucian Arye; Fried, Jesse M. | 2003 | 17(3) | 170 |
| Developments in the Measurement of Subjective Well-Being | Kahneman, Daniel; Krueger, Alan B. | 2006 | 20(1) | 165 |
| The Purchasing Power Parity Debate | Taylor, Alan M.; Taylor, Mark P. | 2004 | 18(4) | 164 |
| What Really Matters in Auction Design? | Klemperer, Paul | 2002 | 16(1) | 157 |
| Prediction Markets | Wolfers, Justin; Zitzewitz, Eric | 2004 | 18(2) | 149 |
| The Impact of Social Structure on Economic Outcomes | Granovetter, Mark | 2005 | 19(1) | 147 |
| Identity and the Economics of Organizations | Akerlof, George A.; Kranton, Rachel E. | 2005 | 19(1) | 120 |
| Does Culture Affect Economic Outcomes? | Guiso, Luigi; Sapienza, Paola; Zingales, Luigi | 2006 | 20(2) | 120 |
| Deciphering the Liquidity and Credit Crunch 2007–2008 | Brunnermeier, Markus K. | 2009 | 23(1) | 111 |
| Income, Health, and Well-Being around the World: Evidence from the Gallup World Poll | Deaton, Angus | 2008 | 22(2) | 46 |
| The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics | Angrist, Joshua D.; Pischke, Jörn-Steffen | 2010 | 24(2) | 17 |

the *JEP* articles they've read because the topics are outside their usual spheres of research. But we suspect that these articles are often cited verbally in collegial conversations, in classroom lectures, at dinner table discussions, and in political debate. Such verbal citations are one of the core missions of *JEP*, even if we cannot enumerate them.

## Contributions

### Opening New Topics

While *JEP* is not intended as a frontier research journal, it is hard to escape the impression that it has nevertheless shaped the frontier—not, however, by the usual method of unleashing a trove of regression output or an exquisite new proof. Rather, it has moved the frontier by capturing the scarcest of all academic resources: attention. Perhaps the best example of this phenomenon is the "Anomalies" feature,

*Table 3*
**JEP Citations on Class Syllabi at Top 100 Research Universities**

| *All top 100* | |
|---|---|
| *University* | *Citations* |
| Mean cites | 43.3 |
| Median cites | 22.5 |
| *Twenty-five highest citation counts* | |
| *University* | *Citations* |
| Harvard | 456 |
| NYU | 241 |
| MIT | 222 |
| UC Berkeley | 191 |
| U Wisconsin–Madison | 158 |
| Princeton | 153 |
| American U | 149 |
| Stanford | 143 |
| U Michigan–Ann Arbor | 129 |
| U Maryland–College Park | 121 |
| UC San Diego | 119 |
| Columbia | 114 |
| U Colorado–Boulder | 111 |
| Duke | 104 |
| U Penn | 99 |
| Boston College | 89 |
| U Chicago | 73 |
| U Minnesota | 70 |
| U Texas–Austin | 70 |
| U Washington | 66 |
| Penn State–U Park | 62 |
| UCLA | 52 |
| U Southern California | 46 |
| Yale | 46 |

written by Richard Thaler and a host of noteworthy coauthors, which drew attention to puzzling regularities in individual or market behavior that seemed to defy ready neoclassical explanation. The 19 articles in this series, most published between 1987 and 1995, seemed mildly heretical at the time, perhaps designed to goad hardened adherents of the *homo economicus* model. Three ("The Endowment Effect . . .," "Ultimatums . . .," and "Cooperation") are among the *JEP*'s 40 most-cited of all time, and seven are at or above the 90th percentile of *JEP* citations.[6] It has long been my hunch

[6] *JEP* has published 1,104 articles since 1987, 19 of which were "Anomalies" articles.

that by making it difficult for mainstream economists to ignore the predictable irrationalities in human behavior, the "Anomalies" feature catalyzed the nascent field of behavioral economics. In writing for this article, I discovered that I'm not alone in this view. In their chapter on "Behavioral Economics" in the *Handbook of Philosophy of Science*, Angner and Loewenstein (forthcoming) write, ". . . through his 'Anomalies' columns published in the widely distributed *Journal of Economic Perspectives* and collected in *The Winner's Curse* (1992), Thaler helped accelerate the awareness and acceptance of behavioral economics among mainstream economists."

Indeed, the list of highly cited articles in Table 1 hints that *JEP* has served as a sort of test kitchen for the expansion of our discipline into topics that traditionally lay within the domains of psychology, sociology, political science, and anthropology. Some examples from Table 1 include: cooperation ("Fairness and Retaliation" and "Does Studying Economics Inhibit Cooperation?"); social norms ("Collective Action and the Evolution of Social Norms" and "Social Norms and Economic Theory"); group behavior ("Learning from the Behavior of Others"); institutions ("Institutions" and "Political Regimes and Economic Growth"); and bargaining behavior ("Explaining Bargaining Impasse: The Role of Self-Serving Biases" and "Bargaining and Distribution in Marriage").

Would the economics profession have moved in these directions absent the *JEP* test kitchen? Surely, the answer is "yes." But the movement might not have been as broad across the profession, and perhaps would not have included the same energy or upwelling of talent. In my observation, it takes at most a handful of unconventional, passionate researchers to ignite a research area and in that way to bend the direction of a field and focus the attention of the profession on the problems they bring to the fore. These upstarts are often born of the marriage between raw talent and intellectual mission consummated during Ph.D. training, when the chance influence of advisors, peers, scholars living and dead, personal drive, and whatever currents are in the air act together to push young scholars towards their life's work. I suspect that *JEP* has played the intellectual matchmaker for many intellectual upstarts in the last 25 years—indeed, I saw this among my peers while in graduate school. By channeling promising but unfinished ideas to young scholars when they were least committed to an ideology and hungriest for an agenda, it is my strong hunch that *JEP* has increased the pace of intellectual ferment in our profession.

### Room for Debate

*JEP* has often staged battles between traditional and revisionist economic viewpoints, and it's far from clear that the revisionists have always prevailed. For example, three of *JEP*'s 40 most highly cited articles stem from a single 1994 symposium on the value of contingent valuation as a tool for pricing environmental amenities. The most cited of these articles ("Contingent Valuation—Is Some Number Better than No Number?" by Diamond and Hausman) savaged the contingent valuation methodology. Similarly, in *JEP*'s most cited paper of the past five years, "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" Levitt and List cast a skeptical eye on the influential body of laboratory experiments that find that in anonymous, one-time interactions, economic actors have a strong preference for altruism, fairness, reciprocity, and inequity aversion.

Indeed, what many of these examples suggest is that a certain amount of controversy is productive, probably because controversy focuses attention. The most-cited *JEP* article of all time, "Toward a New Conception of the Environment-Competitiveness Relationship" by Porter and van der Linde, offered the controversial thesis that environmental regulations can "pay for themselves" by stimulating innovation that generates economic growth. This thesis clearly invites a corrective sermon from the "no free lunch" altar, and I suspect that a significant minority of the cites to this article originate from its critics rather than its admirers. Nevertheless, an article in the *American Economic Review* by Acemoglu, Aghion, Bursztyn, and Hemous (2012) makes a positive case for innovation-inducing environmental regulation, building from the directed technical change literature—a literature that the Porter and van der Linde article predates by some years.[7]

**Public School**

The examples above primarily encompass intradisciplinary debates—instances in which the profession reflects on itself. Finding these are among *JEP*'s most cited papers suggests that the first role that Stiglitz articulated for *JEP*—"providing perspective on current economic research"—is vital. *JEP* has also sought to fulfill what Stiglitz identified as its other mission, "explaining how economics provides perspective on questions of general interest." *JEP* provides a venue where economists can explain current events and the world to one another.

For example, as health and healthcare finance ascended the public agenda over the last two decades, *JEP* has helped bring the profession up to speed. Some widely cited articles include "Healthy Bodies and Thick Wallets" by Smith, "Medical Care Costs—How Much Welfare Loss?" by Newhouse, "Why Have Americans Become More Obese?" by Cutler, Glaeser, and Shapiro, and a more recent paper that is rapidly gathering citations: "Income, Health, and Well-Being around the World" by Deaton.

During the recent financial crisis, then-Editor Andrei Shleifer anticipated rapidly rising demand for professional education on the financial architecture of modern economies. In response, *JEP* commissioned three symposia (15 articles in total) that offered a three-part curriculum: "Early Stages of the Credit Crunch" (Winter 2009); "Financial Plumbing" (Winter 2010); and "Financial Regulation after the Crisis" (Winter 2011).[8] These essays surely helped many faculty members, students, and real-world practitioners get up to speed on these issues, and some of them may have a more lasting impact on the literature as well. For example, Brunnermeier's 2009 article "Deciphering the Liquidity and Credit Crunch 2007–2008" has already accrued more than 100 published citations.

[7] For a rigorous discussion of the Porter and van der Linde hypothesis in a model of directed technical change, see Acemoglu (2010).

[8] In fact, the first of these symposia was in the works well before the crisis came to a head in September 2008.

Perhaps surprisingly, some of *JEP*'s most-cited papers reside in a subject area where one might expect the almost-no-equations format of the *JEP* to be crippling: econometric methodology. The papers "Quantile Regression" by Koenker and Hallock, "Instrumental Variables and the Search for Identification" by Angrist and Krueger, and the very recent (2010) paper, "The Credibility Revolution in Empirical Economics" by Angrist and Pischke demonstrate otherwise. While we would be somewhat taken aback (perhaps even horrified) to think that practitioners are getting their econometrics training from *JEP*, we infer that at a minimum, students view the journal as something of a user's guide to current practice, and even practitioners are interested in comparing their intuition about econometric procedures with a *JEP*-style explanation.

Why would someone turn to *JEP* for guidance on a technical subject like econometrics? The reason is precisely that *JEP* privileges lucid explanation and good writing over technical exegesis. Done right, a *JEP* article does not sacrifice scientific rigor in the service of accessibility; it simply uses logic and clear language in the place of formalism to make its points. This format will not suffice for every topic in economics. But due in no small part to the singular editorial craft of Managing Editor Timothy Taylor, who has served the journal since its inception, the *JEP* has regularly amazed its readers (myself among them) with the technical depth it can reach with some well-wrought prose, a few tables, and some artful figures.

## Where Next?

In assuming the editorship of the *JEP*, I faced an intimidating question: given the journal's unique niche in our profession and its stellar track record in filling it, was there anything I could do other than carry on the journal's traditions and hope not to tarnish its reputation? Three years into my editorship, I've concluded that I won't know the answer to this question for some years after my editorship ends in 2014. Clearly, the vitality of the *JEP* depends on the originality of the articles it attracts and the quality of their exposition. Whether the articles we're publishing now are prescient contributions or merely flashes in the pan is not something I am equipped to judge in real time. Humbled by this dose of realism, I've refocused some of my energy from the sublime to the mundane by taking a few concrete steps to improve the journal in areas where progress is easier to judge.

### Originality of Contributions

As Joseph Stiglitz notes in his companion essay, the *JEP* faces the ongoing challenge of maintaining quality while maximizing the diversity of viewpoints. When successful, this model yields articles that are uniformly good and consistently eclectic. But it's easy to veer too far towards either safe choices—conventional ideas, well exposited—or toward heterodoxy for its own sake, meaning unconventional views that are not widely subscribed for good reason. One way to continually mine fresh intellectual veins is to fill the journal's editorial team with interesting, opinionated people who are not allowed to stay too long. The modal member of our editorial team serves three years, and none serves more than six.[9]

In addition, although most *JEP* articles are solicited by the editorial team, I view it as essential that the *JEP* is open to proposals from those who don't have a personal pipeline to a member of the editorial group. The *JEP* has always looked at proposals sent to our offices, but without publicly enunciating the process. Specific guidelines for proposals to *JEP* are now posted at the journal's website, with the overall goals of minimizing authors' sunk costs and maximizing transparency. We ask authors sending unsolicited work to submit short 2–3 page proposals rather than completed manuscripts. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article. For those proposals that do not clear this bar, we at least offer a quick turnaround. Approximately 20 percent of the articles currently in our publication pipeline began life as unsolicited over-the-transom proposals.

### Empirical Papers

The *JEP* is clearly not intended as an outlet for original, frontier empirical contributions—that's what refereed journals are for—but we nevertheless do occasionally publish original empirical work that seems to fit our broader intellectual mission. To clarify for ourselves and our readers what it means for an empirical paper to "fit," we again adopted some specific rules of thumb for judging empirical papers, rules that are available at the journal's website.

For a primarily empirical paper to work well in *JEP*, the paper's main topic and question must not already have found fertile soil in refereed journals. In addition, if the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. "Strong evidence" in a *JEP* paper will refer to findings that are almost immediately apparent from a scatter plot or a table of means. Although *JEP* papers occasionally include regressions, the main empirical inferences should not be dependent on functional forms or choices of control variables. Indeed, new empirical findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*. "Weak claims" in an empirical *JEP* paper means that while the empirical findings should be robust and thought provoking, the discussion should focus on the range of possible interpretations.

### Open Access

In 2010, the Executive Committee of the American Economic Association decided to make all *JEP* articles publicly accessible at no charge via the Association's website. Back issues from 1994 up through the present are online now; soon, the first seven volumes will be available, too.

[9] Tim Taylor excepted of course!

The AEA has also altered its journal distribution practices so that no member receives a paper copy of any journal without requesting it and paying for it. This practice makes sense for most journals: the readership of most AEA journals is inelastic because, in a nutshell, practicing economists need to know what in their field is published in the *American Economic Review* and the four *American Economic Journals*.

However, the *JEP* may be different. After all, it is the journal where practicing economists read about topics that aren't in their area of specialization. *JEP*'s outsized mindshare probably relies in part on the fact that intellectually curious people just can't help but read the articles once the journal is sitting on their desk, nightstand, or kitchen table.

In the hopes of maintaining the happenstance contact between the journal and its readers, we are investing in several additional methods of distribution. In 2012, *JEP* will also become available in e-reader format (for example, for Kindle or Nook) as well as issue-length PDF files. These formats will be downloadable from the *JEP* website. Those who prefer that their electronic content be automatically delivered to their devices will be able to subscribe to the *JEP* via Amazon and Barnes & Noble at nominal annual cost. Amazon and Barnes & Noble do not currently allow for no-cost subscriptions to copyrighted material, but the AEA is currently negotiating to set the *JEP*'s online subscription price to as close to free as these vendors will allow.

Finally, we will soon experiment with sending postcard mailings of *JEP*'s tables of contents to AEA members to see whether this increases *JEP* downloads relative to the regular AEA table-of-contents e-mail "blast."

## Conclusion

In reflecting on what *JEP* has accomplished in its first 25 years, I see four distinct contributions that the journal has made to our discipline: bringing research topics into the mainstream before they found purchase in refereed journals; informing, disciplining, and shaping debate on public issues; hosting and promoting intradisciplinary communication; and exposing generations of undergraduates, graduate students, and faculty members to the power, utility, and beauty of well-exposited economic insight. I hope that I have provided an inkling of *JEP*'s success on each count.

A final question I have asked myself in the last few years is how the experience of editing *JEP* has shaped my own scholarship. The answer brings me back to a theme of this essay: the scarcity of attention. Although academics are often depicted as leading lives of sober meditation akin to Rodin's *The Thinker*, my experience of academia is anything but contemplative. The rate of production of scholarship is frenetic. The number of articles I flag each day that I "should" read vastly exceeds my available waking hours. I struggle to both stay atop of my tiny corner of scholarship and also to not lose sight of the vast landscape of intellectual activity in which it's situated. This affliction is not uncommon. A *New York*

*Times* blog reported, based on analysis of the National Health Interview Survey, that economists are the fifth most sleep-deprived of U.S. occupations, only beaten out by (in order) home health aides, lawyers, police officers, and physicians and paramedics (Rampell 2012).[10]

Editing the *Journal of Economic Perspectives* has helped me to find some peace with my incurable time shortage in at least two ways. First, developing, editing, and reading *JEP* symposia furnishes me with the enforced luxury of bearing down on a new topic, absorbing some of its distilled wisdom, and developing a sense of its key open questions. As *JEP* editor—and unlike an editor of a top refereed journal—I know that I can't push the frontier of a literature by shepherding path-breaking ideas into public view. However, the *JEP* editors can offer our authors during our commenting process the benefit of the wide-eyed ignorance of a set of smart, unschooled outsiders—ultimately representing our broad intended audience of generalist economist readers. Precisely because the *JEP* editorial team is *not* expert on most of the vast set of topics in which we publish, we focus instead on asking illuminating questions, such as "Why would you assume that?" or "You call that evidence?" or "What do you say to the three obvious objections everyone makes to that claim?" This activity is much more intellectually nutritious than skimming a stack of abstracts or struggling with a couple of small ideas dressed up with a lot of intimidating math.

Editing *JEP* has also provided me with a broader perspective on the life-cycle of research. Economic research often begins with a big interesting question, which also tends to be sprawling and unmanageable. So the researcher breaks down the question into chunks, carefully examining assumptions and interpretations along the way, diving deeply into the analysis. Papers in the refereed literature result from such deep dives. But as these papers are discussed and digested, their lessons are brought back up from the deep where they can be more broadly appreciated. This process is as indispensible for scholars as it is for end users. Academics master and ultimately digest frontier scholarship by distilling its insights down to a few big facts, simple models, and reliable predictive relationships. Indeed, I have sometimes watched my scholarship condensed into single sentences and memes—and I can't honestly say that these distillations do my work gross injustice.

Seeing this process culminate at *JEP*—deep divers resurfacing—has emphasized to me that the life-cycle of scholarship should not, and does not, end with the deep dive of technical research. Continuing the process to draw robust insights—and to delineate the boundaries over which those insights apply—is one of the burdens, or privileges, of bringing an area of research to maturity. For 25 years, *JEP* has offered a unique outlet for scholars who want to do just that. Judging from the wide readership that the journal has attracted and the ongoing enthusiasm of scholars for publishing in its pages, I conclude that both producers of primary research, and the many lay and professional readers who wish to understand its contours, recognize the value of this endeavor.

[10] Not by coincidence, the analysis was funded by the mattress company Sleepy's.

## References

**Acemoglu, Daron.** 2010. "When Does Labor Scarcity Encourage Innovation." *Journal of Political Economy* 118(6): 1037–78.

**Acemoglu, Daron, Philippe Aghion, Leonardo Bursztyn, and David Hemous.** 2012. "The Environment and Directed Technical Change." *American Economic Review* 102(1): 131–66.

**Angner, Erik, and George Loewenstein.** Forthcoming. "Behavioral Economics." In *Philosophy of Economics,* vol. 13, edited by Uskali Mäki. Amsterdam: Elsevier.

**Rampell, Catherine.** 2012. "America's 10 Most Sleep-Deprived Jobs." *New York Times,* February 22. http://economix.blogs.nytimes.com/2012/02/22/americas-10-most-sleep-deprived-jobs/.

**Stiglitz, Joseph.** 1987. "Report of the Editor: *Journal of Economic Perspectives.*" *American Economic Review* 77(4): 388–89.

*Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 19–26*

# *The Journal of Economic Perspectives* and the Marketplace of Ideas: A View from the Founding

Joseph E. Stiglitz

I welcome the opportunity to join in the celebration of the twenty-fifth birthday of the *Journal of Economic Perspectives*. It is wonderful to see how this "baby," which I, along with Carl Shapiro and Timothy Taylor, nurtured through its formative years—from 1984 (three years before the first issue in 1987) until I left to serve in the Council of Economic Advisers in 1993—has grown up and become an established part of the economics profession.

Some of the little and big questions we asked at the founding have now been answered. We spent what seemed, back then, like an inordinate amount of time choosing not just the title, but the colors and design of the journal. For academics whose focus is on content and exposition, this seemed a distraction. We worried: what would it look like on the bookshelf? We wanted a modern, breezy look—but one that would also have staying power. Were our colors too reminiscent of the 1980s television hit "Miami Vice"? Would our style seem, in a few years, as old-fashioned as the covers of nineteenth-century journals—dating us and our ideas? Evidently not, for a succession of editors have chosen to keep the cover and layout essentially unchanged. The format that we established then—symposia (many focusing on policy), articles, and features—has also persisted.

So too, we were worried about whether a journal with an emphasis on good writing (as opposed to jargon) could find a place in the profession and be not only read for amusement but also widely cited. Again, the answer is "yes." While sharing some skepticism about the value of citations as a metric of impact, a quick look through Google Scholar reveals many *JEP* papers with more than 1,000 citations. (Not that we had expected to excel in that distorted measure, since we had

■ *Joseph E. Stiglitz is University Professor at Columbia University, New York City, New York. He was the founding editor of the* Journal of Economic Perspectives.

http://dx.doi.org/10.1257/jep.26.2.19. doi=10.1257/jep.26.2.19

anticipated that our articles would have more influence indirectly, through shaping *perspectives.)* We wanted something that had a greater depth of economic analysis than that provided by the general press, but was less embedded in the dictates of the models and jargon of the sub-disciplines. And I think we succeeded.

In founding the journal, we had many objectives, hopes, and ambitions. We were concerned about the increasing specialization within the economics profession. Individuals in different specialties, or sub-specialties, or sub-sub-specialties, had increasing difficulty talking to each other. They didn't share a language or a body of knowledge. Recognizing the need for a common medium of expression, we sought to have complex and sometimes arcane or highly mathematical ideas translated into plain English, or at least that dialect of the language known as "Economese"—and in a way that was not only informative but engaging. Especially because of Timothy Taylor's hard work and mastery of language, I believe the journal has succeeded in that goal.

We were worried too about a growing distance between economics and policy. We were not seeking to be a policy journal. We were deeply rooted in economic theory and empirics. But I believed that at least a portion of economic research should be related to ideas that were, or should or would be, part of the national and global policy debates. That's why we chose as our initial board of editors a diverse group, including some committed to abstract theory, but also a disproportionate number of young scholars who seemed interested in policy. (Of the initial group of 14 editors and associate editors, three became chairmen of the Council of Economic Advisers and one a member, three became chief economists of the World Bank, one became a central bank governor and another a vice-governor, two became chief economists of the IMF, and one became U.S. Secretary of Treasury—allowing for double counting of a few that held more than one position.) Our first symposium, on tax reform, reflected this commitment to policy. That we did not fully resolve all the disputes over tax policy should be obvious from the ongoing debates. But to do that was beyond our ambition.

Indeed, we began with an explicit commitment to present a diversity of viewpoints, hence the word "perspectives" in the title. As editors, we were committed to publishing ideas with which we disagreed—not that we believed they contained errors in analysis so much as we may have felt they were based on hidden or unpersuasive assumptions, or gave too much weight to some evidence, too little to others. We believed strongly in "letting a hundred flowers bloom"—a free market for ideas. Our job as editors was simply to make sure that the argument was presented as clearly as it could be. Secretly, I hoped that clarity of exposition would suffice to undermine certain of the ideas that I thought were foolish. Sometimes that was the case, but not always. When there are widely shared assumptions, it is hard to change opinions. Even dramatic events, like the Great Recession of 2008, have left many economists still wedded to macroeconomic models that seem so out of touch with what is going on.

We were lucky with timing in the founding of the journal. The American Economic Association was flush with money at that time. There was unhappiness on the part of some members of the Association with the directions in which the profession was going, as reflected in its existing journals. When I proposed the journal, it was quickly realized that it could simultaneously serve multiple functions: it would be a journal that was more accessible to more of the members of the Association, and it might help bridge some of the divides (between theory and policy, and across different sub-disciplines) in the profession.

It is not easy to manage a journal, and it has become increasingly difficult. The supply of papers has increased super-exponentially (not necessarily in tandem with the supply of new ideas), and the editors of journals like the *American Economic Review* are inundated with submissions. While the editors shape the journals—and thereby, to some extent, the profession—by their selections, much of their energy goes to separating the wheat from the chaff, just sifting through the papers and trying to make what they believe is a fair and informed decision. We wanted to take a more proactive role, for instance in identifying authors with a gift for exposition and in anticipating what might be the ideas at the center of the profession in the future.

There was another reason for our approach. One of our concerns, as noted, was to provide a closer link between policy and theory. But policy issues change quickly—much faster than the pace of the standard academic journal. We wanted a structure that enabled us to target policy issues while they were still the subject of debate. For instance, we entered the fray on the economics of transition quickly after the collapse of the Berlin Wall with a Fall 1991 symposium. While our focus was on the economics profession itself (in shaping, for instance, the research agenda), the fact that we made our articles accessible meant that they were able to have a greater impact on policymakers and play a more important role in public discourse.

This goal forced us to make an unpleasant decision: we would solicit papers. I say unpleasant, because I would have preferred more openness. Indeed, the allegedly more open journals were already being accused of being closed, giving preferential treatment to friends and colleagues. That was why we put such stress on having a diverse group of associate editors, giving them responsibility for soliciting articles, and on our commitment to publish articles with which we disagreed. Still, we worried about accusations of elitism, and not surprisingly, charges were levied. Of the 11 articles in the first issue, somewhat less than 50 percent were from the elite universities (say the top ten), while a quick perusal of the most recent issues suggests the percentage has crept up to something more like two-thirds.

Making the journal one where the editors solicited papers gave us time to think more deeply about what it was we should publish and to work harder to make sure that the papers did what we wanted them to do and were accessible. Our approach ensured that the average paper quality was higher and that the papers came out more quickly. I don't think any economics journal had such intense editorial intervention—we rewrote large fractions of many of the papers, and reshaped many more. In the end, I think the approach worked well, enabling the journal to become (by many accounts) the most successful new entrant in the world of academic journals in decades.

But at the same time, this approach put a special burden on us—one that I hoped our successors as editors would take on board. We had to take care to make sure that the journal didn't just reflect our own perspectives, our own views about where the economics profession should be directing its attentions, our own judgments of what was important and what was not. Thus, even more important than a diversity of backgrounds is a diversity of perspectives. This, I thought, was especially important in a field known for having certain orthodoxies—orthodoxies that dominate for a while and then fade, making the profession sometimes look less like a science than it would pretend to be. A case in point is the well-known and widely documented belief within the profession as the economy entered the Great Depression that markets were self-correcting and government intervention would be a mistake. Another is the monetarist fad a half-century later.

We worked hard to challenge these orthodoxies, with some success—and to challenge them *before* it became the fad to do so. In the first issue of Summer 1987, for instance, we published Gavin Wright's wonderful piece on "The Economic Revolution in the American South," where he argues that imposing national wage norms and labor standards on the South played a pivotal role in its transformation out of backwardness. The recognition that minimum wages might not have the adverse effects that economists had widely presumed was, of course, given further impetus in later work by Alan Krueger (editor of this journal from 1996 to 2002) and David Card.

The rational choice model had long been the basis of standard economic analysis. My own work on the economics of information had shown that many of the results of the standard model were not valid *even with rational expectations,* so long as there were, say, important information asymmetries. My own work had uncovered a large number of phenomena that were hard to reconcile with the rational choice model plus rational expectations and even information asymmetries. From its inaugural Summer 1987 issue, the journal explored both these anomalies and the problems posed for the underlying theoretical structures (the former in a special section we had regularly on "Anomalies," edited and often written by Richard Thaler, and the latter in an article by Mark Machina titled "Choice under Uncertainty: Problems Solved and Unsolved").

One of the goals we set out for ourselves in the foreword published in the first issue of the journal was to disseminate developments within economics more rapidly. This posed a challenge: separating what might be a short-run fad (an intellectual bubble) from what might be a transformative idea. In retrospect, I think we did reasonably well. The critique of standard theory that we advanced in our anomalies section evolved into an important strand in modern economics: behavioral economics. We would like to think that perhaps the attention we gave it played a role in its quick rise in prominence.

Other ideas would take longer: we featured a symposium on bubbles in the Spring 1990 issue. If only policymakers had paid more attention, instead of claiming (based on "rational and efficient markets") that bubbles can't occur! So too, in our symposium on the economies in transition, in the Fall 1991 issue, Peter Murrell provided perspectives that I wished our policymakers had listened to more closely. He questioned basing that transition on "neoclassical economics," anticipating the problems that would be posed by shock therapy—the enormous decline in GDP experienced by the transition countries following the standard model—when moving toward market economy was supposed to bring unprecedented prosperity. His paper also anticipated the long-term problems in establishing a rule of law in Russia, problems that I believe were worsened by the misguided policies of the first decade of transition.

Staying ahead of the game—anticipating where the meanderings of the economics profession will lead—is no easy task. Who would have anticipated, for instance, that the conventional wisdom on capital controls would change so quickly and dramatically, with the IMF now supporting their imposition, at least under certain circumstances (and with a much broader set) than was the case even a short while ago? But as far back as a dozen years ago, there was a large body of nonmainstream thought and research—much of it solidly grounded in theory and empirics (not necessarily the "standard theory" with infinitely-lived individuals with complete risk markets and no information imperfections)—pointing out that capital controls sometimes make sense. It is important that the journal be a "big tent," even when the editors subscribe (or not) to the effervescent conventional wisdom.

By the same token, the relationship between inequality and fluctuations is, once again, becoming the subject of attention. Obviously, representative agent models in which distribution plays no role are unable to shed light on this issue, but there have been alternative macroeconomic traditions that have stressed these issues. This is a big deal. I played a role in organizing a UN Commission of Experts on Reforms of the International Monetary and Financial System, which called attention to this and other issues in *The Stiglitz Report: Reforming the International Monetary and Financial Systems in the Wake of the Global Crisis* (New Press 2010). The *Journal of Economic Perspectives* should be calling attention to this kind of issue, one hopes before it becomes painfully obvious to everyone else.

We never shied away from controversy at the journal, but we did try to ensure that the discussion was balanced. In the tax symposium in the first issue of Summer 1987, we included voices from Buchanan to Musgrave and Pechman, with important contributions from then-emerging young scholars such as Auerbach, Rubinfeld, Courant, and Poterba. By the time that symposium was commissioned, it was already clear that the supply-side advocates of Reagan's tax changes had been proven wrong. But we wanted to dig deeper, to look more closely at what had actually happened, for instance, to labor supply or savings. And we wanted to see what, in retrospect, the advocates of those reforms had to say. In rereading this symposium (as in the case of so many of the others in those early years), it is striking how relevant the issues addressed are to what is happening today; and I can't help but feel that current policy debates might be improved if more attention were paid to these earlier analyses.

We knew that it would be hard to maintain that balance, and there are reasons for concern in this area. In the Fall 2010 symposium on "Macroeconomics after the

Financial Crisis," insufficient attention was paid to the voices of macroeconomists critical of the reigning paradigm *before* the crisis (say in the tradition of Hyman Minsky, or those who had explored debt-deflation models, financial instability models, and models with greater emphasis on credit and credit interlinkages). In the journal's Winter 1993 symposium titled "Keynesian Economics Today," attention had been given to some of these alternative perspectives, including the debt-deflation approach—where more wage and price flexibility could exacerbate unemployment (a paper I wrote with Bruce Greenwald)—and an approach focusing on wage and price rigidities (a paper by David Romer).

In the Winter 2011 symposium on "Financial Regulation after the Crisis," again the voices of those who had called for more and better regulation *before the crisis* were either absent or underrepresented. As another example: only one article in that symposium is devoted to consumer protection, arguably one of the most significant parts of the Dodd–Frank financial reform bill. While that article firmly sets the need for consumer protection within a behavioral economics context, it gives short shrift to the predatory and (sometimes borderline) fraudulent and deceptive practices that have marked the financial sector. While it is, perhaps, unfair to pick on the lacuna in any single article, perspectives on whether markets work or not, and if so, why they fail, are reflected in the attention paid to one problem versus another, and the arguments *not* noted. Does the reluctance of banks to restructure arise out of the difficulties of identifying strategic defaulters (an issue noted in the paper)? Or from the fact that, given deficiencies in accounting standards, with restructuring, losses have to be recognized, with consequent implications for capital adequacy? Or because of conflicts of interest between holders of first and second mortgages and service providers (another issue not mentioned)? The fact that there were such deficiencies in risk analysis on the part of so-called experts raises questions about the ability of the financially unsophisticated to manage risk, so while the paper's discussions of financial literacy were interesting and welcome, the problems go beyond "literacy." At the end, the article devotes a couple of pages to the dangers of regulatory interventions.

It might have been good to have an article from a law and economics scholar—perhaps even Elizabeth Warren, viewed as the initiator of the idea, someone aware of the dangers of regulation, but even more aware of the dangers of underregulation. The perspectives of non-American regulators, who have taken somewhat different views than their American counterparts, might be informative: for example, in the United Kingdom alone, there is Adair Turner, former head of their Financial Service Authority regulatory agency, who has clearly articulated the view that the losses from underregulation outweigh, by orders of magnitude, the costs of regulation; Mervyn King, head of the Bank of England, who has warned strongly of the dangers of too-big-to-fail banks; and Andy Haldane, Executive Director for Financial Stability at the Bank of England, who has detailed the risks of financial systems that are too intertwined.

When we initiated the journal, we did not necessarily expect any single article to be balanced—indeed, the notion of the *Journal of Economic Perspectives* was that different articles would take different perspectives—but that the *collection* of articles would be balanced. We wanted articles with a viewpoint, and we wanted a clash of viewpoints. We wanted microeconomic perspectives to be balanced with macroeconomic perspectives—for example, some kinds of mortgage systems may not only be less exploitive of uninformed consumers, but may have performed better systemically, across countries, and over time, perhaps partly because they are less exploitive.

The explosion of research in economics, the imperial successes of economics in making inroads into neighboring disciplines, and the increasing awareness of the limitations of the standard economics paradigm makes the importance of the journal even greater today than it was a quarter–century ago. Today, there is interesting work going on, for instance, in economic anthropology—work, for instance, viewing central bankers through the eyes of a cultural anthropologist—or on the border of sociology and economics, exploring how social constructions like race and caste arise and affect behavior. Given the scarcity of time, most economists simply can't explore this exciting terrain, and the *Journal of Economic Perspectives*, breaking down the boundaries not only within but across disciplines, can help make these ideas acceptable.

Economists often think of their task as making markets work, or at least work better. The *Journal of Economic Perspectives* has helped make the marketplace of ideas work much better. It is not an easy task. It is a task which is becoming increasingly difficult. The journal is to be congratulated on the enormous success it has attained.

*Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 27–40*

# From the Desk of the Managing Editor

Timothy Taylor

Editing isn't "teaching" and it isn't "research," so in the holy trinity of academic responsibilities it is apparently bunched with faculty committees, student advising, and talks to the local Kiwanis club as part of "service." Yet for many economists, editing seems to loom larger in their professional lives. After all, EconLit indexes more than 750 academic journals of economics, which require an ever-shifting group of editors, co-editors, and advisory boards to function. Roughly one-third of the books in the annotated listings at the back of each issue of the *Journal of Economic Literature* are edited volumes.

Editors are gatekeepers, and editors are road-blocks—or perhaps these are essentially the same task. Editors shape "the literature," both what and who is included and how it is presented. I've come to believe that "editing" is no more susceptible to a compact single definition than "manufacturing" or "services." But here is one take on the enterprise of editing from someone who has been sitting in the Managing Editor's chair for all 100 issues of the *Journal of Economic Perspectives* since before the first issue of the journal mailed in Summer 1987.

## Early Developments

I was hired by Joseph Stiglitz back in 1986 to be the Managing Editor of an indeterminate new journal of academic economics. At that time, the name of the

■ *Timothy Taylor is Managing Editor of the* Journal of Economics Perspectives, *based at Macalester College, St. Paul, Minnesota. He blogs at ⟨http://conversableeconomist.blogspot .com⟩. His e-mail address is ⟨taylort@macalester.edu⟩.*

http://dx.doi.org/10.1257/jep.26.2.27. doi=10.1257/jep.26.2.27

journal had not yet been chosen, nor had the layout, cover design, weight of the paper, font, kerning, or formatting.

But one key element of the new journal—indeed, its main justification—was predetermined. The journal would not seek to publish the latest research articles in all their pyrotechnical glory. Such articles would continue to be the purview of the *American Economic Review* and other blind-refereed research journals. Instead, the new journal would seek to publish articles that would be broadly readable by all economists—including, particularly, those who do not specialize in the subject of the article.

To produce articles exposited in the desired style, the editors of the new journal would solicit them, and then rely on in-house reviewing. In particular, the editorial process of this new journal would rely on comments from the professors who would serve as editor and coeditor, but also on an in-house Managing Editor who would perform detailed and hands-on editing of all the articles. We would take advantage of the newfangled "word processing" technology, and mail floppy disks with the articles and revisions back and forth! We would even send floppy disks with electronic files of papers directly to the typesetter! This seemed like hot stuff in 1986.

The job of Managing Editor at *JEP* is, unsurprisingly, part managing and part editing. The management portion involves overseeing the editorial process to assure that we are inviting a steady stream of authors and following up on deadlines. It means assuring that correspondence reaching the office receives an answer. It means, with the help of an assistant (at the start of the journal, Caroline Moseley, now Ann Norman), dealing with the typesetter and printer, annual budgets, office equipment, and the like. I try to be diligent at this part of my job description, but frankly, these tasks are the vegetables I need to eat before dessert—which in this case is the actual editing of manuscripts.

The editing portion of this job called for someone who was tooled up on their technical economics but who had a demonstrated ability to write for a broader audience. On a Venn diagram, the intersection of people with those skill sets isn't large. Joe Stiglitz was spending a chunk of his summer at Stanford in 1985 at what was then called the Institute for Mathematical Studies in the Social Sciences (since 1989, it has been the Stanford Institute for Theoretical Economics), and he asked around the faculty to see if they knew any likely candidates. I had spent a couple of years in Stanford's Ph.D. program, before dropping out with a master's degree, and I had been working for a couple of years at the *San Jose Mercury News* as an editorial writer and columnist. Some of my old faculty connections recommended me to Joe.[1] I truly enjoyed being a newspaper opinion writer, but the chance to get in on the ground floor of the new journal seemed an opportunity not to be missed.

[1] For more on my background, my reactions to graduate school in economics, my early days at the *Journal of Economic Perspectives*, and the task of editing, see my essay in the Summer 2009 issue of the *American Economist* (Taylor 2009).

The decision to have a journal primarily consisting of solicited articles has been a long-simmering source of controversy. To us, "solicited" meant that the editors would come up with ideas, approach potential authors, and make decisions without a standard double-blind refereeing process. After all, the kind of papers we wanted for *JEP* didn't have other obvious publication outlets, and we feared that if the journal encouraged open submission of papers, many authors would spend time and effort writing something to fit the *JEP* style—and then not have other outlets for their paper if we turned them down. By pre-committing to authors, we encouraged authors to put in the time for a *JEP* draft, and discouraged others from gambling their time in an effort that, given our page limits, was unlikely to end in publication.

But to some, "solicited" meant "deliberately intended to shut me out personally." From the start, the *JEP* has been open to suggestions for potential papers from anyone, whether in the form of general ideas or already-written papers. But we publish only about 35–40 papers per year, and given that the editors are reaching out to potential authors for most of those papers, there just isn't room for much else. Over time, I'd guess that maybe one or two papers per issue originally arrive "over the transom," as we used to say back in the twentieth century.

Any form of selection implies a possibility of error or bias. However, with more than 750 refereed journals of economics in existence, having one journal of primarily solicited articles hardly seems like a severe restriction on publication options. Indeed, I've noticed that many of those who complain about the existence of a solicited journal also fervently believe that standard refereed journals are unfair to their work, which suggests that their issue is not actually with how a solicited journal works, but reflects a broader issue over the openness of academic journals to certain kinds of papers.

A final concern about a solicited journal is that when the requested article arrives, its quality may be dismal. The *JEP* approach is that once we have solicited a paper, we have made a commitment to work with the author to produce a paper of the sort we would like to publish. If solicited authors don't want to work with us to improve a paper, as happens once every few years, we are not obligated to publish it. Brad DeLong once memorably observed after some years of participating in the *JEP* editorial process as a co-editor: "We can't always make a silk purse out of a sow's ear, but we can usually make a rayon handbag."

I am not an evangelist for the concept of solicited journals. I believe that the overwhelming majority of academic journals should be blind-refereed. But a solicited journal does have certain advantages. It can reach out to authors who otherwise would be unlikely to submit a paper; sometimes those authors are well-known, or they may be more obscure choices. It can elicit papers or groups of papers on desired topics.

A solicited journal with an activist in-house editor can also speed up the process of convergence from first draft to final version. Ellison (2002) documents that the time from submission to acceptance at economics journals has risen from 6–9 months 30 years ago to an average of about two years today. The main cause is not that referees are slower, but rather that multiple rounds of more extensive revisions are now more commonly required. With a solicited journal, the topic and general approach of the paper has been agreed upon beforehand. Our editorial process offers sets of comments as well my own first stab at actually making many of the requested changes. If we want a paper shortened or a passage clarified, I suggest in my edited draft how this might be done. The author often revises my edited draft substantially, and sometimes we proceed through more than one iteration. But with a hands-on editor, the process of convergence to a final draft can often be more direct and rapid.

## The Special Challenge of Editing Economics

Two of the more admirable traits of economic discourse also make *JEP* articles a genuine challenge to edit. First, economists have long shown a willingness to reach outside their field and learn what seems useful: that is, economists dig down not just into economics itself, but also into math and statistics, history, game theory, sociology, psychology, business and management, philosophy, demography, epidemiology, evolutionary biology, genetics, brain science, and other areas. Second, economics earns the moniker of being a "discipline" by its unrelenting insistence on spelling out underlying assumptions and intermediate steps, and then testing those links against alternative approaches and evidence. For the uninitiated, this insistence on spelling out every step in an argument—along with many of the possible alternatives— may seem pedantic at best and anal-retentive at worst. But economists recognize how this approach aids in developing a nuanced understanding of the exact conditions under which results hold. In addition, it helps to focus research and discussion on precise issues under dispute—rather than just spit-balling back and forth over opposing conclusions. Thus, a well-constructed economic argument often has both an interdisciplinary richness and a logical tautness.

Some economists like to believe (although this belief has blessedly faded in recent decades) that economics is an edifice built on the rocks of mathematical theory and statistical empiricism, and everything else is superfluous fluff. McCloskey (1983) thoroughly strafed that conceit, pointing out in "The Rhetoric of Economics" that the research and analysis of economists is built on uncertain and subjective judgments, and often uses, among its rhetorical tools, analogy and metaphor, appeals to authority and to commonsense intuition, and the use of "toy models" counterbalanced with the choice of supposedly illustrative real-world episodes. Economic arguments rooted purely in mathematical formalism or statistical analyses are superb at specifying the steps leading to the particular conclusion. However, cynical economists (but I repeat myself) know that a model can be built to illustrate any desired conclusion, and that if the data are tortured for long enough, they will confess to anything. Persuasiveness requires a multidimensional argument that reaches beyond formalism. As McCloskey (1983) wrote: "There is no good reason to wish to make 'scientific' as opposed to plausible statements."

Indeed, one of the original justifications for *JEP* was a sense that its articles could offer broader access to the kinds of conversations that surround formal or technical economic results. Our comments and editing often push an author to enunciate the strengths and limitations of an argument, and to respond to questions that we suspect will be asked.

For my hands-on editing of *JEP* articles, the wide-ranging rhetorical tools of economic argument pose a substantial challenge. I try to read widely, but I can't possibly keep up with the field of economics alone, much less with all the other areas with which economists interact. Instead, I have long viewed myself as the designated dummy of the *JEP* operation: that is, a big part of my job is to be willing to raise a hand and say: "I don't understand." If I can't make sense of the links in the argument, or if I'm not confident that others can make sense of the argument, then it's not going into the journal.

Of course, it's not productive for me to be too dumb. Before confessing my ignorance in my comments, I make a genuine effort to understand each paper on its own terms. With every *JEP* paper, I read it once or twice, set it aside, and come back to it a few days later before working through it. When necessary, I'll take my questions to the papers cited in the reference list, to textbooks, and to colleagues, before going back to the author. The Internet has been a godsend for the type of editing that I do: I used to take an afternoon or so each week to haunt the stacks of libraries, hunting through background materials. Now most of what I need is online and searchable. Editing in this spirit benefits from a grasshopper mind: a willingness to hop headlong into new subjects once or twice every week. I then revise and edit, line by line, in a way that seeks to clear up gray areas and can serve as an intermediate draft on which the author can build with additional revisions.

## Sweating the Small Stuff

In Walter Salant's (1969) classic article on "Writing and Reading in Economics," he begins (p. 545) with a passage that could serve as part of my job description: "In the past several months I have spent much time reading manuscripts written by my professional colleagues. Although this activity has taught me some economics, as one might expect, it has not been an unmixed pleasure. At some times, to be frank, it has been rather trying. What has made it trying is that too much of the writing I have read is clumsy or worse: nearly incomprehensible. Crimes of violence are committed daily against the English language and the helpless reader is too often frustrated in his effort to understand the message."

There is surely an essay to be written on the nuts-and-bolts of editing expository academic prose, although the details would probably be of more interest to teachers of English composition than to professional economists. Here, I will just offer a few items of advice for economists, based on some of the more common problems I see.

*1) Develop self-awareness about the technical level of your exposition.* A wise colleague once said to me that jargon and technical expression always have two uses: 1) to streamline and clarify the communication of concepts between specialists, and 2) for specialists to identify themselves as an in-group, while impressing and excluding outsiders. All of us are prone to the belief that we are only using jargon from the highest motives, to pursue intellectual clarity, while quietly feeling the inner glow of being within the charmed circle of jargon-users. But communication with others means tailoring the jargon to the audience, whether in a seminar room, classroom, or faculty meeting, speaking with a reporter, or at a Kiwanis club lunch talk.

*JEP* is a journal for nonspecialist economists, which means that we presume our readers understand the basic jargon of undergraduate economics. We find no need to define terms like "deadweight loss" or "ordinary least squares" or "comparative advantage." However, we do ask authors to explain the more specific terminology of their field, and to do so mostly in *words* rather than algebra. A few times a year, we receive a solicited first draft from an author which, despite all our admonitions, is a blizzard of display equations. I feel that for some authors it is nearly a physical effort to unclench from their jargon.

*2) Shorter is better.* A standard *JEP* article is about 6,000–7,500 words of text, although our articles are often 1,000 words or so above or below this limit when it seems appropriate. Still, we often receive articles that are well above the target length. It's fairly standard for me to cut *JEP* drafts by 1,000–3,000 words. But some examples are more extreme: in the last few years, I have several times trimmed first drafts that exceeded 80 pages to under 30 pages—and in some cases to less than 20 pages.

While spending a few days leaning on the "delete" key, I find solace in believing that shorter articles are more widely read. I always smile at the story in E. B. White's introduction to the *Elements of Style* (1979), in which he remembers his old teacher William Strunk striding up and down in front of the class repeating: "Omit needless words! Omit needless words!" Repeating the admonition does not violate the rule—because for almost all writers and essays, the marginal effect of the final repetition remains significant and positive. One of my messages to authors is: You are not Charles Dickens and you are not getting paid by the word (or at all, in fact!), so get on with it already. In what is said to be the shortest book review on record, the American humorist Ambrose Bierce once wrote, "The covers of this book are too far apart" (as quoted in Aristides, 1996, p. 167). The distance between the covers of *JEP* is predetermined by a budget constraint, but whether a given issue contains six articles or a dozen is largely up to me. I believe that shorter articles are more widely sampled and read, so authors who want to disseminate the central themes of their work have an interest in tighter presentation.

*3) Invest more time in the stepping-stones of exposition: introductions, opening paragraphs of sections, and conclusions.* Introductions of papers are worth four times as much effort as they usually receive. The opening paragraph of each main section of a paper is worth three times as much effort as it usually receives. Conclusions are worth twice as much effort as they usually receive. This recommendation emphatically does not call for long introductions with a blow-by-blow overview of each subsection of the paper to come. It doesn't mean repeating the same topic sentences over and over again, in introduction and section headings and conclusion. It means making a genuine effort to attract the attention of the reader and let the reader know what is at stake up front, to signpost the argument as it develops, and to tell the reader the state of the argument at the end.

*4) Program yourself to recognize vague and fluffy phrases.* As a starting point, think twice, and then again, before ever starting a sentence with "There is . . .", "That is . . ." or "It is . . ." If you have forgotten what the "passive voice" is, or why it saps your writing of vigor and force, time for a quick review of Rule #14 of Strunk and White's (1979) *Elements of Style.*

*5) Old fusty references and quotations should be taken out and shot.* If you ever feel a desire to quote Lewis Carroll or Arthur Conan Doyle, stab yourself in the hand with a pencil until the desire passes. No economics paper should ever again quote "a word means precisely what I want it to" or "when you have eliminated the impossible, whatever remains, however improbable, must be the truth." Unless you are writing for freshmen and sophomores, it is risible to tell your readers that "economics is the study of choice" or that "in the long run, we're all dead." Never again assume a can-opener, or watch sausage being made, or look for your keys under the streetlamp. Remove all references to Thomas Kuhn's paradigm shifts or Karl Popper's positivism. Conversely, when you hear a funny comment in a seminar, or read a good line in an article, or a quirky association crosses your mind, hoard the rhetorical nugget for future use.

Most editing along these lines is an endless parade of small-time changes: adding an extra sentence of explanation here, tightening two sentences over there, taking five paragraphs of section-by-section overview in an introduction and shrinking it to five sentences, spotting a nice example buried in the middle of page 9 that could work well as an introduction, rearranging sentences to avoid the passive voice, reconnecting split infinitives, and slashing all references to *Alice in Wonderland* and Sherlock Holmes. None of these changes is individually crucial. But I'm professionally committed to the belief that the accumulation of such changes makes a substantial difference. Readers have budget constraints of time and attention. One of the main justifications for intensive editing is that readers can consume a fixed quantity of intellectual content with deeper understanding in less time.

## The Editor as Intermediary

Communication is hard. The connection between writer and reader is always tenuous. No article worth the reading will ever be a stroll down the promenade on a summer's day. But most readers of academic articles are walking through swampy woods on a dark night, squelching through puddles and tripping over sticks, banging their shins into rocks, and struggling to see in dim light as thorny branches rake at their clothing. An editor can make the journey easier, so the reader need not dissipate time and attention overcoming unnecessary obstacles, but instead can focus on the intended pathway.

Obstacles to understanding arise both in the form of content and argument and also in the nuts and bolts of writing. An editor needs a certain level of obsessiveness in confronting these issues, manuscript after manuscript, for the 1,000 pages that *JEP* publishes each year. Plotnick (1982, p. 1) writes in *The Elements of Editing*: "What kind of person makes a good editor? When hiring new staff, I look for such useful attributes as genius, charisma, adaptability, and disdain for high wages. I also look for signs of a neurotic trait called compulsiveness, which in one form is indispensable to editors, and in another, disabling."

The ultimate goal of editing is to strengthen the connection between authors and readers. Barney Kilgore, who was editor of the *Wall Street Journal* during its time its circulation expanded dramatically in the 1950s and 1960s, used to post a motto in his office that would terrify any editor (as quoted in Crovitz 2009): "The easiest thing in the world for a reader to do is to stop reading." An editor can help here, by serving as a proxy for future readers.

For many *JEP* authors, my extremely hands-on editing comes as a surprise. After all, most of those who are listed as editors of journals or conference volumes don't actually "edit" in the commonplace meaning of the term. McCloskey (1985, p. 188) made this point in the mini-classic article on "Economical Writing": "Occasionally an editor will receive and pass along to the author a report by a referee that criticizes style in more detail than 'I found this difficult to read' or 'There's a typo on p. 6.' The editors themselves do not edit."

Editors for a specialized academic journal, for example, focus more on choosing among submissions for those of higher quality, rather than trying to boost the quality of accepted articles with detailed revisions. In an essay about his experiences as a co-editor of the *American Economic Review* and editor of *Economic Inquiry*, Preston McAfee (2010, p. 2) described the task this way: "Not all authors agree, of course, but in my view, we are in the business of evaluating papers, not improving papers. If you want to improve your paper, ask your colleagues for advice. When you know what you want to say and how to say it, submit it to a journal."

Similarly, many editors of a conference volume know that they face a situation in which A) the papers are of middling but acceptable quality; B) the papers will have a small audience limited mainly to other specialists in the field; and C) the length of the editorial process determines whether the volume will be published next year or five years from now. They quite reasonably focus on making sure that the papers are complete and reasonably clear, while not sweating the details, and then push them into print sooner rather than later.

But the *Journal of Economic Perspectives* is intended to be a fundamentally different kind of journal. Most journals and conference volumes are essentially large file drawers, organizing content so that you can find it again and certifying that the content is professionally approved. The idea behind *JEP* was for articles to achieve a standard of exposition that is "readable," as opposed to "decipherable given sufficient investment of time and energy." I carry with me mental archetypes of possible *JEP* readers: for example, the 55 year-old professor who is now primarily teaching, rather than doing research; the senior undergraduate or first-year graduate student trying to get traction on a possible research topic; or the sophisticated financial journalist or Congressional staffer who wants to get up to speed on a topic. More broadly, everyone is a nonspecialist in most areas—and thus we hope that all economists will find articles of interest in every issue of *JEP*.

This broader audience drives me to do editing that can be highly interventionist; this is why I do detailed hands-on editing and revising for (nearly) every paper that has ever appeared in *JEP*. Peter Drucker (as quoted by Jenkins 2007) wrote in his memoirs about newspaper editors: "Every first-rate editor I have ever heard of reads, edits and rewrites every word that goes into his publication. . . . Good editors are not 'permissive'; they do not let their colleagues do 'their thing'; they make sure that everybody does the 'paper's thing.'" In that spirit, most of my days are spent working through drafts, tweaking, revising, and reorganizing, one sentence, paragraph, and section at a time. I find the work always intriguing, exhausting, and often downright fun.

## Are Economists Willing to be Edited?

When I took the job of editing back in 1986, one of my main worries was that because the journal was chartered on a two-year trial basis, I might be out of a job by 1988. Another worry was that my job would be a series of pitched battles with authors. After all, McCloskey (1985) had warned in the opening paragraphs of "Economical Writing" (p. 187):

> Most people who write a lot, as do economists, have an amateurish attitude toward writing. Economists do not mind criticism of their facts or their formalisms, because they have been trained in these to take criticism, and to dish it out. Style in writing is another matter entirely. They regard criticism of their drafts the way a man unfamiliar with ideas regards criticism of his ideas: as an assault. . . . The economic writer, therefore, cherishes his habits of style as matters God-given, or at the least highly personal. One cannot change one's body-type or basic character, and it is offensive for some creep to criticize them . . .

The sensitivity of authors to being edited has some rational basis. Over the years, I've also heard plenty of anecdotes about editors who viewed authors as little more than a necessary evil. Humorist Andrew Ferguson (2007) once described the editor's "Platonic ideal of the perfect contributor—the writer who hands in his article and is then run over by a bus before he can complain about the editing . . ."

When you deal with academics, you aren't necessarily working with a group of great team players. However, my experience in editing economists over the last quarter century has been overwhelmingly positive, and certainly has not borne out fears of an inevitable clash between authors and editors. With remarkably few exceptions, *JEP* authors have been receptive, respectful, and even on occasion grateful for our comments and my detailed editing. They often say that it was a pleasure to receive constructive feedback—which I interpret as a not-so-veiled comment on the editorial feedback they are accustomed to receiving.

I do try to follow the "spoonful of sugar" approach to editing. When I can praise parts of an article, I do. I don't pull any punches about what I think needs to happen—but how you ask can be important. Alexander Gerschenkron once wrote a note to Abram Bergson, asking for comments on a paper, and wrote (as described in Dawidoff, 2002, p. 142): "Let me have your criticism, general and particular, and let me have it promptly"; a postscript added, "Criticisms are to be submitted in the form 'I suggest the following change' never in the form: 'This does not make sense' or similar."

However, I suspect that the nature of *JEP* as a journal also contributes to the perhaps surprising cheerfulness of *JEP* authors after they look over pages of comments and see how I've chopped and rearranged their drafts. After all, they know that we solicited their paper and are committed ultimately to publishing. Moreover, I'm not a fellow specialist challenging them on their own intellectual turf, but someone trying to bring their work to a broader audience. Our editing process is not intended to be adversarial, but to help the author's light to shine more broadly and brightly.

In a long-ago essay on editing, James Thurber (1959) wrote: "Editing should be, especially in the case of old writers, a counseling rather than a collaborating task. The tendency of the writer-editor to collaborate is natural, but he should say to himself, 'How can I help this writer to say it better in his own style?' and avoid 'How can I show him how I would write it, if it were my piece?'" Editing in Thurber's spirit requires checking my ego at the door. I often remind myself that the goal isn't a plain-vanilla style, nor my personal style, but an improved version of the author's own style. Not everything needs changing. Concerns about exposition matter, but my opinions about the content don't.

When the editorial process wobbles, I've been very fortunate in those who have served as Editors of *JEP*: Joseph Stiglitz, Carl Shapiro, Alan Krueger, Andrei Shleifer, and now David Autor. I have said to every Editor that I can do most of the work on exposition, but every now and then, I will need them to back me up and play the hammer with a few recalcitrant authors. When necessary, they have each stepped up and done so.

## Editing and the Production of Knowledge

Every professor has had a student ask: "Are you going to grade us on our writing or on our ideas?" Of course, a fair translation of this question might be: "Can we get full credit on your exams if we give unclear answers?" Every few years I tack up on my door this comment from the sixteenth-century French philosopher Montaigne (1579–1580, as translated by Frame, 1971, p. 125): "I hear some making excuses for not being able to express themselves, and pretending to have their heads full of many fine things, but to be unable to bring them out for lack of eloquence. That is all bluff. Do you know what I think these things are? They are shadows that come to them of some shapeless conceptions, which they cannot untangle and clear up within, and consequently cannot set forth without: they do not understand themselves yet."

The underlying lesson here is that knowledge and the exposition of that knowledge are not fully separable, and that lesson applies to faculty as well as to students. Lack of clarity usually reflects a less-than-full understanding. As the professors of communications say: "Rhetoric is epistemic." That is, the way in which you express yourself is actually part of the knowledge-content of what you say.

Perhaps I'm a little overemphatic or defensive on this theme, but it touches near the heart of what a journal like *JEP* can hope to accomplish. One vision of "knowledge" is that it all appears in the specialized literature, and dissemination of that knowledge—whether through *JEP*, policy reports, articles in the popular press, teaching, or textbooks—cannot add to knowledge. However, I believe that knowledge is multidimensional: for example, as ideas and applications are applied and considered and explained in various contexts, new strengths and weaknesses are ever-emerging. One of the nicest compliments our *JEP* editorial process ever received was from a prominent author who sent us a first draft that, by *JEP* standards, was overly technical. We pushed him to scale back the algebra and to explain in words. With his revision, he sent along a note saying as he had worked to explain the material in a way appropriate for *JEP*, he had also come to a better understanding of his original technical demonstration.

Knowledge doesn't end with the QED at the end of a proof or with the publication of a regression table. An editing process that produces an accessible discussion of results is part of knowledge, too.

## Permanence and Serendipity in Web-World

Some days, working on an academic journal feels like being among the last of the telephone switchboard operators or the gas lamplighters. Printing on paper is a 500 year-old technology. When the first issue of *JEP* was printed in 1987, the print run was nearly 25,000 copies. Now, as readers shift to reading online or on CD-ROM, the print run has fallen to 13,000. The American Economic Association has shifted its membership rules toward a model where all dues-paying members have online access to the AEA journals at zero marginal cost, but need to pay extra for paper copies. Thus, in the next few years, I wouldn't be surprised to see the *JEP* print run fall by half again. The smaller print run means substantial up-front cost savings for the AEA: paper and postage used to amount to half the journal's budget. But for anyone sitting in a managing editor's chair, the shorter print runs also raise existential questions about your work: in particular, questions about permanence and serendipity.

Back in 1986, when we were choosing paper stocks for the journal, "permanence" meant acid-free paper that would last 100 years or more on a library shelf. I'm still acculturating myself to the concept that in the web-world, permanence has little to do with paper quality, but instead means a permanent IP address and a server with multiple back-ups. As a twentieth-century guy, pixels seem impermanent to me. I still get a little shock seeing a CD-ROM with back issues of *JEP*: almost two decades of my work product condensed down to a space about the size of a lettuce leaf.

But in a world of evanescent interactive social media, there remains a place for publications that are meant to lay down a record—to last. It pleases me enormously that the American Economic Association in 2010 made all issues of *JEP* freely available online at ⟨http://e-jep.org⟩. Archives are available back to 1994; the complete journal back to 1987 will eventually become available. The *JEP* now has a combination of permanency and omnipresence.

The other concern about the gradual disappearance of paper journals is the issue of serendipity—the possibility of *accidentally* finding something of interest. In the old days, serendipity often happened when you were standing in the library stacks, looking up a book or paging through a back issue of a journal, and you ran across another intriguing article. The *Journal of Economic Perspectives* was founded on the brave and nonobvious assumption that busy-bee academic economists are actually interested in cross-pollination—in reaching out beyond their specialties.

As the *JEP* makes a gradual transition from paper to pixels, I hope it doesn't become a disconnected collection of permanent URLs. When you hold a paper copy of an issue in your hands, the barriers to flipping through a few articles are low. When you receive an e-mail with the table of contents for an issue, the barriers to sampling are a little higher. But perhaps my worries here betray a lack of imagination for where technology is headed. Soon enough, I expect many of us will have full issues of our periodicals delivered directly to our e-readers. When these are tied together with the connectivity of weblinks and blogs, the possibilities for serendipity could easily improve. Starting with the Winter 2012 issue, entire issues of *JEP* can be downloaded in pdf or e-reader formats.

My job as Managing Editor of *JEP* has been a pride and a pleasure for these last 25 years. It's consistently interesting work: after all, my job is to do close readings of the highly varied work of a succession of prominent economists who are trying to explain their thinking—and then to ask them questions until they explain it all to me! Editing an academic journal also offers the psychic frisson of leaving something behind: 100 issues and counting, to be precise. When I visit another college or university, I sometimes walk through the periodical stacks just to see *JEP* on the shelf. Running an academic journal for a long time offers a pleasing sense of place within the discipline of economics, spinning a web of personal contacts from the up-and-comers to the well-established in academic institutions around the world. Some of my friends refer to my job at the journal as "the guy who gets thanked" at the end of articles. There are worse epitaphs.

## References

**Bierce, Ambrose.** 1911. *The Devil's Dictionary.* http://www.thedevilsdictionary.com/.

**Crovitz, L. Gordon.** 2009. "Making Old Media New Again." *Wall Street Journal*, April 12. http://online.wsj.com/article/SB123958338833312319.html.

**Dawidoff, Nicholas.** 2002. *The Fly Swatter: How My Grandfather Made His Way in the World.* New York: Pantheon.

**Ellison, Glenn.** 2002. "The Slowdown of the Economics Publishing Process." *Journal of Political Economy* 110(5): 947–93.

**Aristides [Epstein, Joseph].** 1996. "A Real Page-Turner." *American Scholar* 65(2): 167–74.

**Ferguson, Andrew.** 2007. "Dear Mr. Buckley, Cancel My Subscription." *Wall Street Journal,* December 20, p. D9.

**Jenkins, Holman W., Jr.** 2007. "My Sweet Press Lord: We'll Take the *Washington Post,* Please." *Wall Street Journal,* June 6, p. A18.

**McAfee, R. Preston.** 2010. "Edifying Editing." *American Economist* 55(1): 1–7.

**McCloskey, Donald N.** 1983. "The Rhetoric of Economics." *Journal of Economic Literature* 21(2): 481–517.

**McCloskey, Donald.** 1985. "Economical Writing." *Economic Inquiry* 23(2): 187–223.

**Montaigne, Michel de.** 1579–1580 [1971]. Translation published as "On the Education of Children," in *The Complete Works of Montaigne: Essays, Travel Journal, Letters,* translated by Donald M. Frame, pp. 106–31. London: Hamish Hamilton.

**Plotnick, Arthur.** 1982. *The Elements of Editing.* Macmillan Publishing Company.

**Salant, Walter S.** 1969. "Writing and Reading in Economics." *Journal of Political Economy* 77(4, Part 1): 545–58.

**Strunk, William, Jr., and E. B. White.** 1979. *The Elements of Style.* Macmillan. Third edition.

**Taylor, Timothy.** 2009. "An Editor's Life at *The Journal of Economic Perspectives.*" *American Economist* 53(1): 48–59.

**Thurber, James.** 1959. "The Theory and Practice of Criticizing the Criticism of the Editing of *New Yorker* Articles." May 18, manuscript. (Published in *Collecting Himself: James Thurber on Writing and Writers, Humor, and Himself,* edited by Michael J. Rosen. HarperCollins, 1989.)

Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 41–64

# The Rise of Middle Kingdoms: Emerging Economies in Global Trade

Gordon H. Hanson

In the recent global financial crisis, there was a sharp divide in the economic performance of high-income and emerging-market nations. The United States, the countries of the European Union, and Japan suffered most. They have been slow to recover, with heavy debt burdens and enfeebled banks promising continued sluggishness. Many emerging economies, in contrast, hardly paused during 2008 and 2009. Led by China and India, their robust growth is now fueling the recovery of the global economy. The shift in economic power is palpable. Brazilian, Chinese, and Indian multinational firms are eagerly acquiring assets abroad; U.S. and European leadership in the World Trade Organization, once unassailable, has failed to consummate the Doha round of global trade negotiations; and the IMF has been spending more time worrying about the balance sheets of high-income nations than of lower-income ones.

Although events since 2007 have brought these changes in the world economy into sharp focus, the rise of low- and middle-income countries in global trade has been decades in the making. China's economic transition, which accelerated in the 1990s, allowed the country to realize a latent comparative advantage in labor-intensive products (Amiti and Freund 2010; Harrigan and Deng 2010). India's surge of growth, which began even before the reforms it initiated in 1991 (Rodrik and Subramanian 2004), was, like China's, aided by industries beginning far inside the technology frontier (Hsieh and Klenow 2009). The result of China's

■ *Gordon H. Hanson is Professor of Economics and Director of the Center on Emerging and Pacific Economies, University of California–San Diego, La Jolla, California. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His e-mail address is ⟨gohanson@ucsd.edu⟩.*

http://dx.doi.org/10.1257/jep.26.2.41. doi=10.1257/jep.26.2.41

and India's openings has been an immense global export supply shock. Between 1992 and 2008, average annual growth in exports was 18 percent in China and 14 percent in India. These two are not the only significant new players in global trade. Consider the next 15 middle-income countries, which (in order of market size) are Brazil, Korea, Mexico, Russia, Argentina, Turkey, Indonesia, Poland, South Africa, Thailand, Egypt, Colombia, Malaysia, the Philippines, and Chile. In 2008, they each had a GDP above $100 billion; as a group, their collective GDP is 1.4 times China and India's combined total. From 1992 to 2008, these 15 countries had average annual export growth of 8 percent. During this period, low- and middle-income countries overall saw their share of global exports more than double, from 21 to 43 percent.

In this paper, I examine changes in international trade associated with the integration of low- and middle-income nations into the global economy. From the 1950s to the 1980s, trade was dominated by flows between high-income countries both because they accounted for most of global GDP and because many developing countries maintained high barriers to imports. In the international economics literature, the exchange of goods between the United States, Canada, the nations of Western Europe, and Japan is often referred to as North-North trade. However, we are moving toward a world in which South-South commerce (trade between developing countries) and North-South commerce (trade between developed and developing countries) are overtaking North-North flows. Whereas high-income economies accounted for four-fifths of global trade in 1985, they will account for less than half by the middle of this decade.

I start by focusing on the growth in South-South trade. In part, this pattern has arisen because urbanization and industrialization in China and India have contributed to strong demand for raw materials needed to build cities and factories. Other developing economies are abundant in these materials and have become important sources of global supply. Growth in low- and middle-income economies has also deepened global production networks, with lengthening production chains in the South increasing cross-border flows of parts and components.

I then turn to the rise in North-South trade. During the 1980s, when North-North trade was dominant, explanations for global trade patterns shifted away from classic theories of comparative advantage. Models explaining trade as the result of differences in national factor supplies, as in the Heckscher–Ohlin tradition, or differences in sectoral labor productivities, as in the Ricardian tradition, seemed incapable of accounting for substantial trade flows between high-income countries. The literature instead explained these types of trade flows using models based on product differentiation and economies of scale (Helpman and Krugman 1985).

The rise in North-South trade has rekindled interest in the role of comparative advantage in global production. Countries export different types of goods at different stages of development, with low-income countries producing a narrow range of goods (Imbs and Wacziarg 2003; Cadot, Carrère, and Strauss-Kahn 2011) and moving up the product ladder in terms of capital intensity and quality as their incomes rise (Schott 2003, 2004). For some countries, and for China in particular, industrial specialization evolves rapidly (Rodrik 2006), revealing a capacity to speed up product ladders. For other countries, specialization in primary commodities, long seen as a hallmark of underdevelopment, has been a pathway to economic growth. Concomitant with recent changes in global trade, trade theorists have revived the Ricardian model (Eaton and Kortum 2002), which for years was used as little more than a tool for introducing undergraduates to international economics. The new theories rightly emphasize differences in national industrial capabilities as a driver of trade, but are not yet sufficiently developed to account for the full richness of the trade patterns that we see emerging.

## Growth in South-South Trade

As a starting point, I describe growth in trade between nations grouped by income level. I assign countries to income categories based on their per capita GDP in 1990, which characterizes their level of development at the beginning of the recent global trade surge.[1] Low-income countries are those with per capita GDP of less than $800 (in 2000 U.S. dollars); middle-income countries have per capita GDP of $800 to $10,000; and high-income countries have per capita GDP of $10,000 to $33,000. These categories correspond closely to World Bank definitions of country income status in 1990. China and India have their own category (whereas by the above cutoffs they would be defined as low-income countries).

Table 1 shows trade flows, normalized by (regional) GDP, between country income groups over the period 1994 to 2008 based on data from UN Comtrade (as are all other figures presented in the paper). Two properties of global trade are immediately apparent. One is that for low- and middle-income countries, trade as a share of regional GDP has grown sharply. Exports over GDP rise from 26 to 55 percent in low-income countries, from 25 to 55 percent in middle-income countries, and from 25 to 44 percent in China and India. For high-income countries, the change is much smaller, from 17 to 26 percent. Changes in imports as a share of GDP are similar.

The shifting pattern of international trade involves much larger South-South flows. Between 1994 and 2008, the share of exports from low-income countries going to low- and middle-income markets (including China and India) rose from 24 to 42 percent, with China and India accounting for about half of this growth. The share of exports from middle-income countries going to low- and middle-income markets (including China and India) rose from 33 to 46 percent, with China and India accounting for two-fifths of this growth.

Growth in trade shares for low- and middle-income countries far exceeds the increase in their relative economic size. Between 1994 and 2008, the share of

[1] I calculate per capita GDP in 1990 as the average over 1985 to 1995 to account for missing annual data in some countries and the creation of new nations after 1990.

*Table 1*
**Exports and Imports Relative to GDP by Regional Trading Partner**

| | | *Exports to partner relative to regional GDP* | | | *Imports from partner relative to regional GDP* | | |
|---|---|---|---|---|---|---|---|
| *Region* | *Trade partner* | *1994* | *2008* | *Percentage point change* | *1994* | *2008* | *Percentage point change* |
| Low-income countries | Low-income countries | 0.8% | 3.2% | 2.4 | 0.8% | 3.2% | 2.4 |
| | Middle-income countries | 4.5% | 11.6% | 7.1 | 6.0% | 17.1% | 11.1 |
| | China, India | 1.1% | 8.3% | 7.2 | 1.8% | 10.7% | 8.9 |
| | High-income countries | 20.0% | 31.8% | 11.8 | 15.1% | 23.0% | 7.9 |
| | World | 26.3% | 55.0% | 28.6 | 23.7% | 54.0% | 30.4 |
| Mid-income countries | Low-income countries | 0.7% | 2.1% | 1.4 | 0.5% | 1.4% | 0.9 |
| | Middle-income countries | 5.3% | 15.6% | 10.3 | 5.3% | 15.6% | 10.3 |
| | China, India | 2.2% | 7.5% | 5.3 | 2.4% | 7.4% | 5.0 |
| | High-income countries | 16.9% | 29.6% | 12.7 | 18.6% | 26.0% | 7.4 |
| | World | 25.1% | 54.8% | 29.8 | 26.8% | 50.4% | 23.6 |
| China and India | Low-income countries | 0.8% | 2.7% | 1.9 | 0.5% | 2.1% | 1.6 |
| | Middle-income countries | 9.5% | 15.2% | 5.7 | 8.6% | 15.4% | 6.8 |
| | China, India | 0.1% | 1.2% | 1.1 | 0.1% | 1.2% | 1.1 |
| | High-income countries | 14.3% | 25.3% | 11.0 | 9.8% | 14.1% | 4.3 |
| | World | 24.8% | 44.4% | 19.6 | 19.0% | 32.7% | 13.7 |
| High-income countries | Low-income countries | 0.3% | 0.7% | 0.4 | 0.5% | 1.0% | 0.5 |
| | Middle-income countries | 3.7% | 6.6% | 2.9 | 3.4% | 7.5% | 4.1 |
| | China, India | 0.5% | 1.7% | 1.2 | 0.7% | 3.1% | 2.4 |
| | High-income countries | 12.8% | 16.9% | 4.1 | 12.8% | 16.9% | 4.1 |
| | World | 17.4% | 26.0% | 8.6 | 17.4% | 28.6% | 11.2 |

*Source:* UN Comtrade, ⟨http://comtrade.un.org/⟩.

low- and middle-income countries (including China and India) in global GDP increased from 22 to 29 percent. The gravity model of trade, which is a workhorse for empirical research on trade flows, expresses exports from one country to another as a function of the countries' GDPs, bilateral trade costs, and relative prices (Anderson and van Wincoop 2004). Following the gravity logic, the share of low- and middle-income countries in global trade should increase in rough proportion to their share of global income. But Southern trade has grown much faster than Southern GDP.

What accounts for the surge in South-South commerce? A first possible explanation is falling trade costs in emerging economies, resulting from unilateral trade reform, growth in World Trade Organization membership, or reduced costs of shipping goods. But these explanations are not well-supported by more detailed research. Hummels (2007) documents in this journal that while the costs of air transport have fallen significantly in recent decades, the costs of ocean transport, the mode of transport for most developing-country trade, have not. Between 1994 and 2008, policy barriers to trade have fallen, with the average applied tariff rate across all goods (weighted by imports) declining from 12 to 4 percent in middle-income countries and from 29 to 8 percent in China. In high-income nations, already-low tariffs meant further reductions were small, with average tariffs falling from 5 to 3 percent. However, estimates by Yi (2003), using data for an earlier period, suggest that such modest tariff changes are too small to explain the more than doubling of trade as a share of GDP in developing economies.

The importance of the World Trade Organization in expanding trade appears to be overblown. Since 1995, when the WTO was created out of the General Agreement on Trade and Tariffs, 41 new nations have joined the organization, bringing membership to 153 countries. Yet, the literature provides ambiguous support for the idea that WTO membership expands trade. Rose (2004a) finds that, conditional on GDP, WTO members do not trade significantly more than nonmembers, which he attributes to WTO members not having more liberal trade policies (Rose 2004b), partly as a result of the WTO placing weak demands on developing countries to liberalize trade.

An alternative explanation is that the growth in Southern trade is a result of expanding multistage global production networks. Much of the recent increase in trade appears to be the result of offshoring, in which firms fragment manufacturing across borders by locating individual production stages in the countries where they can be performed at least cost (Hummels, Ishii, and Yi 2001). A consequence may be that *gross* trade flows (that is, total exports) overstate *net* trade flows (that is, exports net of imported intermediate inputs); if true, this would imply that the expansion of South-South trade is in part a statistical artifact. If goods are produced through a sequence of stages, as modeled by Costinot, Vogel, and Wang (2011), each country will add value as a product is transformed from raw inputs into a final output along a production chain than spans national borders; the value added of countries participating earlier in the chain will therefore be counted in trade flows multiple times. The Intel Corporation, for example, produces semiconductors by first manufacturing silicon wafers in the United States, Ireland, and Israel and then assembling and testing integrated circuits made out of these wafers at plants in China, Costa Rica, Malaysia, and the Philippines. Silicon wafers are counted in trade twice, first in shipments of the raw wafers from the United States to Costa Rica, and again in the shipment of integrated circuits that embody the wafers from Costa Rica to the final destination market.

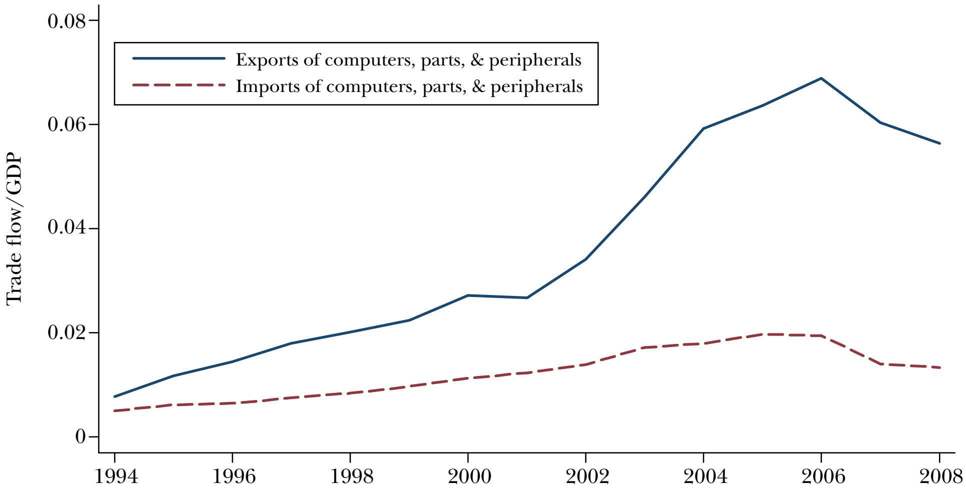
Global production networks, however, need not be based on sequential manufacturing. Dell follows an alternative model in making computers, in which it subcontracts the production of parts and components to suppliers in many countries and has these parts shipped to final destination markets, where they are assembled into computers for final consumers. If production networks tend to follow the Intel model, total exports may greatly exceed trade in value added, whereas if they tend to follow the Dell model they may not.

What fraction of measured trade flows is subject to concerns over double counting? In the case of China, half of its manufacturing exports in the late 1990s and early 2000s were produced by processing plants that assemble imported parts and components into final goods for export, primarily in consumer products (Feenstra and Hanson 2005). Within export processing plants, value added in China as a share of total exports is only 36 percent. However, domestic value added in China's exports outside of processing plants appears to be much higher. Koopmans, Powers, Wang, and Wei (2010) estimate that for China's overall exports, domestic value added accounts for 65 percent of total exports. For the world as a whole, 75 percent of exports consists of value added in the country of export. Among middle-income countries, the share of domestic value added in total exports is relatively low in Malaysia (59 percent), Mexico (62 percent), the Philippines (58 percent), and Thailand (60 percent) and relatively high in Brazil (87 percent), Indonesia (77 percent), Russia (89 percent), and South Africa (81 percent). The latter group of countries specializes in producing commodities, in which fragmentation of production is less feasible, whereas the first group specializes in manufacturing, where production chains are more common. In Mexico, for instance, half of manufacturing exports are by *maquiladoras,* plants that assemble final goods in electronics, automotive goods, and machinery from parts and components imported from the United States (Bergin, Feenstra, and Hanson 2009).

Other evidence confirms that double counting in recorded total trade flows is more severe for manufactured goods. Johnson and Noguera (2012) find that the ratio of export value added to total exports is lower in manufacturing (44 percent) than in agriculture and natural resources (109 percent, where a value of greater than 100 percent indicates that a large fraction of exports in the sector are indirect in the form of value added embodied in the final exports of other sectors). Many middle-income countries specialize in manufacturing exports, suggesting that global production networks are part of the reason their trade is expanding so rapidly. But double-counting as a result of such networks cannot be the entire story. Most low-income and some middle-income nations specialize in primary products, including minerals and farm goods. For these countries, the increase in exports to GDP is not an artifact of total exports overstating value added in exports but is instead a reflection of increasing specialization for global markets.

Evidence on the share of value added in total exports is based largely on cross-sectional data; there is little work providing clues on the expansion path of production chains. We do not know whether the share of value added in total exports is rising or falling. There is some evidence that emerging economies are deepening their productive capacity, capturing production of intermediate inputs that they previously imported from abroad, a phenomenon that is pronounced in China. By the mid-2000s, exports of completed computers had become China's top export good. Many of these computers are assembled in export processing plants, requiring China to import components from abroad.

*Figure 1*
**China's Imports and Exports of Computers and Parts**



*Source:* Author's calculations using (World Bank) World Development Indicators and UN Comtrade.
*Note:* The figure shows exports and imports of computers, computer parts, and computer peripheral devices (Standard International Trade Classification products 752, 7512, and 7519) in China over the period 1994 to 2008.

However, over time China's reliance on imports in the sector has declined markedly. Figure 1 shows exports and imports of computers, computer parts, and computer peripheral devices (Standard International Trade Classification products 752, 7512 and 7559) in China over the period 1994 to 2008. In 1994, exports were 1.6 times imports in the sector; by 2008, they were 4.2 times imports. While it is unclear how much one can generalize from China's experience, growth in trade involving middle-income manufacturers does not necessarily go hand in hand with greater back and forth flows of intermediate inputs.

As China develops, it may continue to take over the manufacture of inputs, making production in some sectors less fragmented globally. This experience is similar to that of Hong Kong, Korea, Singapore, and Taiwan, which also entered global production networks by specializing in product assembly and later expanded into input production and the design and distribution of goods. But not all countries that begin as assemblers succeed in graduating into other manufacturing stages. After nearly three decades of global manufacturing, most exporters in Mexico and Central America remain in the assembly stage.

Growth in South-South trade is a major part of the recent global trade boom. Falling trade barriers and expanding global production networks have surely contributed to Southern trade growth. However, they appear insufficient on their own to explain why trade to GDP ratios have risen so much in low- and middle-income countries. What else could be behind the rapid expansion of trade relative to GDP? One possibility, as yet unexplored in the literature, is that the greater role of emerging economies in global trade is inducing a much finer degree of international specialization than occurred during previous decades, in which North-North trade predominated. As I discuss next, such an explanation would require that comparative advantage is assuming a larger role in determining global trade flows than it did in the past.

## The Return of Comparative Advantage

The 1980s and 1990s were not kind to theories of trade based on comparative advantage. The dominance of high-income countries in global commerce led international economists to develop models that explain trade as the result of increasing returns to scale and monopolistic competition in differentiated products (Helpman and Krugman 1985). The Heckscher–Ohlin model, once a staple of graduate training in international trade, failed repeatedly to explain prevailing trade patterns (Feenstra 2003). And the Ricardian model (based on sectoral labor productivities) remained little more than an intellectual curiosity, given its tendency to predict extreme patterns of industry specialization, seemingly at odds with the data (at least for Northern countries). Further, the robust success of the gravity model, in which country size and trade costs are the primary determinants of trade flows, seemed to defy a significant role for comparative advantage (Anderson and van Wincoop 2003). If we can explain much of bilateral trade using the size and location of the importer and exporter, why do we need comparative advantage at all?

Much has changed in the last decade. On the empirical side, China's and India's growth are powerful reminders that cross-country differences in technology and resources are potent motivations for commerce. On the theory side, Eaton and Kortum (2002) show that Ricardian comparative advantage is consistent with the gravity model. In their formulation, an exporter captures a share of an importer's market according to its technological capability and its trade costs in delivering goods. They avoid the knife-edge feature of the textbook Ricardian model, in which a country tends to supply either the entire market for a good or none, by having productivity vary randomly across firms within a country and the position of the country's productivity distribution in an industry be given by its predetermined technological capability, which they liken to absolute advantage. These technological capabilities (along with the dispersion of productivity across firms) are the key primitives of the model. Other models have comparative advantage arising from country differences in factor supplies that result in trade in intermediate inputs associated with the formation of global production networks (Feenstra 2010). In either set of theories, it is differences between countries, whether in terms of their technological capabilities or their factor supplies, that cause trade.

## International Specialization

A role for comparative advantage in trade is evident in the pattern of net exports by sector across country income groups, as shown in Figures 2A–D. I group non-oil exports into nine categories: agriculture and food products; minerals and other raw materials; apparel, footwear, and textiles; metals and metal products; chemicals; machinery; electronics and electrical machinery; transportation equipment; and other manufactures.[2] Agriculture and raw materials are intensive in the use of land, mineral reserves, and in some cases raw labor. Apparel, footwear, textiles, some electronics, and other manufactures (which include furniture, toys, and games) are intensive in the use of low-skilled labor. And chemicals, machinery, some electronics, and transportation equipment are intensive in human and physical capital.

International specialization, at least in broad overview, follows perceived patterns of comparative advantage. Low-income countries (Figure 2A) have positive net exports in three resource- or labor-intensive sectors—agriculture, raw materials, and apparel and shoes—and negative net exports in other sectors. China and India (Figure 2B) have positive net exports in three labor-intensive sectors—apparel and shoes, electronics, and other manufactures—and negative or negligible net exports in other sectors. Middle-income countries (Figure 2C) have negative net exports in the three capital-intensive sectors—chemicals, machinery, and transportation equipment—and export strength in other goods. And high-income countries (Figure 2D) have positive net exports in the three capital-intensive sectors and negative net exports elsewhere.

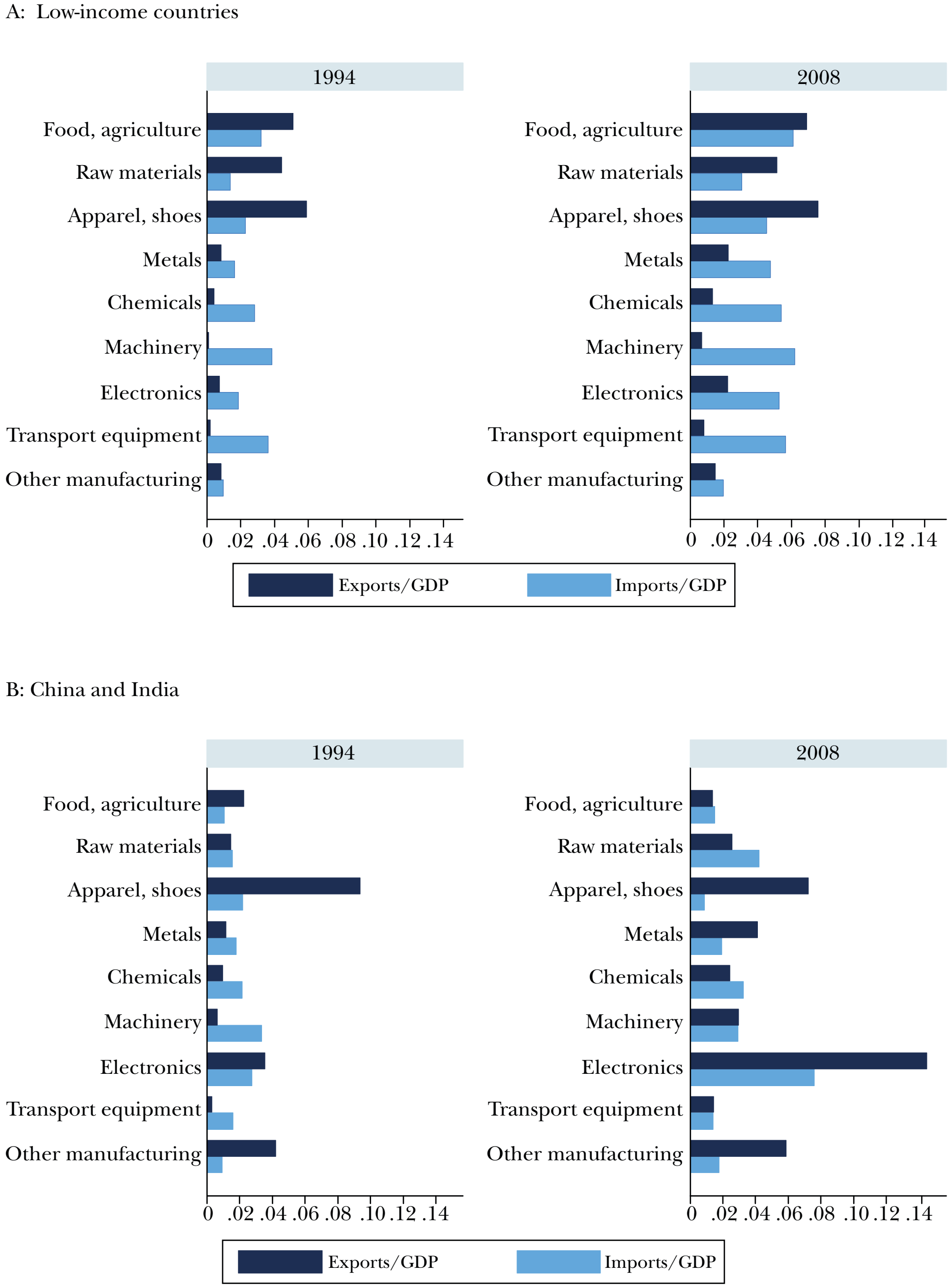
Underlying these specialization patterns is growing South-South trade along comparative advantage lines, with resource-poor emerging economies purchasing raw materials from ones that are resource rich. For low-income countries over the period 1994 to 2008, shipments to low- and middle-income markets (including China and India) accounted for 70 percent of their export growth in agriculture and food products and 73 percent of their export growth in raw materials. China and India are particularly important sources of raw material demand, absorbing 40 percent of low-income country growth in exports of these goods over the period. Not surprisingly, emerging economies are a relatively unimportant source of demand for apparel and textiles, absorbing only 25 percent of low-income country export growth in that sector. Low-income countries send most of their output of clothing and shoes to high-income markets.

Middle-income countries export a diverse set of goods, owing to the heterogeneity of countries within the group. Individual countries tend to specialize in a

[2] The corresponding one- and two-digit Standard International Trade Classification (SITC) products are: agriculture and food products (SITC 0, 1, 4, 21, 22, 29, and 94); raw materials, which include rubber, wood, paper, iron ore, and other minerals (SITC, 23–25, 27–28, 62–64, and 66); apparel, footwear, and textiles (SITC 26, 61, 65, and 83–85); metals and metal products (SITC 67–69), chemicals (SITC 5), machinery (SITC 71–74), electronics and electrical machinery (SITC 75–77), transportation equipment (SITC 78–79, 95), and other manufactures, which include toys and games, plumbing and light fixtures, furniture, professional and scientific equipment, photographic and optical equipment, watches, and miscellaneous goods (SITC 81–82, 87–89). Petroleum, coal, and natural gas (SITC 3) are excluded.

*Figure 2*

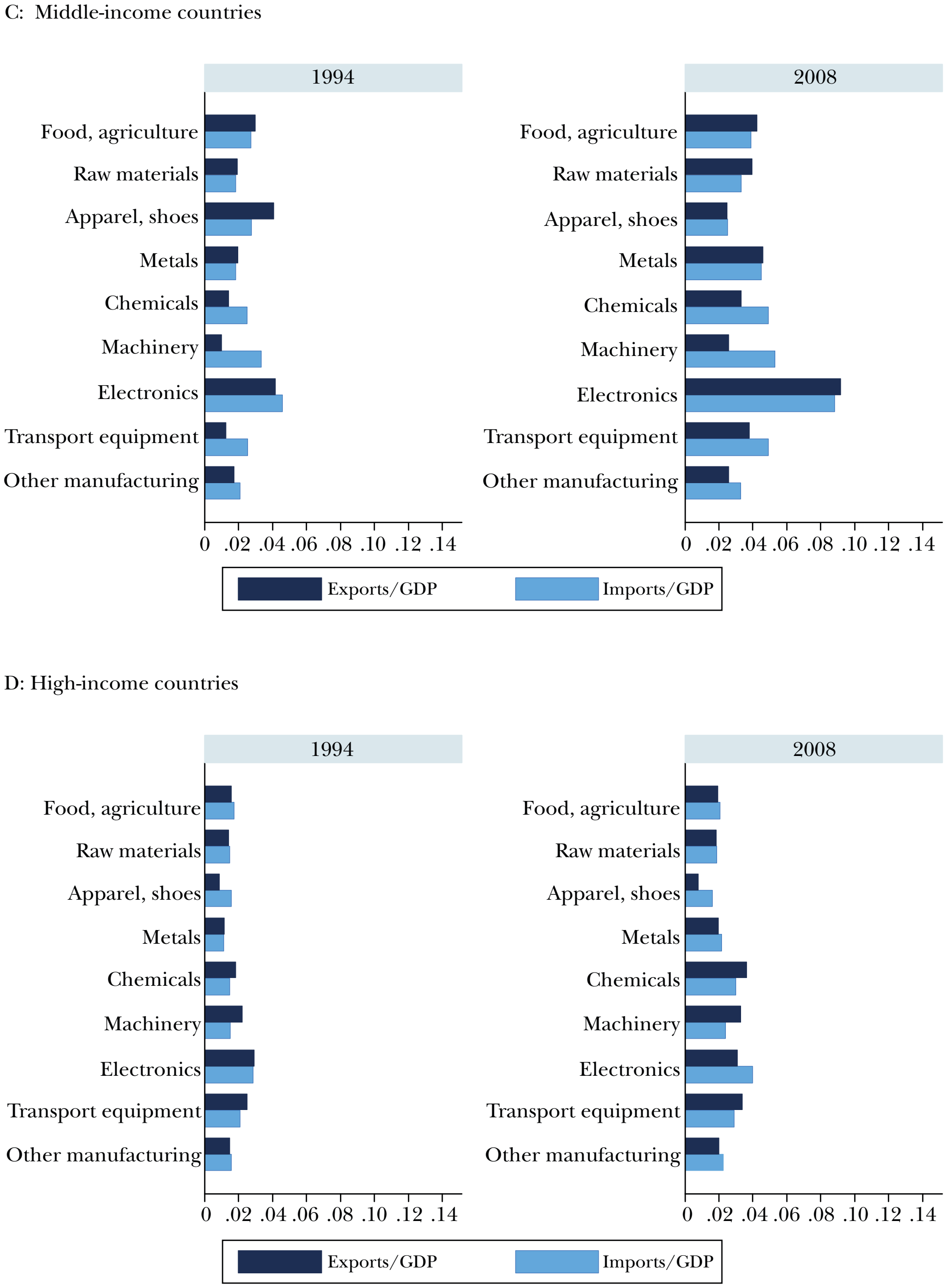
**Sector Trade Shares of GDP, by Income Group**



A: Low-income countries

B: China and India

*Figure 2 (continued)*



C: Middle-income countries

D: High-income countries

*Source:* Author's calculations using (World Bank) World Development Indicators and UN Comtrade.

subset of sectors. Overall, middle-income countries have had strong export growth in agriculture, led by Argentina and Brazil; metals, led by Russia, Korea, South Africa, and Chile; electronics, led by Korea, Malaysia, Thailand, and the Philippines; and transportation equipment, led by Korea, Mexico, Poland, and Turkey. In each of these sectors, except autos, low- and middle-income markets absorbed 50 percent or more of middle-income country export growth over the period 1994 to 2008. China and India alone accounted for more than 25 percent of absorption of the export growth of middle-income countries in raw materials and electronics. Their raw material demand reflects the need for iron ore, copper, and other minerals they require to build their economies; their electronics demand, coming primarily from China, represents the deepening of production networks among emerging economies. China and India are distinct among low- and middle-countries for being reliant on high-income markets to absorb their ever-growing exports. High-income countries absorbed over 70 percent of China's and India's export growth in apparel, footwear, and other manufactures and over 55 percent in electronics (one of China's strengths) and metals (one of India's).

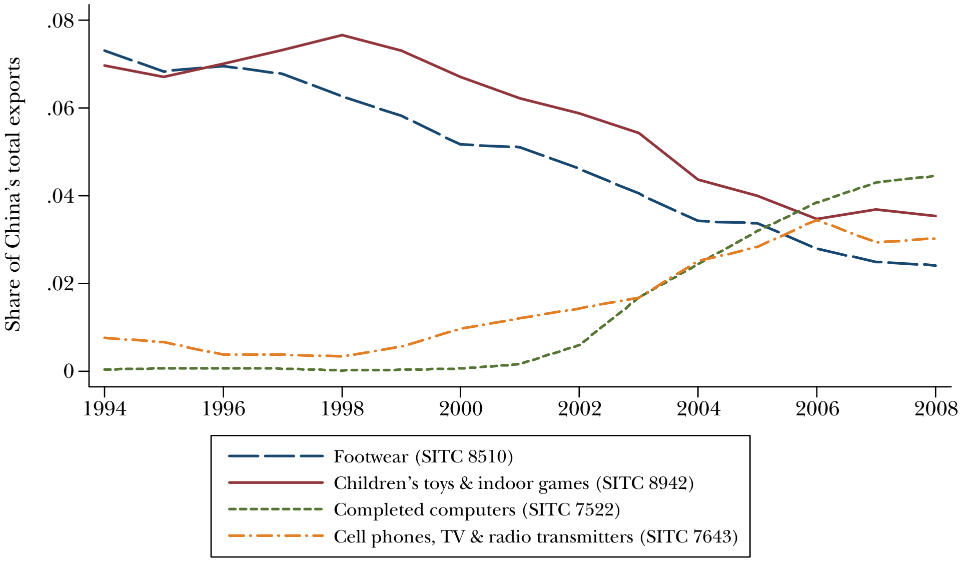
Foreign direct investment is abetting the growth of emerging-economy trade. North-to-South flows of foreign direct investment are well known to international economists. A large literature documents the importance of these flows in building global production chains (for example, Hanson, Mataloni, and Slaughter 2005; Harrison and MacMillan 2011; Becker and Muendler 2010). Between 1994 and 2008, inflows of foreign direct investment as a share of GDP rose from 2.1 to 3.4 percent in low-income countries, rose from 1.3 to 4.4 percent in middle-income countries, and held steady at 3 percent in China and India. Much less appreciated is the growth in *outward* foreign direct investment by emerging economies. Outflows of foreign direct investment as a share of GDP rose over the 1994 to 2008 period from 0.2 to 2.2 percent of GDP in middle-income countries and from 0.2 to 1.2 percent of GDP in India and China. For comparison, outflows of foreign direct investment from high-income countries were 3.6 percent of GDP in 2008.

## Dynamics in Specialization

The cross-sectional view of trade data seen in the last section highlights what appears to be specialization according to comparative advantage, whether Heckscher–Ohlin (resource-based) or Ricardian (technology-based) in origin. A dynamic view of trade reveals that specialization in low- and middle-income countries can change rapidly over time.

The data from Figure 2C show that for middle-income countries in 1994, apparel and footwear was the top sector for net exports, but that by 2008, electronics had displaced it. This change is consistent with middle-income economies accumulating human and physical capital, pushing them out of labor-intensive clothes and shoes and into more capital-intensive goods (Schott 2003). Low-income countries, including Bangladesh and Vietnam, are moving in to fill the space vacated by middle-income countries in apparel. The largest changes in specialization occur in China and India. As shown in Figure 2B, China's and India's net exports as a

*Figure 3*
**China's Top Export Products, 1994–2008**



*Source:* Author's calculations using (World Bank) World Development Indicators and UN Comtrade.

share of GDP decline over 1994–2008 in apparel and footwear, the sector with their largest net exports in 1994. Since the early 1990s, China has been moving into more sophisticated products, including cellphones and computers. The sector with the largest growth in net exports is electronics, with an increase in net exports to GDP of 10 percentage points, followed by metals and machinery, each with increases of net exports to GDP of 2.8 percentage points. In the future, China and India may acquire comparative advantage in new sectors, such as chemicals or transportation equipment, as occurred in the last decade with machinery.

China's recent experience is worth a closer look. Figure 3 shows the share of four (Standard International Trade Classification four-digit) products in China's total exports. The first two, footwear and children's toys, were China's top two exports in 1994, the former accounting for 7.3 percent of total shipments and the latter for 7.0 percent. By 2008, the export shares of these two products had declined to 2.4 and 3.5 percent, respectively. Joining shoes and toys among China's top exports are completed computers, with 4.5 percent of total shipments in 2008, making this the country's top export good, and cellphones, TVs, and radio transmitters, with 3.0 percent of total shipments in 2008, making this category China's third–largest export. China's transition into computer production happened virtually overnight, with most of its export growth in the product occurring between 2002 and 2005.

Of course, if China is just progressing from assembling tennis shoes to assembling laptops, the change in its export patterns would not be all that impressive.

While export processing plants continue to account for a large share of China's total shipments abroad, as Figure 1 showed earlier, China's exports of computers and computer parts have grown much more rapidly than its imports of final and intermediate goods in the sector, suggesting that at least part of China's graduation from apparel to electronics also involves capturing more links in global production chains. Huawei and Lenovo, companies based in China and controlled by Chinese investors, are respectively the world's second-largest manufacturer of mobile telecommunications equipment and the world's fourth-largest manufacturer of laptops and personal computers. Over time, China is both manufacturing more technologically advanced goods and accounting for a larger share of value added in their production.

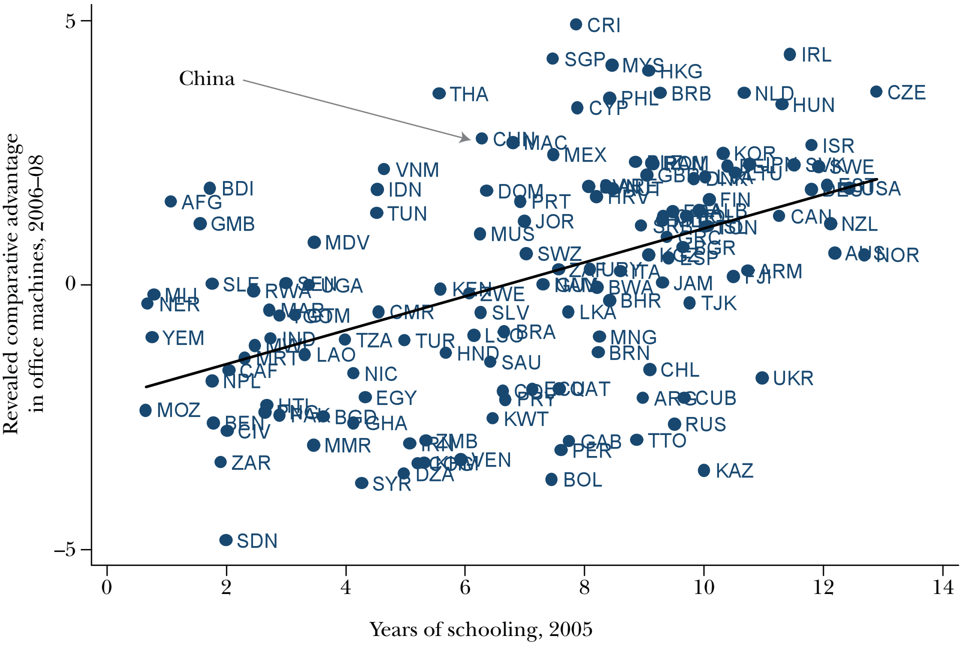
Is China's specialization in computers unusual for a country at its level of development? Rodrik (2006) doubts that China's export strength in electronics is attributable to comparative advantage, suggesting instead that the country has used industrial policy to expand high-tech production. Figure 4 plots countries' revealed comparative advantage in office machines—Standard International Trade Classification (SITC) industry 75—averaged over 2006 to 2008, against the average years of schooling of the adult population in 2005 (Barro and Lee 2010). Revealed comparative advantage in computers is defined as the log ratio of a country's share of world exports of SITC 75 to its share of world exports of all merchandise.[3] China is above the regression line, indicating that its specialization in the sector is greater than one would expect given its level of education, but it is hardly an extreme outlier. Other middle-income countries—including Costa Rica, the Philippines, Malaysia, and Thailand—have larger positive residuals. While China's rapid export growth in electronics grabs one's attention, its current specialization in the sector does not seem unwarranted given its stock of human capital.

The results in Figure 4 suggest that international specialization in computers is associated with cross-country variation in the supply of skilled labor. More generally, we would like to know whether the accumulation of human and physical capital by middle-income countries explains their rapid progression into electronics, which is visible in Figure 2. The literature has yet to address the issue—which is surprising given that the growth of international trade in electronics is perhaps the single most important factor behind the expansion of global production networks.

Another perspective on China's evolution highlights that while the country is shifting into more advanced sectors, it remains locked into producing low-priced goods. Schott (2008) finds that the overlap between the products that the U.S. imports from China and from OECD countries is substantial. Between 1983 and 2005, the similarity of the U.S. import bundle from China relative to that for the OECD jumped from rank 13 among developing countries to rank 4, behind Korea, Mexico, and Taiwan. As of 2005, U.S. imports from China covered 89 percent of

[3] To purge the revealed comparative advantage index of the effects of country size, I use the residuals from a regression of the index on log country population.

*Figure 4*
**Education and Exports of Office Machines**



*Source:* Author's calculations using (World Bank) World Development Indicators and UN Comtrade.
*Notes:* Figure 4 plots countries' revealed comparative advantage in office machines—Standard International Trade Classification (SITC) industry 75—averaged over 2006 to 2008, against the average years of schooling of the adult population in 2005. Revealed comparative advantage in computers is defined as the log ratio of a country's share of world exports of SITC 75 to its share of world exports of all merchandise. The countries are indicated by their World Bank abbreviations.

all ten-digit products in the Harmonized System,[4] compared to 97 percent for the OECD as a whole. Despite the breadth in the goods that China exports, and the similarity of its product categories with far richer countries, China appears to occupy a down-market niche within narrowly defined goods. Schott (2008) finds a large price discount on Chinese imports in the United States. He regresses the unit value (average price) of U.S. imports on product-year dummy variables, the distance between the United States and the exporting country, the U.S. applied tariff rate on the product, and the exporting country's per capita GDP. In the 2000s, unit values on Chinese imports are 48 log points lower than those of other countries.

[4] The trade data used in this paper are based on the Standard International Trade Classification system developed by the United Nations (see ⟨http://unstats.un.org/⟩). Recently, trade data have become available based on the Harmonized System of product classification developed by the World Customs Organization (see ⟨http://www.wcoomd.org/⟩), which provides more disaggregated product categories than are available in the SITC system.

Lower unit values for U.S. imports from China may indicate that the country produces goods of inferior quality, leading to lower market prices. However, China's experience isn't all that different from some comparison countries. For the 2000s, the discount on unit values for U.S. imports from Korea is 45 log points, from Japan is 33 log points, and from Mexico is 59 log points (Schott 2008). Whatever accounts for the relatively low average prices of U.S. imports from China, its unit values are similar to its neighbors in either geographic or product space. Further, lower average import prices do not necessarily imply lower quality. Japan and Korea have a strong reputation for quality. And the success of Lenovo, the Chinese manufacturer of laptops, demonstrates that the country is capable of producing high-quality goods.

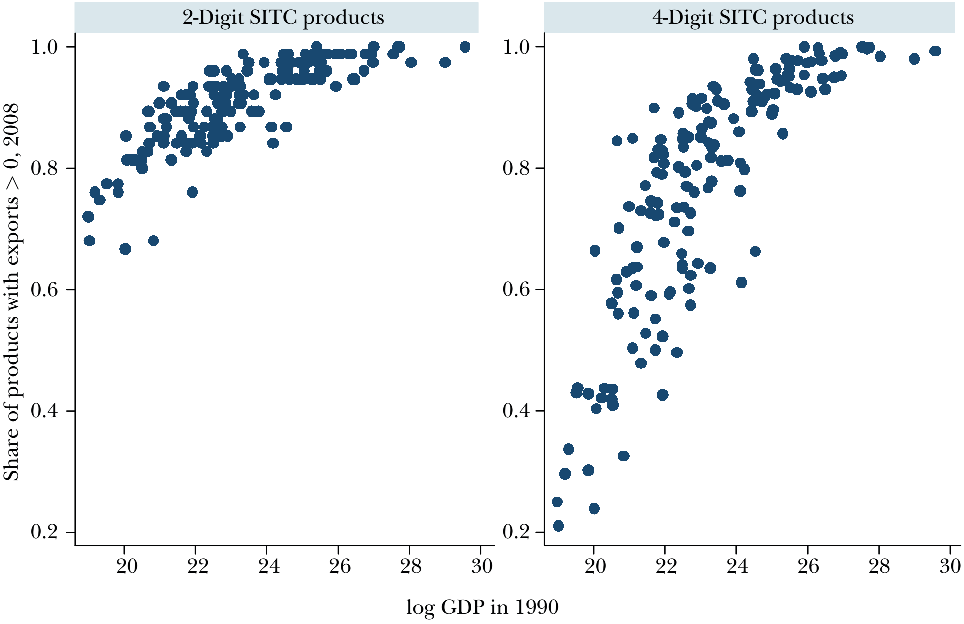
## Hyper-specialized Exporters

We have now seen that: 1) at any moment in time, country specialization by broad sector appears to be consistent with standard models of comparative advantage, and 2) over time specialization patterns evolve rapidly, perhaps in line with factor accumulation. Missing in the discussion is information on what is happening inside the broad sectors. When we drill down, we observe a fine degree of specialization in which the exports of low- and middle-income countries are concentrated in a relatively small number of products (Easterly and Reshef 2009). Such hyper-specialization is harder to explain with standard trade models.

Many developing countries have zero exports in broad swaths of product space. Using data for 2008, Figure 5 shows the fraction of products in which countries have greater than zero exports, plotted against log real GDP, as a measure of country size. At the two-digit Standard International Trade Classification level, there are 69 products (examples would include cereals, pharmaceutical products, electrical machinery), and at the four-digit SITC level there are 786 products (for example, milled rice, antibiotics, semiconductors). Most countries with GDP of less than $3.6 billion (an example would be Senegal) have positive exports in fewer than 60 percent of four-digit products and in fewer than 80 percent of two-digit products. It is not until countries reach a GDP of $72 billion (Iran) that they tend to export the full range of two-digit goods and not until $195 billion (Sweden) that they export the full range of four-digit products.

Further, exports of most countries are concentrated in a small number of goods. Figure 6 is concerned with the share of a country's top products in total trade (its single top product, its four top products, and its eight top products at the three-digit SITC level). It shows an export-weighted average of these shares by country income category for 2000 and 2008. I exclude petroleum exports (although graphs including them are similar), and I aggregate up to the three-digit level (of which there are 238 products) to account for reporting anomalies in some countries. For low-income countries in 2008, the share of exports accounted for by the single largest three-digit good is a whopping 21 percent; for the top four goods, the share of exports is 45 percent; and for the top eight goods, the share is 58 percent. Hyper-specialization extends beyond poor nations. In middle-income countries, the one-, four-, and eight-good export concentration ratios are 16, 37,

*Figure 5*
**Share of Products with Positive Exports by Country**



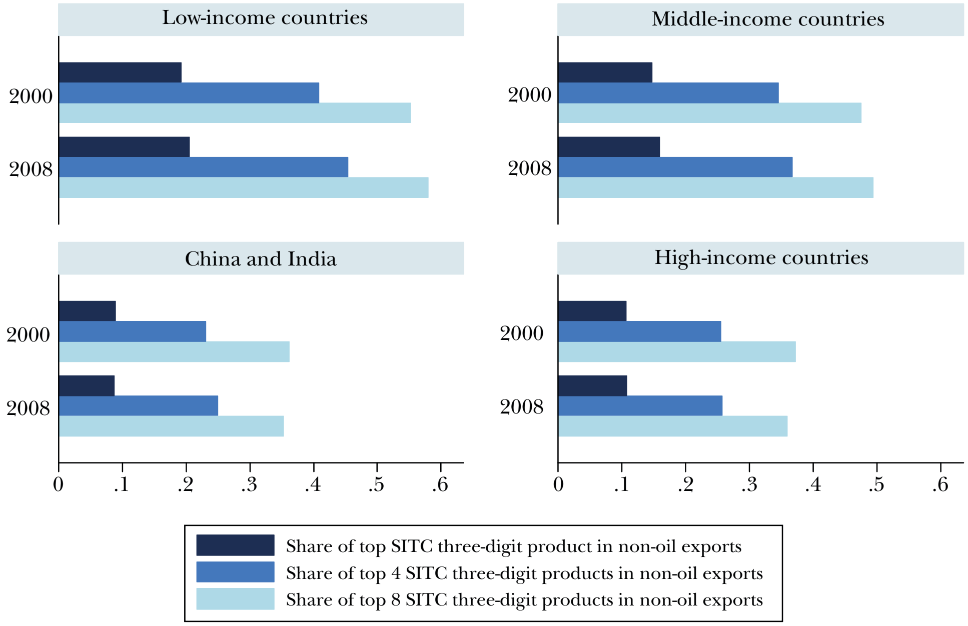
*Source:* Author's calculations using (World Bank) World Development Indicators and UN Comtrade.
*Note:* Using data for 2008, Figure 5 shows the fraction of products in which countries have greater than zero exports, plotted against log real GDP, as a measure of country size.

and 49 percent, respectively, and in high-income countries they are 11, 26, and 36 percent. For comparison, the U.S. ratios are 5, 17, and 28 percent.

For low-income countries, the most common top four-digit SITC export products (in order of frequency) are petroleum, unwrought aluminum, tea, coffee, edible nuts, raw cotton, diamonds, copper, and knitted apparel. For middle-income countries, the most common export products are petroleum, semiconductors, autos, knitted apparel, frozen fish, cane sugar, aluminum ore, diamonds, ferro-alloys, copper, and ships. The exports from the two groups overlap, with middle-income countries adding goods intensive in human or physical capital (microchips, cars, metals, and boats) to the list. Differences between the low- and middle-income lists may reflect countries investing in education and machinery as they develop.

Specialization in a small number of exports is not simply a function of developing countries having small economies and therefore producing a relatively narrow range of goods. Even the largest middle-income economies hyper-specialize. Taking the largest middle-income economy in each geographic region, Brazil's top 2008 exports are iron ore (11 percent of total exports), petroleum (8 percent), and soya beans (6 percent); Korea's are semiconductors (11 percent), autos (7 percent), and ships (7 percent); Mexico's are petroleum (14 percent), televisions (8 percent), and autos (8 percent); Indonesia's are coal (9 percent), palm oil (8 percent), and

*Figure 6*
**Concentration of Non-Oil Exports in Top Products**



*Source:* Author's calculations using (World Bank) World Development Indicators and UN Comtrade.
*Notes:* Figure 6 is concerned with the share of a country's top products in total trade (its single top product, its four top products, and its eight top products at the three-digit SITC level). It shows an export-weighted average of these shares by country income category for 2000 and 2008. I exclude petroleum exports (although graphs including them are similar), and I aggregate up to the three-digit level (of which there are 238 products) to account for reporting anomalies in some countries.

petroleum (7 percent); Poland's are autos (6 percent), auto parts (5 percent), and televisions (3 percent); and South Africa's are platinum (13 percent), coal (7 percent), and diamonds (6 percent). Of this group, only Poland's top three exports account for less than 25 percent of the country's total foreign shipments. For comparison, the top U.S. exports are aircraft (5 percent), petroleum products (4 percent), and microcircuits and transistors (4 percent).

What explains hyper-specialization in exporting? One account comes from the booming literature on firm heterogeneity and trade. Following the empirical findings of Bernard and Jensen (1999) that most firms do not export, Melitz (2003) develops a model in which firm productivity is a random variable (drawn from a Pareto distribution that is identical across countries) and firms face fixed costs in exporting goods abroad. Helpman, Melitz, and Rubinstein (2008) extend the Melitz model to account for the fact that the majority of bilateral trade flows are zero (that is, most pairs of countries do not trade) (Santos Silva and Tenreyro 2006; Baldwin and Harrigan 2011). Key to their explanation is the perhaps strong assumption that the distribution of firm productivity is bounded from above, meaning that in each industry there is a maximum level of productivity that a firm can attain. Consequently, for pairs of countries in which the importer has high trade barriers or a small market, no firm in the exporting country will be productive enough to justify the fixed cost of shipping to that market. We will therefore observe zero trade from the exporter to the importer.

Returning to Figure 5, the striking fact is that smaller countries have positive exports of fewer goods. In the Helpman, Melitz, and Rubinstein (2008) model, it is the size of the *importer's* market and not the size of the exporter's market that predicts zero trade. Further, in this model exports will be concentrated *within* industries (with more productive firms doing the lion's share of the trade) with no obvious pattern *between* industries (see Bernard, Redding, and Schott 2007, for an extension of the Melitz model that incorporates Heckscher–Ohlin features). The Melitz model therefore does not offer an obvious account of the hyper-specialization that we see in Figure 6.

Can the Eaton and Kortum (2002) Ricardian model explain hyper-specialization? Similar to Melitz, the Eaton–Kortum framework allows for heterogeneous firm productivity, but in a Ricardian setting such that country differences in technological capabilities dictate the share of import markets captured by exporting countries. Eaton–Kortum would ascribe the patterns of sector specialization in Figure 2 to country differences in these capabilities (Costinot, Donaldson, and Komunjer 2011). To account for the hyper-specialization seen in Figure 6, these capabilities would have to differ sharply across countries. The framework is silent about where technological capabilities come from, though Chor (2010) finds that in cross-section data these capabilities are correlated with country supplies of capital and labor and with country institutional characteristics such as financial development and the legal environment. However, there are limits to the applicability of the Eaton and Kortum approach. It predicts a smoothness to bilateral trade flows that does not allow for the preponderance of zeros at the exporter-product level seen in Figure 5. (Eaton, Kortum, and Sotelo 2012 attempt to extend the model to allow for zeros in trade flows.) While the Eaton–Kortum model gives us an elegant way of linking the gravity model of trade to comparative advantage, along the way it loses the extreme specialization of the simple Ricardian model, which matched well with trade patterns in many emerging economies.

A second explanation for hyper-specialization comes from Rodrik (2006) and Easterly and Reshef (2009), who suggest that exports are subject to externalities in production. The logic of external economies of scale is that when one firm expands production in an industry, it lowers costs for other firms through knowledge spillovers or through pecuniary externalities associated with making inputs available at a lower cost. Externality-based explanations for the location of production date back to Alfred Marshall's (1920) discussion of the development of the English textile industry and are typically associated with understanding where and how manufacturing begins and develops. However, the products that dominate exports for low- and middle-income countries include many primary commodities. It is not obvious why externalities should be important in the export of soya beans, edible nuts, aluminum ore, or copper. If they are, they have yet to be documented in the literature.

Available theories of trade are capable of explaining specific features of global commerce, such as why trade has a gravity structure, why countries specialize, or why so few firms export, but they do not yet appear capable of explaining the rich tableaux of trade patterns that we observe through the growing importance of low- and middle-income countries in the world economy. Countries at different income levels produce different types of goods; specialize according to broad sector and, within these sectors, in a relatively small number of products; export many goods not at all; and are capable of progressing rapidly up the ladder in terms of product sophistication. A model that could explain these outcomes would need prominent roles for comparative advantage, for extreme specialization even in larger countries, and for rapid changes in specialization associated with factor accumulation or technological progress.

## Final Discussion

The dramatic growth of China, India, and other middle-income nations is transforming the global economy. It is changing who trades with whom, how production is organized across borders, and how the global gains from trade are distributed. Research is just beginning to take stock of the emerging-economy trade boom. An active body of work seeks to explain China's growth (Song, Storesletten, and Zilibotti 2011), its implications for global welfare (Hsieh and Ossa 2011; Levchenko and Zhang 2011), and its effect on economies of high-income countries (for example, Bernard, Jensen, and Schott 2006; Bloom, Draca, and Van Reenan 2011; Autor, Dorn, and Hanson 2011). Other literature examines motivations for offshoring and global production networks (for example, Yi 2003; Feenstra and Hanson 2005; Hanson, Mataloni, and Slaughter 2005; Grossman and Ross-Hansberg 2008; Costinot, Vogel, and Wang 2011). We know less about the empirical determinants of export specialization, the dynamics of specialization patterns, or why South-South trade looks so different from North-North trade.

As we look ahead to this research agenda, what are the questions that need to be answered? I can think of at least four. First, how much of the recent growth in global trade represents real value added? We know that for trade in manufactures, the share of national value added in export shipments is relatively low. China's export success is based in part on export processing plants that import parts and components and assemble them into final goods to ship abroad. Is the value-added share in exports rising or falling over time? Is production becoming more or less fragmented across borders? How much of the rising share of exports to GDP in developing countries represents a true increase in specialization for foreign markets?

Second, what explains hyper-specialization in exporting? The tendency for countries to rely on a handful of products for most of their exports makes it tempting to see nonconvexities at work, such as informational spillovers in learning about foreign markets or industry-level distribution networks whose creation entails substantial up-front investments. But the goods that tend to top the list of developing country exports look less like products that are information-intensive and more like ones that require the availability of mineral reserves (copper, iron ore) or specific types of agricultural land (soya beans, tea). Is the high degree of export specialization in developing countries consistent with comparative advantage?

Third, is China's government pushing the country up the product ladder? China's rapid transition from producing shoes and dolls to laptops and cellphones has created anxiety both in high-income countries that see their competitive advantage in high-tech goods eroding and in emerging economies that fear being left in China's shadow. But along with its changing export specialization, China has increased its supply of educated labor, attracted investment by multinational firms, and improved its transportation and communications infrastructure, making it plausible that its advantage in electronics is natural and not induced by an industrial policy that selectively supports certain industries. Since economists know little about the deep determinants of national export specialization patterns, we have little basis for saying whether or not China's export success is unwarranted.

Finally, what effect has the global commodity boom had on living standards in low-income countries? China's and India's immense demand for raw materials has meant a sharp increase in the terms of trade for commodity exporters, including many countries in Sub-Saharan Africa. Has the commodity export boom reduced absolute poverty or otherwise improved the quality of life in the developing world? Conversely, have low-income countries that are commodity importers seen a decline in their material well-being? Since the 1980s, the World Bank and the IMF have been preaching that trade liberalization is part of the path to prosperity for developing economies. Yet the literature so far (for example, Goldberg and Pavcnik 2007) has not established that lower trade barriers make the poor in those countries better off.

## References

**Amiti, Mary, and Caroline Freund.** 2010. "The Anatomy of China's Export Growth" with a comment by Bin Xu. In *China's Growing Role in World Trade,* edited by Robert Feenstra and Shang-Jin Wei, 35–61. Chicago: NBER and University of Chicago Press.

**Anderson, James E, and Eric van Wincoop.** 2004. "Trade Costs." *Journal of Economic Literature* 42(3): 691–761.

**Autor, David, David Dorn, and Gordon H. Hanson.** 2011. "The China Syndrome: Local Labor Market Effects of Impacts of Import Competition in the United States." http://irps.ucsd.edu/assets/001/502434.pdf.

**Baldwin, Richard E., and James Harrigan.** 2011. "Zeros, Quality, and Space: Trade Theory and Trade Evidence." *American Economic Journal: Microeconomics* 3(2): 60–88.

**Barro, Robert J., and Jong-Wha Lee.** 2010. "A New Data Set of Educational Attainment in the World, 1950–2010." NBER Working Paper 15902.

**Becker, Sascha O., and Marc-Andreas Muendler.** 2010. "Margins of Multinational Labor Substitution." *American Economic Review* 100(5): 1999–2030.

**Bergin, Paul R., Robert C. Feenstra, and Gordon H. Hanson.** 2009. "Offshoring and Volatility: Evidence from Mexico's Maquiladora Industry." *American Economic Review* 99(4): 1664–71.

**Bernard, Andrew B., and J. Bradford Jensen.** 1999. "Exceptional Exporter Performance: Cause, Effect, or Both." *Journal of International Economics* 47(1): 1–25.

**Bernard, Andrew B., J. Bradford Jensen, and Peter K. Schott.** 2006. "Survival of the Best Fit: Exposure to Low-Wage Countries and the (Uneven) Growth of U.S. Manufacturing Plants." *Journal of International Economics* 68(1): 219–37.

**Bernard, Andrew B., Stephen Redding, and Peter K. Schott.** 2007. "Comparative Advantage and Heterogeneous Firms." *Review of Economic Studies* 74(1): 31–66.

**Bloom, Nicholas, Mirko Draca, and John Van Reenen.** 2011. "Trade Induced Technical Change? The Impact of Chinese Imports on Innovation, IT, and Productivity." NBER Working Paper 16717.

**Cadot, Olivier, Céine Carrère, and Vanessa Strauss-Kahn.** 2011. "Export Diversification: What's behind the Hump?" *Review of Economics and Statistics* 93(2): 590–605.

**Chor, Davin.** 2010. "Unpacking Sources of Comparative Advantage: A Quantitative Approach." *Journal of International Economics* 82(2): 152–67.

**Costinot, Arnaud, David Donaldson, and Ivana Komunjer.** 2011. "What Goods Do Countries Trade? A Quantitative Exploration of Ricardo's Ideas." *Review of Economic Studies,* published online September 28.

**Costinot, Arnaud, Jonathan Vogel, and Su Wang.** 2011. "An Elementary Theory of Global Supply Chains." NBER Working Paper 16936.

**Easterly, William, and Ariell Reshef.** 2009. "Big Hits in Manufacturing Exports and Development." http://williameasterly.files.wordpress.com/2010/08/4_easterly_reshef_bighitsinmanufacturingexportsanddevelopment_wp.pdf.

**Eaton, Jonathan, and Samuel Kortum.** 2002. "Technology, Geography, and Trade." *Econometrica* 70(5): 1741–79.

**Eaton, Jonathan, Samuel S. Kortum, and Sebastian Sotelo.** 2012. "International Trade: Linking Micro and Macro." NBER Working Paper 17864.

**Feenstra, Robert C.** 2003. *Advanced International Trade: Theory and Evidence.* Princeton: Princeton University Press.

**Feenstra, Robert C.** 2010. *Offshoring in the Global Economy: Microeconomic Structure and Macroeconomic Implications.* Cambridge, MA: MIT Press.

**Feenstra, Robert C., and Gordon H. Hanson.** 2005. "Ownership and Control in Outsourcing to China: Estimating the Property-Rights Theory of the Firm." *Quarterly Journal of Economics* 120(2): 729–61.

**Goldberg, Pinelopi Koujianou, and Nina Pavcnik.** 2007. "Distributional Effects of Globalization in Developing Countries." *Journal of Economic Literature* 45(1): 39–82.

**Grossman, Gene M., and Esteban Rossi-Hansberg.** 2008. "Trading Tasks: A Simple Theory of Offshoring." *American Economic Review* 98(5): 1978–97.

**Hanson, Gordon H., Raymond J. Mataloni, and Matthew J. Slaughter.** 2005. "Vertical Production Networks in Multinational Firms." *Review of Economics and Statistics* 87(4): 664–78.

**Harrigan, James, and Haiyan Deng.** 2010. "China's Local Comparative Advantage." In *China's Growing Role in World Trade,* edited by Robert C. Feenstra and Shang-jin Wei, 109–136. Chicago: University of Chicago Press.

**Harrison, Ann, and Margaret McMillan.** 2011. "Offshoring Jobs? Multinationals and U.S. Manufacturing Employment." *Review of Economics and Statistics* 93(3): 857–75.

**Helpman, Elhanan, and Paul R. Krugman.** 1985. *Market Structure and Foreign Trade: Increasing Returns, Imperfect Competition, and the International Economy.* Cambridge, MA: MIT Press.

**Helpman, Elhanan, Marc J. Melitz, and Yona Rubinstein.** 2008. "Estimating Trade Flows: Trading Partners and Trading Volumes." *Quarterly Journal of Economics* 123(2): 441–87.

**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 Ralph Ossa.** 2011. "A Global View of Productivity Growth in China." NBER Working Paper 16778.

**Hummels, David.** 2007. "Transportation Costs and International Trade in the Second Era of Globalization." *Journal of Economic Perspectives* 21(3): 131–54.

**Hummels, David, June Ishii, and Kei-Mu Yi.** 2001. "The Nature and Growth of Vertical Specialization in World Trade." *Journal of International Economics* 54(1): 75–96.

**Imbs, Jean, and Romain Wacziarg.** 2003. "Stages of Diversification." *American Economic Review* 93(1): 63–86.

**Johnson, Robert C., and Guillermo Noguera.** 2012. "Accounting for Intermediates: Production

Sharing and Trade in Value Added." *Journal of International Economics* 86(2): 224–36.

**Koopmans, Robert, William Powers, Zhi Wang, and Shang-Jin Wei.** 2010. "Give Credit where Credit Is Due: Tracing Value Added in Global Production Chains." NBER Working Paper 16426.

**Levchenko, Andrei A., and Jing Zhang.** 2011. "The Evolution of Comparative Advantage: Measurement and Welfare Implications." NBER Working Paper 16806.

**Marshal, Alfred.** 1920. *Principals of Economics.* New York: MacMillan.

**Melitz, Marc J.** 2003. "The Impact of Trade on Intra-industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71(6): 1695–1725.

**Rose, Andrew K.** 2004a. "Do We Really Know that the WTO Increases Trade?" *American Economic Review* 94(1): 98–114.

**Rose, Andrew K.** 2004b. "Do WTO Members Have More Liberal Trade Policy?" *Journal of International Economics* 63(2): 209–35.

**Rodrik, Dani.** 2006. "What's So Special about China's Exports?" *China and the World Economy* 14(5): 1–19.

**Rodrik, Dani, and Arvind Subramanian.** 2004. "From Hindu Growth to Productivity Surge: The Mystery of the Indian Growth Transition." NBER Working Paper 10376.

**Santos Silva, J. M. C., and Silvana Tenreyro.** 2006. "The Log of Gravity." *Review of Economics and Statistics* 88(4): 641–58.

**Schott, Peter K.** 2003. "One Size Fits All? Heckscher–Ohlin Specialization in Global Production." *American Economic Review* 93(3): 686–708.

**Schott, Peter K.** 2004. "Across-Product versus Within-Product Specialization in International Trade." *Quarterly Journal of Economics* 119(2): 647–78.

**Schott, Peter K.** 2008. "The Relative Sophistication of Chinese Exports." *Economic Policy* 22(53): 5–49.

**Song, Zheng, Kjetil Storesletten, and Fabrizio Zilibotti.** 2011. "Growing Like China." *American Economic Review* 101(1): 196–233.

**Yi, Kei-Mu.** 2003. "Can Vertical Specialization Explain the Growth of World Trade?" *Journal of Political Economy* 111(1): 52–102.

Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 65–90

# Putting Ricardo to Work[†]

Jonathan Eaton and Samuel Kortum

When presented with the opportunity to trade, countries benefit by specializing in the activities they do relatively better. This finding, the principle of comparative advantage, is one of the first analytic results in economics. While Adam Smith (1776) made a much earlier case for free trade, he based it on increasing returns to scale, and provided no formal demonstration. In contrast, David Ricardo (1817) provided a mathematical example showing that countries could gain from trade by exploiting innate differences in their ability to make different goods.

In the basic Ricardian example, two countries do better by specializing in different goods and exchanging them for each other, even when one country is better at making both. This example typically gets presented in the first or second chapter of a text on international trade, and sometimes appears even in a principles text. The reason is to demonstrate the gains from specialization and trade in a way that at least a bright student can absorb quickly. But having served its pedagogical purpose, the model is rarely heard from again. As one example, Feenstra (2004), the leading Ph.D. text in international trade, devotes only three pages to the Ricardian model. During the twentieth century, the theoretical and quantitative analysis of international trade turned first to differences in factor endowments and then to increasing returns to scale as explanations for trade and its benefits. The Ricardian model became something like a family heirloom, brought down from the attic to show a new generation of students, and then put back, allowing them to pursue more fruitful lines of study and research.

■ *Jonathan Eaton is Professor of Economics, Pennsylvania State University, University Park, Pennsylvania. Samuel Kortum is Professor of Economics, University of Chicago, Chicago, Illinois. Their e-mail addresses are ⟨jxe22@psu.edu⟩ and ⟨kortum@uchicago.edu⟩.*

[†] To access the Appendix, visit
http://dx.doi.org/10.1257/jep.26.2.65.

doi=10.1257/jep.26.2.65

Nearly two centuries later, however, the Ricardian framework has experienced a revival. Much work in international trade during the last decade has returned to the assumption that countries gain from trade because they have access to different technologies. These technologies may be generally available to producers in a country, as in the Ricardian model of trade, our topic here, or exclusive to individual firms, as Marc Melitz and Daniel Trefler discusses in the companion paper in this issue. This line of thought has brought Ricardo's theory of comparative advantage back to center stage. Our goal is to make this new old trade theory accessible and to put it to work on some current issues in the international economy.

## Revisiting Ricardo's Example

Ricardo (1817) posited a world of two countries, England and Portugal, which can make each of two goods, cloth and wine. What he assumed about how many workers it takes to make a unit of each good in each country appears in Table 1. Since the workers required to make one unit of a good are the same no matter how many units are produced, Ricardo was assuming constant returns to scale.

Ricardo argued that trade could allow England to obtain a unit of wine with the effort of only 100 workers (instead of 120) and Portugal to obtain a unit of cloth with the effort of only 80 workers (instead of 90)—the outcome if international trade established an international price of 1 unit of cloth exchanging for 1 unit of wine.

Of course, to our twenty-first century eyes, Ricardo's example is very incomplete. For example, he does not explain what assumptions about tastes, endowments, or competition are needed for this world price ratio of 1 to arise. However, in using this example Ricardo was advocating policy in a very modern way. He compared an actual world with one policy—trade prohibited—with a counterfactual world of free trade. In making the comparison, he described each world in terms of a common set of parameters, the labor requirements in Table 1, that are plausibly exogenous to the policy in question, thus immunizing himself to the Lucas critique (1976) of the following century.[1]

Why, when the Ricardian model delivers such a slick demonstration of the gains from trade, did it hit such a dead end in terms of providing a framework for more sophisticated and quantitatively meaningful analysis? A major reason is that even this basic formulation gives rise to different types of equilibria that need to be analyzed separately. Even in Ricardo's minimalist setting, three types of outcomes are possible: 1) England makes only cloth and Portugal only wine, 2) England makes both cloth and wine and Portugal only wine, or 3) England makes only cloth and Portugal both

[1] Chipman (1965), in his magnificent three-part survey of the theory of international trade, attributes the first complete statement of a Ricardian equilibrium to Mill (1844), who implicitly assumed what we now call Cobb–Douglas preferences, with equal shares for each good. Using the labor requirements in Table 1, the reader can verify that Ricardo's posited price of 1 will then emerge if Portugal has 80 percent as many workers as England.

*Table 1*
**Ricardo's Example: How Many Workers to Make a Unit of a Good**

| | *Cloth* | *Wine* |
|---|---|---|
| England | 100 | 120 |
| Portugal | 90 | 80 |

cloth and wine. In case 1, the one assumed in Ricardo's example, outputs can be immediately solved for from labor endowments, with prices then determined by demand. In the second two, relative prices are given by the relative labor requirements in the incompletely specialized country, with demand then determining outputs. At the intro level, the lesson from his sort of example is that gains from trade are possible, although we can only put bounds on what the gains are. At a more advanced level, students are told to solve for the equilibrium outcome by assuming one case and then checking that it satisfies the requirement that prices don't exceed costs or that labor is fully employed. Already the model has to confront a clumsy taxonomy.

International trade is a field rich in data. United Nations COMTRADE, currently the major source of statistics on merchandise trade, reports the annual value of bilateral trade between over 242 countries (making for $242 \times 241 = 58{,}322$ bilateral pairs) in 776 product categories going back to 1990. Given that even the two-country, two-good example is awkward to work out, what hope does the Ricardian model have of sorting out data of this complexity?

In fact, a handful of developments have recently culminated in a formulation of the Ricardian model that is highly amenable to exploiting exactly such data. This formulation has spawned a surge of studies to address various policy questions quantitatively. We chart this evolution and show where it has led.

## Ricardian Trade Theory: From Textbook Example to Practical Tool

To begin, let's reformulate Ricardo's example in terms of England's wage $\omega$ relative to Portugal's, setting the Portuguese wage to 1. Making a unit of cloth in England will then cost $100\omega$, while making it in Portugal will cost 90. Making a unit of wine in England will cost $120\omega$, while making it in Portugal will cost 80. With free trade and perfect competition, the prices of cloth and of wine are the same in each location and constitute the lowest-cost way of producing each good. Say that $\omega$ is bigger than 90/100, the ratio of Portuguese to English workers required to make cloth. Then, since

$$\underset{\text{(cloth)}}{\frac{90}{100}} > \underset{\text{(wine)}}{\frac{80}{120}}$$

both cloth and wine will be produced more cheaply in Portugal, leaving English labor out of work. Hence an English wage that is more than 90 percent of the Portuguese wage is not compatible with employment in England. At the other extreme, if $\omega$ is smaller than 80/120, then both cloth and wine will be cheaper if made in England, putting Portuguese labor out of work. Hence we need $\omega$ to be somewhere in between 2/3 and 9/10. (Because Ricardo granted Portugal an absolute advantage in both goods, he doomed English workers to a lower wage in order to be employed.) The idea that a Ricardian equilibrium involves identifying the source that can supply a good at minimum cost is at the heart of taking the model to more goods and countries.

Any hope of applying this example to actual world trade requires adding more goods and countries. How can we do that? Let's proceed step by step.

### More Goods

Let's add another good, linen, while sticking with just our two countries. Say England needs 100 workers to make a unit of linen, and Portugal needs 100 workers as well. These numbers grant England an even stronger comparative advantage in linen than in cloth. We can extend the previous inequality to:

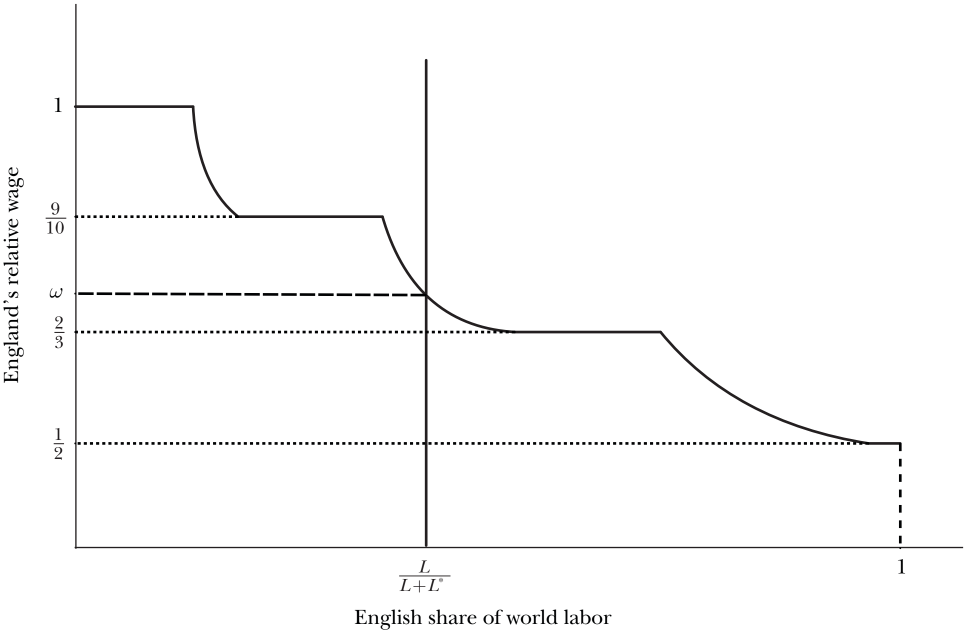
$$\underset{\text{(linen)}}{\frac{100}{100}} > \underset{\text{(cloth)}}{\frac{90}{100}} > \underset{\text{(wine)}}{\frac{80}{120}}.$$

This ordering of goods in terms of England's relative productivity is called a chain of comparative advantage. Under free trade, the English relative wage $\omega$ breaks this chain between goods for which England's relative productivity is above or below its relative wage. The goods to the left of the break are produced more cheaply in England and those to the right of the break are produced more cheaply in Portugal. For example, an $\omega$ of .95 breaks the chain between linen (produced more cheaply in England) and cloth and wine (cheaper from Portugal). An $\omega$ of .9 breaks it at cloth (costing the same from either country, with linen cheaper from England and wine cheaper from Portugal).

What determines the relative wage $\omega$ that breaks the chain? In general, finding it can be quite complicated but, if the two countries spend their income the same way (specifically, if tastes are identical and homothetic), the problem simplifies. We can then use the chain of comparative advantage to construct the demand curve for English labor relative to world labor (on the $x$-axis) as it varies with the English wage $\omega$ (on the $y$-axis).

If $\omega > 1$, then English labor has priced itself out of all goods. Hence, the demand curve is just a vertical line at zero for $\omega$ above England's relative productivity for good 1. At a wage $\omega = 1$, England is competitive in linen, and buyers are indifferent between England and Portugal as a source. The demand curve for English labor is then flat (perfectly elastic) between zero and the point at which the demand for linen is saturated at the price of 100. A decline in $\omega$ from this point renders England the sole producer of linen. Since the price of linen is $100\omega$, a drop in $\omega$ lowers the price

*Figure 1*
**Wage Determination in the Many Good Model**



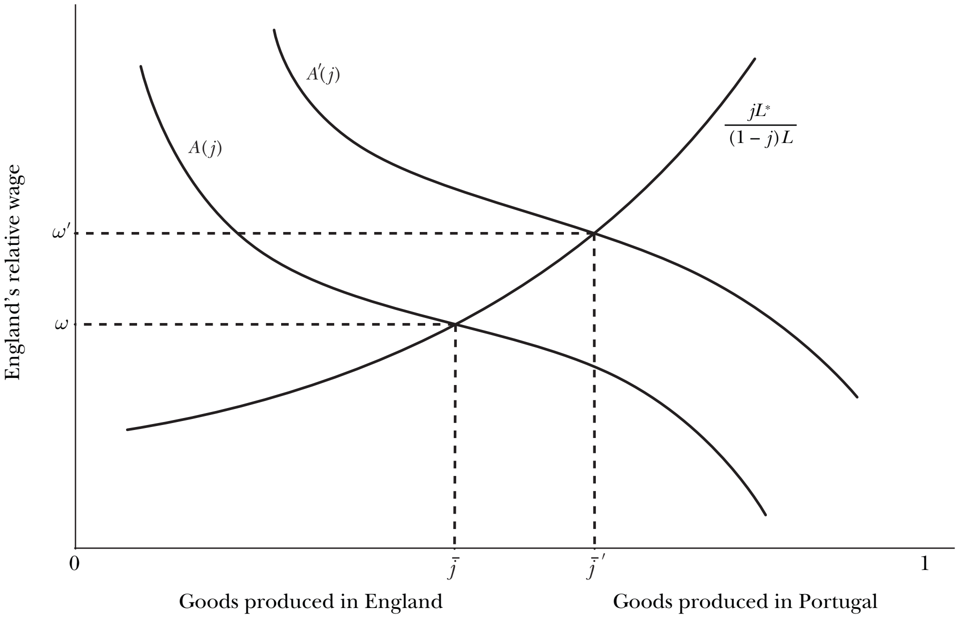
*Source:* Authors.
*Note:* The solid downward-sloping line is the relative demand curve for English labor, and the solid vertical line is the relative supply curve for English labor.

of linen, increasing demand for it and hence for English labor. At the point $\omega = .9$, England becomes competitive in cloth as well as linen. The demand curve for English labor thus hits another flat zone as world buyers are indifferent between England and Portugal as sources of cloth (continuing to buy all their linen from England and wine from Portugal). Proceeding along the chain, the demand curve for English labor is a downward stairway with treads along which England and Portugal share production of a good connected by risers along which England and Portugal specialize in producing distinct sets of goods. The treads are horizontal, as with a standard staircase, but the risers are vertical only in an extreme case. Otherwise they slope downward to the next tread. The equilibrium can be found by imposing the vertical supply curve for English labor as a share of the world's, which could cut the demand curve along a tread (corresponding to a good for which England and Portugal share production) or through a riser (with no shared goods).

We count five possible types of outcomes, going from linen, cloth, and wine made in England and wine elsewhere, to linen, cloth, and wine made in Portugal and linen elsewhere. Of course, more goods can be added by inserting them into the chain, raising the number of types of outcomes.

Figure 1 illustrates the case for four goods, adding one product to the example above—say, anchovies—for which England requires twice as many workers as

*Figure 2*
**Wage Determination with a Continuum of Goods**



*Source:* Authors.
*Notes:* On the *x*-axis is a continuum of goods from 0 to 1 with England having the strongest comparative advantage in goods nearer 0 and Portugal in goods nearer 1. England produces the goods from 0 to $\bar{j}$. Portugal produces the goods from $\bar{j}$ through 1. The figure illustrates how a shift up in the productivity curve $A(j)$, meaning that England gets relatively more productive at making every good, raises England's relative wage $\omega$ and expands the share of goods it produces. A partial derivation for the equation describing the upward-sloping curve is provided in footnote 2.

Portugal to produce a unit. Changing the English labor supply involves sliding the English relative labor supply curve $L/(L + L^*)$ along the *x*-axis where $L$ is English labor and $L^*$ Portugal's.

Trade economists now speak frequently of the extensive and intensive margins of trade. A country's exports can increase on the intensive margin, exporting more of a given set of goods, or on the extensive margin, exporting a wider range of goods. The stairway shows how the two operate in a Ricardian framework. Along a riser, a drop in $\omega$ raises demand for English exports only at the intensive margin, by lowering the price of the given set of goods that England produces. When $\omega$ hits a tread, however, expansion is also at the extensive margin as England expands the set of goods it produces and exports.

An implication of the framework is that, given technologies around the world, having a larger share of the world labor force may require a country to have a lower wage. In order to employ more labor with its given set of technologies, a country needs to sell more of the goods it currently produces (going down a riser) or to take over goods from other countries (reaching a lower step). The result holds even though technologies are constant returns to scale, because larger size reduces the gains from trade. This basic implication of the Ricardian model will survive its modern reincarnation.

While the construct is intuitive, stairways are trouble not only for wheeled vehicles but for comparative statics. Solving for the equilibrium is tedious.

### More Goods than You Can Count

A classic paper by Dornbusch, Fischer, and Samuelson (1977) made life much simpler by replacing the stairway with a ramp. These authors had the insight that inserting more and more goods into the chain of comparative advantage would render the gaps between the ratios of the labor requirements miniscule, in which case the three types of equilibria around any good in the original model collapse to the same outcome. They assumed that the set of goods correspond to all the points on an interval between 0 and 1, and sorted the goods to form a chain of comparative advantage, with England having the strongest comparative advantage in goods closest to zero and Portugal in goods closest to one. They defined a function $A(j)$ as the ratio of Portugal's labor requirements to England's labor requirements for good $j$, hence England's relative productivity, for each $j$ between 0 and 1. They went on to assume that $A(j)$ was smooth and strictly decreasing. The downward sloping curve in Figure 2 illustrates such a function.

For any English wage $\omega$ between $A(0)$ and $A(1)$ there is some good, let's call it $\bar{j}$, satisfying $\mathrm{A}(\bar{j}) = \omega$. This good $\bar{j}$ costs the same whether it is produced in England or Portugal. England produces goods $j \leq \bar{j}$, Portugal goods $j \geq \bar{j}$. Who produces good $\bar{j}$ is irrelevant to anything else, because this good is only an infinitesimal fraction of the total. Because $j$ goes from 0 to 1, $\bar{j}$ is also the share of goods produced in England, for consumption in either England or Portugal. Because $A(j)$ is decreasing, a change that increases England's relative wage $\omega$, given the function $A(j)$, must reduce the share of goods produced in England.

To figure out what $\omega$ will break the chain, we need to look at the demand side. A higher $\bar{j}$ means that England is producing a larger share of goods, increasing demand for its labor and hence its wage $\omega$. Figure 2 depicts this positive relationship between $\omega$ and $\bar{j}$. Where it intersects the downward sloping $A(j)$ curve determines the equilibrium.[2]

[2] To get an exact expression for this upward-sloping relationship requires us to say something about tastes. The simplest assumption is that individuals in either country spread their spending evenly across the goods (as with symmetric Cobb–Douglas preferences). In this case the share of goods produced in England becomes the share of spending devoted to goods produced in England. Labor market equilibrium requires full employment of workers in England and in Portugal at a relative wage $\omega$, with English workers paid a fraction $\bar{j}$ of world income, which is just the wage income in each country added together:

$$\omega L = \bar{j}(\omega L + L^*).$$

A lower English wage $\omega$ increases demand for English labor in two ways: At the intensive margin, a lower $\omega$ lowers the price of all goods England makes, so increases demand for them and thus for English workers. At the extensive margin, a lower $\omega$ increases the range of goods that England exports.

Figure 2 illustrates how a shift up in the productivity curve $A(j)$, meaning that England gets relatively more productive at making every good, raises England's relative wage $\omega$ and expands the share of goods it produces.

In all of the examples so far, if England and Portugal spend their incomes the same way (again, meaning identical, homothetic preferences) there is no reason for English and Portuguese to consume goods in different proportions. But a robust feature of data on trade and production is that countries tend to buy more goods from themselves. We could explain this fact in terms of the basic Ricardian model by assuming that Portuguese like wine more than the English. But it would be coincidental if tastes always happened to align with comparative advantage, and there is little evidence that they do.

A more plausible explanation is that moving goods between countries is costly. Another useful contribution of Dornbusch, Fischer, and Samuelson (1977) is to introduce trade costs into their Ricardian model. Specifically, they make Samuelson's classic iceberg assumption that delivering one unit of any good from one country to the other requires shipping $d$ units, where $d \geq 1$. The specification is consistent with a fraction of the goods getting lost, rotten, or broken in shipment, but admits many other interpretations as well.

Because of iceberg trade barriers, goods no longer cost the same in each location. Consider the case of cloth in Ricardo's example. If the wage in England is .8, then cloth costs 80 if made in England and 90 if made in Portugal. But say that one-third of the cloth shipped from England to Portugal is ruined by saltwater in transport. Then 1.5 units of cloth need to be shipped to deliver 1 usable unit to Portugal, raising the cost of English cloth in Portugal to 120. It no longer pays for Portugal to import cloth from England rather than make it at home.

What happens to the Dornbusch, Fischer, and Samuelson (1977) model if we introduce a trade cost $d$ to all goods? The trade cost creates a range of goods that are not traded as each country makes them more cheaply for itself. As long as $d$ is not too big, there is still a range of goods (with $j$ near zero) that England makes for everyone and another range (with $j$ near one) that Portugal makes for everyone.

An important implication of the trade cost, which we exploit in our applications below, is that it introduces a relationship between any trade deficit that England runs with Portugal and its relative wage. A transfer from England to Portugal diverts spending away from the nontraded goods that England was producing for itself toward the production of those same goods in Portugal. As a consequence, the English wage falls, leading to an expansion of the range of goods that England exports and a contraction of the range that Portugal exports.

The work of Dornbusch, Fischer, and Samuelson (1977) moved the Ricardian framework far forward from being a toy example to becoming a tool that can address a variety of questions. For example, Matsuyama (2008) uses variants of the model to examine the consequences of country size, technological change, and technology transfer on the gains from trade and the distribution of income. But a limitation remains. There are still only two countries.

**More Countries**

It's just as straightforward to add more countries to Ricardo's example as more goods. Let's add a third country, France, with labor requirements 120 in cloth and 60 in wine. Begin by rewriting Ricardo's earlier inequality as

$$\underset{\text{(England)}}{\frac{120}{100}} > \underset{\text{(Portugal)}}{\frac{80}{90}}$$

and then insert France into the chain to get:

$$\underset{\text{(England)}}{\frac{120}{100}} > \underset{\text{(Portugal)}}{\frac{80}{90}} > \underset{\text{(France)}}{\frac{60}{120}}.$$

England, at one end of the chain, will produce cloth and France, at the other end, will produce wine. As before, tastes and the sizes of the labor forces in each country will determine where the chain is broken. As above, we count five types of possible outcomes. We are back to a stairway. More countries can be added, but the number of cases expands. As with two countries and many goods, finding the solution is relatively straightforward but tedious.

**The Challenge: Many Goods with Many Countries**

What about more goods *and* more countries? In this setting, chains no longer work. Jones (1961) provides an example with the following labor requirements for three countries in three goods:

| | *America* | *Britain* | *Europe* |
|---|---|---|---|
| Corn | 10 | 10 | 10 |
| Linen | 5 | 7 | 3 |
| Cloth | 4 | 3 | 2 |

Two assignments—(America, Linen; Britain, Corn; Europe, Cloth) and (America, Corn; Britain, Cloth; Europe, Linen)—each satisfy Ricardo's inequality for any two countries and any two goods looked at in isolation from the third.[3] But only the

[3] Graham (1948) solved for competitive equilibria in numerical examples of the Ricardian model with many countries and many goods. His generalizations from these examples were not always correct. McKenzie (1954) formalized Graham's model and used it in his demonstration of the existence and uniqueness of a competitive equilibrium. In this journal, Weintraub (2011) provides a detailed account of McKenzie's relatively unheralded contribution. McKenzie (1953) established the equivalence between an efficient solution and a competitive equilibrium in Graham's model, and pointed to the inadequacy of bilateral comparisons in determining efficient specialization. The contribution of Jones (1961) is to obtain a simple characterization of efficient specialization in this model.

second is a possible competitive equilibrium. To see why, cross multiply Ricardo's earlier inequality so that it appears a third way as:

$$\underset{\begin{pmatrix}\text{incorrect}\\\text{assignment}\end{pmatrix}}{120 \times 90} > \underset{\begin{pmatrix}\text{correct}\\\text{assignment}\end{pmatrix}}{100 \times 80}.$$

Note that England producing cloth and Portugal wine, the equilibrium assignment in Ricardo's example, minimizes the product of the labor requirements for the technologies used. Generalizing this result, Jones can rule out the first assignment in his example since it involves a higher value for the product of the labor requirements used ($5 \times 10 \times 2 = 100$ versus $10 \times 3 \times 3 = 90$).[4]

Fun as this example is, it doesn't provide much guidance into how to solve for the equilibrium in high-dimensional cases. For one thing, we're still left with the problem of figuring out if the solution is on a tread or a riser. But now we have stairways running in multiple directions in ways that only M. C. Escher could diagram.

### A Solution: Distributions of Worker Requirements

Again, we need a ramp. To construct one, let's return to the Dornbusch, Fischer, and Samuelson (1977) formulation with a continuum of goods, but now allow for an arbitrary (integer) number $I$ of countries. We must deal with unit labor requirements for each good (one for each point on the unit interval) in each country (of which there are $I$), vastly more numbers than Ricardo's four.

To tackle the problem, let's first give up on actual numbers and, following Dornbusch, Samuelson, and Fisher (1977), label the labor requirement for good $j$ in country $i$ by $a_i(j)$. With $I > 2$ countries and lots of goods, it doesn't help to think about ratios of the $a$'s, so chains are out of the question. Instead, we will think about the $a$'s as the realizations of random variables drawn from a particular family of probability

[4] This idea generalizes to the $I$-good, $I$-country case. To see why this rule works it helps to go back to prices and to think about finding the minimum cost source. Let's index countries by $i = 1,\ldots, I$ and goods by $j = 1,\ldots, I$ and denote the amount of labor needed to make good $j$ in country $i$ as $a_i(j)$. Let $w_i$ be the wage in country $i$ and $p(j)$ the world price of good $j$ (as there are no transport costs). Let's also number countries and goods so good $j$ is produced by country $i = j$ in an efficient outcome (so that we can label by $j$ the country producing good $j$ under the correct assignment). Perfect competition then means that $p(j) = a_j(j)w_j$ (zero profits where good $j$ is produced) and $p(j) \leq a_i(j)w_i$ for all other countries $i$ (no profit opportunities anywhere else). Multiplying the equalities together for the correct assignment gives:

$$\prod_{j=1}^{I} p(j) = \prod_{j=1}^{I} a_j(j) \prod_{j=1}^{I} w_j.$$

Multiplying together the inequalities for any other one-to-one assignment $i(j)$ of country $i$ to good $j$ gives:

$$\prod_{j=1}^{I} p(j) \leq \prod_{j=1}^{I} [a_{i(j)}(j)w_{i(j)}] = \prod_{j=1}^{I} a_{i(j)}(j) \prod_{j=1}^{I} w_{i(j)}.$$

Since the terms $\prod_{j=1}^{I} p(j)$ and $\prod_{j=1}^{I} w_{i(j)} = \prod_{j=1}^{I} w_j$ are the same in each, the only way both expressions can be true is if

$$\prod_{j=1}^{I} a_j(j) \leq \prod_{j=1}^{I} a_{i(j)}(j).$$

distributions. This way of thinking about technology (the labor requirements to produce different goods in different locations) has two advantages: First, the distributions themselves can be smooth, giving us our ramp. Second, we don't have to keep track of all the individual $a_i(j)$'s, of which there are many, but only the parameters of the distributions from which they are drawn, which can be small in number.

Before getting into the details, it's useful to step back and articulate some principles that guide the choice of a family of distributions. First, we want to stay within the family when we move from the distribution of labor requirements to the distribution of the costs of producing goods. Second, we want to stay within the family when we consider the distribution of the price of a good in a country, which is the minimum of the cost of acquiring it across all potential source countries. Finally, we want a simple expression for the probability that a particular country is the low-cost source.

These considerations led us, in Eaton and Kortum (2002), to a family of what are called extreme-value distributions. The well-known central limit theorem states that if a large sample is taken from a well-behaved distribution, then the mean of the sample has an approximate normal distribution. Less well-known is that the highest or lowest observations in such a sample also can approach a particular distribution, called an extreme-value distribution. For example, consider the winning (fastest) times in a series of races. If each runner's time is drawn from a particular distribution, such as the lognormal, then the fastest time across a large number of races has an extreme value distribution, which, if the times are lognormal, turns out to be the type-III extreme value, or Weibull distribution.

What's the connection between winning a race and the number of workers needed to make a product? As derived in Kortum (1997) and Eaton and Kortum (1999), if technologies for making a good are the results of inventions that occur over time, and if the output per worker delivered by an invention is drawn from the Pareto distribution, then output per worker using the most efficient (that is, winning) technology discovered to date have a type-II or Fréchet distribution.

The Ricardian language describes a technology by its worker requirement rather than by its reciprocal, output per worker. Translating our results on the Fréchet distribution above into Ricardian, the probability that the labor requirement for producing any particular good $j$ in country $i$ is less than any positive number $x$ forms a Weibull distribution, specifically:

$$\Pr[a_i(j) \leq x] = 1 - e^{-(A_i x)^\theta}.$$

Its two parameters relate to absolute and comparative advantage. The parameter $A_i$ captures country $i$'s absolute advantage: A higher value means that the labor requirement is likely to be lower for any good. Having absolute advantage vary across countries allows us to capture the fact that some countries are much more productive than others across a wide range of activities: for example, in the way Portugal is more productive than England across both goods in Ricardo's example. A country that has accumulated more technology will have a higher $A_i$.

The parameter $\theta$ captures (inversely) how variable the labor requirement is, with a higher value meaning that a country's labor requirement is typically close to its mean, weakening the force of comparative advantage. In Ricardo's example above, suppose Portugal could make cloth with 67 workers rather than with 90. While Portugal would still be better at both goods than England, it's no longer differentially much better at wine. As Ricardo's inequality gets closer to equality, the scope for gains from trade decreases. Similarly, a high value of $\theta$ in our model reduces the gains from trade. Imposing a common $\theta$ across countries makes it easy for us to see how technologies around the world interact through trade.

The extreme value distribution is convenient, but how well does it reflect reality? As described above, a way of generating this distribution is to draw worker efficiencies repeatedly from a Pareto distribution, taking the largest. The upper tail of the distribution, representing the most efficient firms, itself resembles a Pareto distribution. Wilfredo Pareto invented what we now call the Pareto distribution to describe how income was distributed. It turns out that the Pareto distribution, sometimes called a "power law," describes the upper tail of a large number of magnitudes, such as city population and firm sales and employment. Hence the extreme value distribution fits the data quite well.

Since we now have $I$ countries, iceberg trade costs can now vary with the pair of countries in question, so that delivering a unit of a good to country $n$ requires shipping $d_{ni} \geq 1$ units from country $i$ (with $d_{ii} = 1$). These trade costs can capture a well-known regularity in data on trade, which is that the amount of trade between two countries tends to fall as the distance between them rises. This feature is known as "gravity," and gravity models of trade build on this insight. The multi-country framework developed here will display gravity if iceberg costs between any two countries rise systematically with the distance between them. Here, although we incorporate iceberg costs, we steer away from giving them too specific an interpretation. The issue of how well iceberg costs capture reality remains subject to debate: see Anderson and van Wincoop (2004) for further discussion.

Putting all these ingredients together, the cost of producing a good $j$ in country $i$ and delivering it to country $n$ is $c_{ni}(j) = a_i(j) w_i d_{ni}$, the product of the labor requirement in country $i$, the wage in $i$, and the iceberg cost of moving goods from $i$ to $n$. As in Dornbusch, Fischer, and Samuelson (1977), wages and trade costs are the same for all goods produced in a country, and so $c_{ni}(j)$ has the same distribution as $a_i(j)$, only with the absolute advantage parameter $A_i$ replaced by $A_{ni} = A_i/(w_i d_{ni})$. The positive effect of raw efficiency in country $i$ (through a higher $A_i$) on the cost distribution in $n$ is offset by a higher wage and a higher cost of shipping to country $n$.

Just as in the basic Ricardian model, perfect competition guarantees that the price $p_n(j)$ of good $j$ in country $n$ is the lowest cost $c_{ni}(j)$ looking across all potential sources $i$. Unlike the simple Ricardian model with no trade costs, in the more general set-up here, which country $I$ provides the good at lowest cost may differ across destinations $n$. We already saw such an outcome in the two-country Dornbusch, Fischer, and Samuelson (1977) model with trade costs: Each country produced a range of goods for itself while other goods, for which differences in productivity were more extreme, were produced in only one country. While the multicountry formulation here is more complicated, the distribution of the price of a good $j$ in country $n$ is straightforward. It inherits the extreme value distribution from the costs $c_{ni}(j)$ of which $p_n(j)$ is the minimum across all potential sources $I$, with its distribution remaining in the Weibull family.[5]

Aside from telling us about prices, the model can also tell us about trade between any two countries via the probability $\pi_{ni}$ that a particular country $i$ is the lowest cost source of a good in country $n$. This probability is lower the higher $d_{ni}$, the trade barrier in shipping from $i$ to $n$, and the higher the wage in the source country, adjusted for absolute advantage. Since there are a continuum of goods, the probability $\pi_{ni}$ is also the share of goods in country $n$ supplied by country $i$. Furthermore, with symmetric Cobb–Douglas preferences, the $\pi_{ni}$'s also correspond to the fraction of country $n$'s spending devoted to goods bought from country $i$. These purchases are imports if $i$ and $n$ are different, but are domestic sales when $i$ and $n$ are the same.[6] Because data on the value of trade and production are readily available to calculate trade shares, the $\pi_{ni}$'s provide a crucial link between the model and data.

Anything that lowers a country's cost of serving a market (such as a lower tariff) means more purchases are shifted there; how much depends on $\theta$. Remember that a larger $\theta$ means that technologies are more similar across goods from any given country. Hence a given change in costs implies a bigger shift in trade shares when $\theta$ is high, since relative costs don't vary that much across countries.

Trade economists have long sought to measure the elasticity of trade with respect to relative costs, which are affected by such things as changes in tariffs or exchange rates. In our analysis, $\theta$ determines that elasticity. It plays an important role in all that follows. In our numerical analysis below, we use a value of $\theta = 4$ suggested in a recent paper by Simonovska and Waugh (2011). Their recommended value is based on a careful analysis of the prices of 62 manufactured goods across 123 countries, and the estimate is in line with several earlier studies based on other evidence.

How does trade translate into welfare in this framework? The model delivers a handy expression for the real wage in country $I$, which is proportional to $A_i \pi_{ii}^{-1/\theta}$.

[5] In particular, the distribution of prices $p_n(j)$ emerges just by replacing $A_i$ in the distribution of labor requirements with a term $\overline{A}_n$, that aggregates the $A_{ni}$'s from each source $i$:

$$(\overline{A}_n)^{\theta} = \sum_{i=1}^{I} (A_{ni})^{\theta}.$$

The expression for $\overline{A}_n$ shows how higher efficiency, lower wages, and greater proximity of country $n$'s trading partners translates into lower prices.

[6] The trade share turns out to be country $i$'s contribution to the term $\overline{A}_n$ given in the previous footnote: $\pi_{ni} = (A_{ni}/\overline{A}_n)^{\theta}$. With Cobb–Douglas preferences, the ideal price index $p_n$ in country $n$ is the geometric mean of the price distribution, which is simply $\gamma/\overline{A}_n$. The constant $\gamma$ is given as equation (5) in the online appendix available with this paper at ⟨http://e-jep.org⟩. We could be much more general in our specification of preferences, but for our analysis here nothing would be gained. For example, with Dixit–Stiglitz preferences, the only change is in the formula for $\gamma$, which then depends on the elasticity of substitution.

The absolute advantage parameter $A_i$ captures labor productivity in the country. In a closed economy, with $\pi_{ii} = 1$, productivity by itself determines the real wage. The second term $\pi_{ii}^{-1/\theta}$ captures the gains from trade.[7] A country $i$ with a small home share $\pi_{ii}$ makes use (via imports) of technologies from elsewhere for a large range of goods. Without trade, of course, it would be using its own technologies to make these goods. How much a given drop in the home share in moving from autarky to trade raises welfare depends on how different the technologies embodied in imports are from the domestic technologies they replace. The smaller is $\theta$, the bigger the difference, on average, and hence the larger the gains. Hence a country without many advanced technologies itself may nevertheless have a high living standard because it specializes in the technologies in which it is most advanced and purchases the rest from abroad. Using our value of $\theta = 4$, we can infer that a country importing 25 percent of what it consumes from abroad, hence purchasing 75 percent locally, gains about 7.5 percent in real income.

While it's very useful to infer the gains from trade by knowing just the home share in expenditure and the parameter $\theta$, the home share itself depends on wages around the world, which are determined by the labor market equilibrium in each country. In order to solve for wages, we need to know not only the trade costs but the labor endowment $L_i$ in each country and the trade deficit $D_i$ each country runs with the rest of the world.

In general, we can't solve the system of labor market equilibrium conditions for wages analytically, although a computer can spit out the answers rapidly, even with several hundred countries. But with costless trade (that is, no iceberg costs), we can obtain an analytic solution. In this special case, the relative wage between two countries is increasing in the ratio of their productivities (their $A_i$'s), with an elasticity of $\theta/(1 + \theta)$. That this elasticity is below one reflects the fact that in an open economy a country passes on some of the benefits of its own higher productivity to others through lower export prices. In this way, even without international technology diffusion, international trade allows countries to benefit from having trading partners with a high level of technology. The relative wage is decreasing in the ratio of labor endowments (the $L_i$'s), with an elasticity of $-1/(1 + \theta)$. This elasticity is negative, just as in the basic Ricardian model. A country with more workers, in order to employ them, produces more of an existing set of goods (the intensive margin), lowering their relative price. Here, in addition, the country diversifies into additional goods (expanding at the extensive margin) in which its relative productivity is lower. Without trade barriers, the relative wage is independent of trade deficits, just as deficits don't matter for wages in Dornbusch, Fischer, and Samuelson (1977) when there are no trade costs.

In a world of costless trade, we can see how countries' endowments of labor and technology interact to determine their relative welfare. Recognizing that distance matters introduces location as a third major determinant of a country's relative

[7] Arkolakis, Costinot, and Rodríguez-Clare (2012) show how, with $\theta$ suitably reinterpreted, this result on the gains from trade generalizes to a wide class of models.

income and welfare. Proximity to large markets and to inexpensive sources of goods then becomes another important feature of a country in determining its welfare.

To get some sense of the magnitude of geography's role in a country's well-being let's perform a numerical exercise with just two countries. Say that one country is large, with 99 percent of the world's labor, and the other small, with 1 percent of the world's labor. Let's start by assuming free trade and labor efficiencies such that with no trade barriers the two countries have the same wage (and hence the same real wage since prices are the same). In a frictionless world with no trade barriers, the small country would spend only 1 percent of its income on goods from itself.

Now imagine introducing a trade barrier between the two countries, so that the iceberg costs are $d = 2$ for sending goods in either direction. In the resulting equilibrium, the small country spends just under half of its income on goods from itself (a typical amount for an actual small country). While the large country is virtually unaffected by the change, the real wage in the small country falls to 38 percent of that in the large country. This decline is the result of two effects. First, to be competitive in the large country, the small country's wage has to fall to 65 percent of the large country's wage. Second, because goods from the large country are expensive to import, the price index is 70 percent higher.

With these trade barriers in place, how much of a productivity boost would we have to give the small country to bring its real wage back up to the level in the rich one? The answer is so much that under costless trade its wage would be more than double the large country's. An implication of this example is that, by influencing trade costs, geography can play as important a role in determining income differences as technology.

## Applying the Tool

Having shown how the Ricardian model can accommodate a complex world of many goods and many countries separated by trade barriers, we now connect it to data. We can then use it to ask many questions both about the world as it is and what it would look like under different circumstances. In this section, we investigate four particular questions: 1) How much do countries gain from trade, and how have these gains evolved over the last two decades? 2) How much will these gains grow if falling trade costs lead to further increases in world trade? 3) To what extent do countries benefit from the technological improvements of their trading partners? 4) What are the costs to deficit countries of moving to balanced trade?

We fit the model to data on 32 countries (31 actual countries and a "rest of the world" which combines all the others) as listed later in Table 3. The limit on the number of countries arises from the availability of data; adding countries adds little to computational complexity.

While any model is a simplification, we can bring the model we have been discussing here much closer to reality with three embellishments: First, the model applies quite naturally to manufactures, the dominant component of trade for most high-income countries. Indeed, manufactures make up 64 percent of trade in goods and services among our 31 actual countries. It is less clear how well this model applies to services or to products in which natural resources play a major role. To focus on trade in manufactures, we follow Alvarez and Lucas (2007) and divide the economy into two sectors, which we call manufacturing and services, with labor mobile between them. Among our set of countries, manufactures represent only a share $\alpha$ of about 0.2 of final spending.

Second, while manufactures are not a large share of final spending, a great deal of manufacturing output goes into the production of manufactures. Among our countries, the share $\beta$ of labor in manufacturing production is only about 0.3 with most of the rest manufactured intermediates. As pointed out, for example, by Krugman and Venables (1995), recognizing the importance of manufactures as inputs makes location as well as geography an important determinant of manufacturing costs. In Dekle, Eaton, and Kortum (2007), we describe in more detail how we set $\alpha$ and $\beta$.

Third, the textbook Ricardian model typically assumes that trade is balanced. However, we design our model to accommodate deficits both in manufacturing and in everything else. In fact, one of our exercises is to examine the consequence of shifting these deficits in order to balance each country's current account. To finish putting numbers on the model we go to the OECD STAN (STructural ANalysis) Database for data on bilateral trade and production of manufactures and to the Economist Intelligence Unit for data on unilateral trade in goods and services, GDP, and the current account. Along with our values for the three parameters $\theta = 4$, $\alpha = 0.2$, and $\beta = 0.3$, these data tell us all that we need to know about the rest of parameters in order to answer our four questions. We can answer our first question, about the magnitude of gains from trade, directly from data by using a relationship discussed in the previous section. The other three questions force us to consider all of the shifts in wages and prices around the world that would result from a change in trade costs, technology, or trade deficits.

We calculate counterfactuals in the following way: We shock the model by changing the relevant parameters. We denote the new level of a variable or parameter $x$ as $x'$ and the proportional change in it as $\hat{x} = x'/x$. In particular, we consider changes $\hat{d}_{ni}$ in trade costs (keeping them at one when $i = n$), changes $\hat{A}_i$ in technology, or counterfactual deficits $D_n'$ (and $D_i^{M'}$ for manufactures). We then calculate the changes in wages $\hat{w}_i$ and prices $\hat{p}_i$ needed to re-equilibrate the world economy. Our baseline is the world as it was in 2009, the last year for which data are available for all of our countries.

**Gains from Trade**

As discussed above, we can measure the gains from trade using data on only the home share, $\pi_{ii}$. For this exercise we employ a direct measure of the home share: gross manufacturing production less *gross* exports, divided by gross manufacturing production less *net* exports.[8] This statistic deserves some consideration on its own.

[8] In our counterfactual simulations, we use a different measure, as in Dekle, Eaton, and Kortum (2007).

*Table 2*
**The Home Share of Spending on Manufactures and Gains from Trade**

| | | *Home share of spending* | | *Implied gains from trade* | |
|---|---|---|---|---|---|
| *Country* | *World GDP share (%) in 2006* | *Level in 2006 (%)* | *Change since 1996 (percentage points)* | *Level in 2006 (%)* | *Change since 1996 (percentage points)* |
| Austria | 0.66 | 31.4 | –16.2 | 21.3 | 8.1 |
| Canada | 2.60 | 49.1 | –1.5 | 12.6 | 0.6 |
| Czech Republic | 0.29 | 42.6 | –14.7 | 15.3 | 5.5 |
| Denmark | 0.56 | 25.6 | –18.1 | 25.5 | 10.7 |
| Estonia | 0.03 | 2.5 | –19.6 | 85.4 | 56.7 |
| Finland | 0.42 | 58.2 | –7.3 | 9.4 | 2.1 |
| France | 4.60 | 56.9 | –10.3 | 9.9 | 3.0 |
| Germany | 5.94 | 53.7 | –16.4 | 10.9 | 4.8 |
| Greece | 0.54 | 52.7 | –11.6 | 11.3 | 3.6 |
| Hungary | 0.23 | 26.0 | –34.5 | 25.1 | 16.4 |
| Iceland | 0.03 | 27.9 | –10.0 | 23.7 | 6.2 |
| Ireland | 0.46 | 39.6 | 9.9 | 16.7 | –5.7 |
| Italy | 3.80 | 68.9 | –7.1 | 6.4 | 1.7 |
| Japan | 8.88 | 84.9 | –5.6 | 2.8 | 1.1 |
| Korea | 1.94 | 77.2 | –0.7 | 4.4 | 0.1 |
| Mexico | 1.94 | 58.3 | –7.9 | 9.4 | 2.3 |
| New Zealand | 0.22 | 53.6 | –8.2 | 11.0 | 2.6 |
| Norway | 0.68 | 51.9 | –2.5 | 11.6 | 0.9 |
| Poland | 0.69 | 53.4 | –15.8 | 11.0 | 4.7 |
| Portugal | 0.41 | 50.8 | –10.2 | 12.0 | 3.4 |
| Slovenia | 0.08 | 27.2 | –15.5 | 24.3 | 9.0 |
| Spain | 2.51 | 62.8 | –10.2 | 8.1 | 2.7 |
| Sweden | 0.81 | 49.2 | –10.0 | 12.5 | 3.4 |
| Switzerland | 0.80 | 35.3 | –20.0 | 18.9 | 8.6 |
| United States | 27.26 | 73.5 | –8.3 | 5.3 | 1.9 |
| All others | 33.62 | | | | |

*Source:* Authors' calculations from the OECD STAN (STructural ANalysis) Database, the Economist Intelligence Unit, and a model described in the text.
*Notes:* The home share is the share a country spends on domestic manufactures out of total country spending on manufactures. The last two columns calculate the implications of the level of the home share, and its changes over time, for countries' gains from trade and how those gains have evolved. We look at the gains from trade only in manufactures.

Table 2 reports the home share in 2006 for the 25 countries with data on gross manufacturing production. The mean value of the home share is just under 50 percent. In a world of frictionless trade (all $d_{ni} = 1$), there is no reason for a country to spend a larger share of its income on its own goods than any other country. A country's home share, in that case, would correspond to its share in world output. As Table 2 makes clear, for each of these countries the home share is many times larger than the country's share in world GDP: three times higher for the United States, ten times for Germany, 50 times for Denmark, and 100 times for Greece. Such multiples illustrate the extent to which trade barriers continue to chop up world markets. Even though countries buy much more of their manufactures from home than a world of costless trade would predict, in line with theory large countries tend to buy much more from themselves than small countries: The overall correlation between home share and share in GDP is close to 0.5 in 2006.

The third column of Table 2 shows that the home share declined substantially between 1996 and 2006, reflecting globalization of manufactures production over the period. (Only Ireland bucked this trend.) The last two columns calculate the implications of the level of the home share, and its changes over time, for countries' gains from trade and how those gains have evolved. In making these calculations, our first two embellishments to the model require two modifications. Since we look at the gains from trade only in manufactures, the fact that manufactures are only 20 percent of final spending limits the benefit. But since manufactures are a major input into the production of manufactures, there are large indirect benefits of trade in lowering input costs. Putting the two together, the elasticity that translates a smaller home share into larger gains from trade is no longer $1/\theta$ but rather $\alpha/(\beta\theta) = 1/6$ . Thus, we calculate the gains from trade for country $i$ at date $t$ as

$$G_i^t = 100[(\pi_{ii}^t)^{-1/6} - 1]$$

where $\pi_{ii}^t$ is country $i$'s home share at date $t$.

Clearly, gains from trade are substantial, particularly for small countries: for example over 25 percent of income for Denmark, Estonia, and Hungary. For the largest countries, Japan and the United States, gains from trade amounted to 2–3 percent of GDP 20 years ago. But those gains are now over 50 percent higher.

### Benefits of Further Globalization

Our measure of the gains from trade compares where we are now with no trade. We can also consider the gains that would accrue in the future if globalization, driven by lower trade costs, continues. Our counterfactual experiment considers a uniform proportional 25 percent drop in the costs of trade, ($\hat{d}_{ni} = 0.75$ for all foreign-country pairs), a magnitude chosen so that world trade in manufactures approximately doubles relative to world GDP. As a point of reference, world trade in goods and services did double relative to world GDP over the past 30 years.

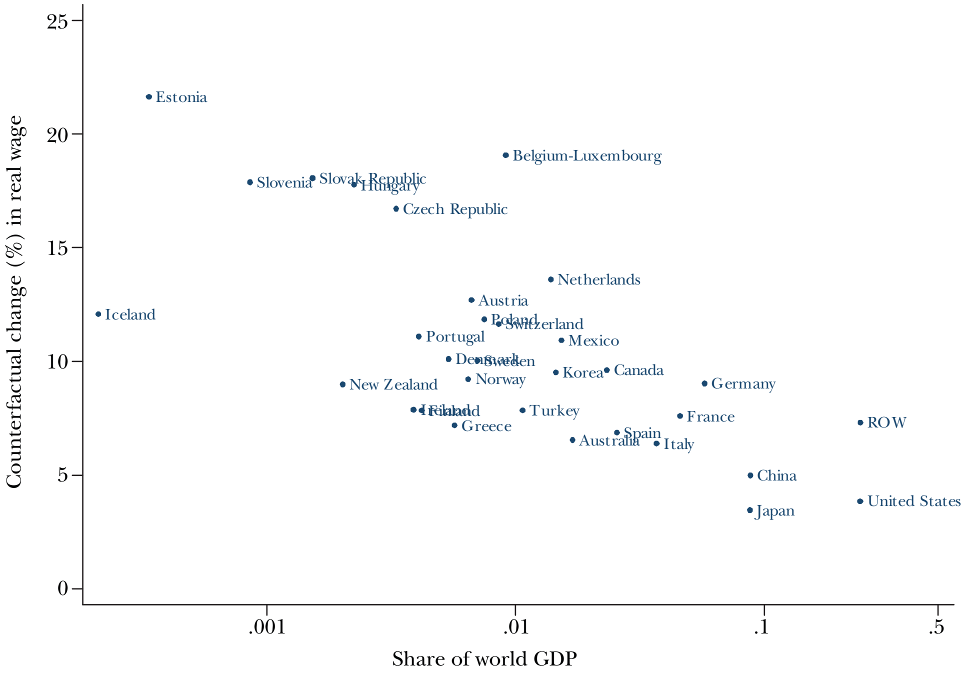
Figure 3 plots the results against each countries' share of world GDP. The gains, measured by the increase in the real wage, are substantial, with a median gain of about 10 percent. The gains are also very heterogeneous, with small countries typically gaining proportionately much more than large countries. Given their size, isolated countries, such as Iceland, New Zealand, and Greece, gain much less than countries proximate to many others, such as Belgium–Luxembourg, the Netherlands, and Germany.

### Benefits of Technological Improvements

As the basic Ricardian model illustrates, trade provides a conduit through which foreign countries benefit from an improvement in a country's ability to produce a good. We can measure the strength of this mechanism by considering the effects

*Figure 3*

**Real Wage Response to a Decrease in Trade Barriers**



*Source:* Authors' calculations using data from the OECD STAN (STructural Analysis) Database and the Economist Intelligence Unit and a model described in the text.
*Notes:* We consider a uniform proportional 25 percent drop in the costs of trade, a magnitude chosen so that world trade in manufactures approximately doubles relative to world GDP. The figure plots the counterfactual change in real wage against each countries' share of world GDP.

on welfare around the world from a shift in the distribution of technologies in a particular country $i$ (as reflected in the parameter $A_i$). Our particular experiment makes the United States 10 percent more productive, so that $\hat{A}_{US} = 1.1$.

The world economy responds in two important ways: First, the U.S. wage rises by about 30 percent relative to other countries' wages. Second, the U.S. real wage (in terms of goods and services) rises by about 6 percent, while real wages in other countries increase by only a small amount, if at all.

The effects of geography are apparent as the greatest foreign beneficiaries are Canada and Mexico, which experience a real wage gain one-tenth that in the United States. A few countries, if they are initially running a trade surplus in manufactures, experience a small real wage decline. (If we first eliminate all trade imbalances and then increase U.S. technology, all foreign countries experience a real wage gain.)

Overall, the increase in U.S. technology raises the GDP-weighted real wage around the world by 1.6 percent, with 8 percent of this gain experienced outside the United States. Foreign countries gain both due to the lower prices of final goods and of intermediate inputs relative to wages. Performing the same experiment, but with China in place of the United States, yields similar results. Better technology in China raises the world's average real wage by 0.6 percent, with 10 percent of this gain experienced outside China. These results reflect the fact that the improvement in technology in China adds to a smaller base, yet China's greater export orientation means the overall benefits are spread somewhat more to foreign destinations.[9]

### Consequences of Eliminating Current Account Imbalances

Our model, like Dornbusch, Fischer, and Samuelson (1977) with trade costs, implies that transfers between countries have implications for relative wages. Our final counterfactual, following Dekle, Eaton, and Kortum (2007), considers exogenous shifts in manufacturing trade deficits that would simultaneously balance every country's current account, holding fixed any deficits outside of manufacturing. We also hold trade costs and technologies fixed. Table 3 shows the results. To undo the huge 2009 U.S. deficit, the wage in most countries rises relative to the U.S. wage by over 13 percent in China and 14 percent in Germany since the large surpluses of these two must decline. The small European deficit countries of Greece and Portugal are the dramatic exceptions, declining 21 and 12 percent relative to the United States.

Figure 4 shows that initial current account balances (as a share of GDP), determining the required adjustment of trade imbalances, go a long way toward explaining the direction (positive) and magnitude of wage adjustment. A question is why Iceland, Portugal, and Greece experience very different wage responses even though their current account deficit to GDP ratios were similar in 2009. It turns out that another important factor in explaining the magnitude of the change in wages is the initial size of the manufacturing sector in a country's GDP. This share is lowest in Greece, worsening the wage decline necessary to bring about current account balance via an increase in net exports of manufactures.

The consequences for the real wage are much more muted than those for the relative wage. Even Greece, the most negatively affected, suffers less than a 4 percent decline. The United States, with its large current account deficit, would see its real wage decline by only half of 1 percent. The reason is a combination of large home shares in manufacturing spending and small shares of manufactures in overall final demand. For goods and services produced at home, prices move in line with wages. The change in the relative wage acts only through import prices.

More dramatic are the changes in the share of manufacturing in GDP required to rebalance current accounts. This share rises by over 10 percentage points in Iceland and by nearly as much in Greece and Portugal. It falls by over 4 percentage

[9] Our results on the benefits of foreign technology are much smaller than some of the results in Eaton and Kortum (2002). The main reason is in that paper the mobile-labor case held wages fixed so that foreign countries did not suffer a decline in their terms of trade. Our results here, in that respect, are more in line with the immobile-labor case of Eaton and Kortum (2002). See Fieler (2011) and Hsieh and Ossa (2011) for a related analysis of the benefits to foreign countries of China's technology gains. Hsieh and Ossa's analysis is retrospective rather than counterfactual.

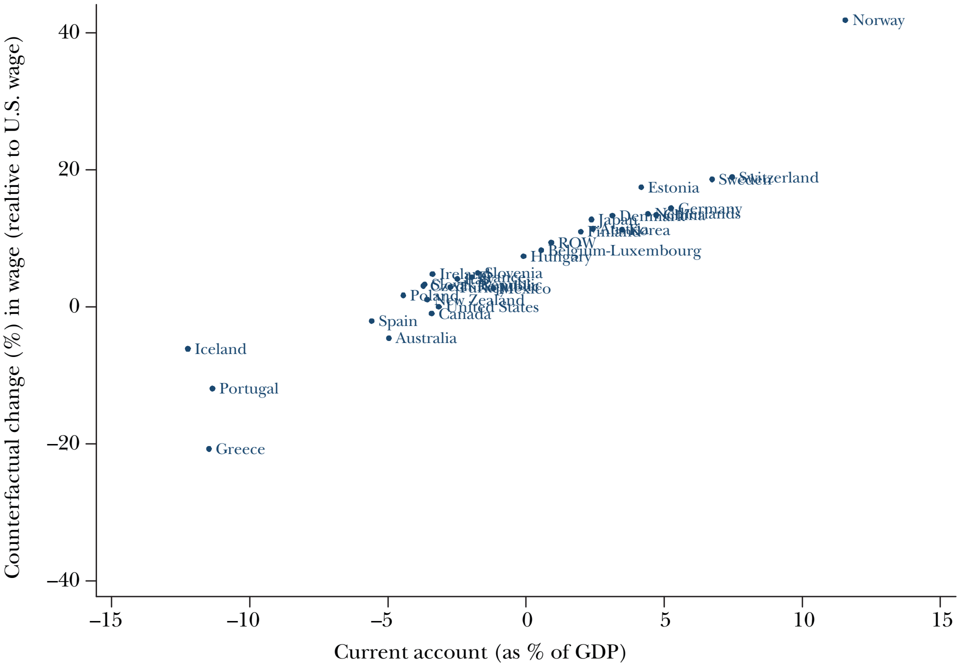
*Table 3*
**Consequences of Eliminating Current Account Imbalances**

| | *Data* | | | *Counterfactuals* | | |
|---|---|---|---|---|---|---|
| | | | | *Change in* | | *Change in mfg share (percentage points)* |
| *Country* | *GDP (US$ billions)* | *Current account balance (% GDP)* | *Manufactures trade balance (% GDP)* | *Relative wage (%)* | *Real wage (%)* | |
| Australia | 973.7 | –5.0 | –8.1 | –4.6 | –1.4 | 3.5 |
| Austria | 382.0 | 2.4 | 1.2 | 11.4 | 0.3 | –1.9 |
| Belgium-Luxembourg | 525.2 | 0.6 | 7.4 | 8.3 | 0.0 | –0.5 |
| Canada | 1337.6 | –3.4 | –4.7 | –1.0 | –0.7 | 2.6 |
| China | 5050.5 | 4.7 | 10.6 | 13.4 | 0.3 | –4.1 |
| Czech Republic | 190.2 | –3.7 | 6.4 | 3.1 | –0.9 | 3.3 |
| Denmark | 308.9 | 3.1 | 1.3 | 13.3 | 0.4 | –2.4 |
| Estonia | 19.3 | 4.2 | –3.9 | 17.5 | 1.5 | –2.6 |
| Finland | 241.3 | 2.0 | 5.6 | 11.0 | 0.1 | –1.7 |
| France | 2632.7 | –2.0 | –1.2 | 4.3 | –0.4 | 1.6 |
| Germany | 3308.3 | 5.2 | 8.6 | 14.4 | 0.7 | –4.4 |
| Greece | 326.4 | –11.5 | –12.3 | –20.7 | –3.7 | 8.7 |
| Hungary | 128.8 | –0.1 | 7.7 | 7.4 | –0.3 | 0.0 |
| Iceland | 12.1 | –12.2 | 3.2 | –6.1 | –2.1 | 11.4 |
| Ireland | 223.8 | –3.4 | 27.9 | 4.8 | –0.2 | 3.2 |
| Italy | 2116.7 | –2.5 | 2.9 | 4.1 | –0.4 | 2.1 |
| Japan | 5031.6 | 2.4 | 3.4 | 12.7 | 0.2 | –2.0 |
| Korea | 834.1 | 3.5 | 16.1 | 11.3 | 0.2 | –3.3 |
| Mexico | 879.2 | –1.2 | –2.9 | 2.7 | –0.4 | 0.9 |
| Netherlands | 796.2 | 4.4 | 7.5 | 13.6 | 0.7 | –3.7 |
| New Zealand | 116.2 | –3.6 | –2.2 | 1.1 | –0.6 | 2.9 |
| Norway | 370.7 | 11.6 | –6.0 | 41.9 | 3.9 | –4.8 |
| Poland | 430.5 | –4.4 | 0.1 | 1.7 | –1.0 | 3.8 |
| Portugal | 234.9 | –11.4 | –7.6 | –11.9 | –2.6 | 9.7 |
| Slovak Republic | 87.8 | –3.7 | 6.5 | 3.3 | –0.9 | 3.3 |
| Slovenia | 49.2 | –1.7 | –1.4 | 5.0 | –0.5 | 1.4 |
| Spain | 1468.4 | –5.6 | –2.4 | –2.0 | –1.0 | 4.7 |
| Sweden | 403.5 | 6.7 | 4.8 | 18.6 | 1.0 | –5.2 |
| Switzerland | 492.3 | 7.5 | 4.6 | 18.9 | 1.3 | –5.7 |
| Turkey | 613.8 | –2.7 | –2.7 | 2.9 | –0.6 | 2.2 |
| United States | 13939.0 | –3.2 | –2.6 | 0.0 | –0.5 | 2.6 |
| ROW | 13961.0 | 0.9 | –5.4 | 9.4 | 0.2 | –0.6 |

*Source:* Authors' calculations from the OECD STAN (STructural ANalysis) Database, the Economist Intelligence Unit, and a model described in the text.
*Notes:* We consider the effects of exogenous shifts in manufacturing trade deficits that would simultaneously balance every country's current account, holding fixed any deficits outside of manufacturing. Data are for 2009. "Relative wage" is the wage relative to the United States. "ROW" is "rest of world."

points in the large, surplus countries (China and Germany) and by 5 percentage points in the smaller ones (Norway, Sweden, and Switzerland). These extreme predictions about the impact on the size of the manufacturing sector follow from our Ricardian assumption that labor can flow seamlessly between manufacturing

*Figure 4*
**Wage Response to Eliminating Current Account Imbalances**



*Source:* Authors' calculations using data from the OECD STAN (STructural Analysis) Database and the Economist Intelligence Unit and a model described in the text.
*Notes:* We consider the effects of exogenous shifts in manufacturing trade deficits that would simultaneously balance every country's current account, holding fixed any deficits outside of manufacturing, as in Table 3. The figure plots the counterfactual change in wage relative to the United States against the initial current account balance as a share of GDP. "ROW" is "rest of world."

and other activities. In Dekle, Eaton, and Kortum (2008), we introduce rigidities and examine their effect.

## Extending and Improving the Tool

Much recent work has extended this new old Ricardian trade theory in various ways, sometimes combining elements of it with other theories to address new questions. Here we briefly discuss a few of these contributions.

The field of international trade has traditionally used industry as its unit of analysis, a natural choice given the heterogeneity of industries and the fact that most trade policy is implemented at the industry level. In moving from a small number of goods, with labor requirements specified in a table, to a continuum of goods, with labor requirements only described probabilistically, we lose track of this industry dimension. A number of papers have brought industries back into the analysis, including Chor (2010) and Shikher (2011). The idea is that each industry $k$ consists of a continuum of differentiated goods and each country $i$ has an absolute advantage parameter $A_{ik}$ in each industry. Costinot, Donaldson, and Komunjer (forthcoming) use this approach to revisit the connection between trade and industry-level productivity implied by Ricardian theory, avoiding ambiguities that plagued the early analysis of MacDougall (1951, 1952). Incorporating input-output linkages between industries, Caliendo and Parro (2010) use the model to explore the welfare gains from tariff reductions under the North American Free Trade Agreement (NAFTA).

While the basic Ricardian trade model treats labor as the only primary factor, many applications require incorporating other factors of production. In his monumental study measuring the gains from rail transport in nineteenth-century India, Donaldson (2010) applies this Ricardian model, replacing labor with land as the primary factor, with land rents appearing in place of wages. Incorporating several factors, and multiple industries, leads to a hybrid Ricardian–Heckscher-Ohlin model, as in Shikher (2011), used by Parro (2012) to account for the rise of the skill premium in developing and developed countries. Burstein and Vogel (2010) interweave these two theories at a deeper level, introducing a correlation between labor requirements and skill intensity at the level of the individual goods on the continuum.

Our applications above were limited to trade in manufactures among OECD countries. Extending the analysis more broadly, the theory has to confront the fact that low-income countries trade less than high-income ones, even taking into account their economic size and location. Waugh (2010) proposes a model in which barriers to exporting are the culprit, consistent with evidence on prices. Fieler (2011) pursues another explanation, introducing different classes of goods with different income elasticities of demand and with different degrees of technological heterogeneity (in the notation in the model we have presented here, different $\theta$'s). She finds that poor countries have a comparative advantage in goods that are both more income inelastic and more technologically homogeneous (that is, with a higher $\theta$). Tombe (2011) finds that barriers to food trade are higher than for manufactures, particularly in poor countries. He also departs from the standard Ricardian tradition by introducing barriers to domestic labor mobility between rural areas (where food is produced) and cities (which produce manufactures and services).

In keeping with Ricardo's original analysis, the models discussed so far mostly assume perfect competition. Breaking with that tradition, in Bernard, Eaton, Jensen, and Kortum (2003), we incorporate Bertrand competition, allowing the theory to make contact with data on individual producers. This extension also opens up the possibility of addressing pricing puzzles in international economics, as explored in Atkeson and Burstein (2008). While the basic model with Bertrand competition yields a distribution of price markups that is invariant to trade, de Blas and Russ (2010) develop a variant of the model that breaks that result. Holmes, Hsu, and Lee (2012) investigate a related model that yields new results on the gains from trade.

Having introduced imperfect competition, in Eaton and Kortum (2001) we show that innovation and growth fit seamlessly into the theory. Incorporating technology diffusion and multinational production has turned out to be more challenging. The problem is that the theory can easily deliver myriad treads and risers again when groups of countries have access to the same technologies for producing some goods. Recent work by Ramondo and Rodríguez-Clare (2009) has begun to map a way through these difficulties. Another promising approach, representing a greater departure from the basic theory, is pursued by Alvarez, Buera, and Lucas (2011).

In short, the framework we present in this paper is tractable, versatile, and amenable to empirical analysis. It is keeping Ricardo busy.

■ *We would like to thank Daisuke Fuji, Brodie Olsen, and Sebastian Sotelo for excellent research assistance.*

## References

**Adam Smith.** 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations.* Available online at the Library of Economics and Liberty: http://www.econlib.org/library/Smith/smWN.html.

**Alvarez, Fernando, and Robert E. Lucas, Jr.** 2007. "General Equilibrium Analysis of the Eaton-Kortum Model of International Trade." *Journal of Monetary Economics* 54(6): 1726–68.

**Alvarez, Fernando, Francisco Buera, and Robert E. Lucas.** 2011. "Idea Flows, Economic Growth, and Trade." Unpublished paper.

**Anderson, James E., and Eric van Wincoop.** 2004. "Trade Costs." *Journal of Economic Literature* 42(3): 691–751.

**Arkolakis, Costas, Arnaud Costinot, and Andrés Rodriguez-Clare.** 2012. "New Trade Models, Same Old Gains?" *American Economic Review* 102(1): 94–130.

**Atkeson, Andrew, and Ariel Burstein.** 2008. "Pricing-to-Market, Trade Costs, and International Relative Prices." *American Economic Review* 98(5): 1998–2031.

**Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum.** 2003. "Plants and Productivity in International Trade." *American Economic Review* 93(4): 1268–90.

**Burstein, Ariel, and Jonathan Vogel.** 2010. "Globalization, Technology, and the Skill Premium: A Quantitative Analysis." NBER Working Paper 16459.

**Caliendo, Lorenzo, and Fernando Parro.** 2010. "Estimates of the Trade and Welfare Effects of NAFTA." Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1916287.

**Chipman, John S.** 1965. "A Survey of the Theory of International Trade: Part 1, The Classical Theory." *Econometrica* 33(3): 477–519.

**Chor, Davin.** 2010. "Unpacking Sources of Comparative Advantage: A Quantitative Approach." *Journal of International Economics* 82(2): 152–67.

**Costinot, Arnaud, Dave Donaldson, and Ivana Komunjer.** Forthcoming. "What Goods Do Countries Trade? A Quantitative Exploration of Ricardo's Ideas." *Review of Economic Studies.* (Published online September 29, 2011: http://restud.oxfordjournals.org/content/early/2011/09/28/restud.rdr033.short?rss=1.)

**de Blas, Beatriz, and Katheryn Niles Russ.** 2010. "Understanding Markups in the Open Economy under Bertrand Competition." NBER Working Paper 16587.

**Dekle, Robert, Jonathan Eaton, and Samuel Kortum.** 2007. "Unbalanced Trade." *American Economic Review* 97(2): 351–55. (Longer version is NBER Working Paper 13035).

**Dekle, Robert, Jonathan Eaton, and Samuel Kortum.** 2008. "Global Rebalancing with Gravity: Measuring the Burden of Adjustment." *IMF Staff Papers* 55(3): 511–40.

**Donaldson, Dave.** 2010. "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure." Asia Research Center Working Paper 41, London School of Economics and Political Science.

**Dornbusch, Rudiger, Stanley Fischer, and Paul A. Samuelson.** 1977. "Comparative Advantage, Trade, and Payments in a Ricardian Model with a Continuum of Goods." *American Economic Review* 67(5): 823–39.

**Eaton, Jonathan, and Samuel Kortum.** 1999. "International Technology Diffusion: Theory and Measurement." *International Economic Review* 40(3): 537–70.

**Eaton, Jonathan, and Samuel Kortum.** 2001. "Technology, Trade, and Growth: A Unified Framework." *European Economic Review* 459(4–6): 742–55.

**Eaton, Jonathan, and Samuel Kortum.** 2002. "Technology, Geography, and Trade," *Econometrica* 70(5): 1741–80.

**Feenstra, Robert C.** 2004. *Advanced International Trade: Theory and Evidence.* Princeton University Press.

**Fieler, Cecilia.** 2011. "Non-Homotheticity and Bilateral Trade: Evidence and a Quantitative Explanation." *Econometrica* 79(4): 1069–1101.

**Graham, Frank D.** 1948. *The Theory of International Values.* Princeton University Press.

**Holmes, Thomas J., Wen-Tai Hsu, and Sanghoon Lee.** 2012. "Allocative Efficiency, Mark-ups, and the Welfare Gains from Trade." http://www.econ.umn.edu/~holmes/papers/alloc_eff_trade.pdf.

**Hsieh, Chang-Tai, and Ralph Ossa.** 2011. "A Global View of Productivity Growth in China." NBER Working Paper 16778.

**Jones, Ronald W.** 1961. "Comparative Advantage and the Theory of Tariffs: A Multi-Country Multi-Commodity Model." *Review of Economic Studies* 28(3): 161–75.

**Kortum, Samuel S.** 1997. "Research, Patenting, and Technological Change." *Econometrica* 65(6): 1389–1419.

**Krugman, Paul R., and Anthony Venables.** 1995. "Globalization and the Inequality of Nations." *Quarterly Journal of Economics* 110(4): 857–80.

**Lucas, Robert.** 1976. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets,* edited by K. Brunner and A. H. Meltzer, 19–46. New York: Elsevier..

**MacDougall, G. D. A.** 1951. "British and American Exports: A Study Suggested by the Theory of Comparative Costs. Part I." *Economic Journal* 61(244): 697–724.

**MacDougall, G. D. A.** 1952. "British and American Exports: A Study Suggested by the Theory of Comparative Costs. Part II." *Economic Journal* 62(247): 487–521.

**Matsuyama, Kiminori.** 2008. "Ricardian Trade Theory." In *The New Palgrave Dictionary of Economics,* 2nd Edition, edited by Lawrence E. Blume and Steven N. Durlauf. Palgrave Macmillan.

**McKenzie, Lionel W.** 1953. "Specialisation and Efficiency in World Production." *Review of Economic Studies* 21(3): 165–80.

**McKenzie, Lionel W.** 1954. "On Equilibrium in Graham's Model of World Trade and Other Competitive Systems." *Econometrica* 22(2): 147–61.

**Mill, John Stuart.** 1844. *Essays on Some Unsettled Questions of Political Economy.* London: John W. Parker.

**Parro, Fernando.** 2012. "Capital-Skill Complementarity and the Skill Premium in a Quantitative Model of Trade." Available at: http://sites.google.com/site/fernandoparro1/home/research-1.

**Ramondo, Natalia, and Andrés Rodríguez-Clare.** 2009. "Trade, Multinational Production, and the Gains from Openness." NBER Working Paper 15604.

**Ricardo, David.** 1817. *On the Principles of Political Economy and Taxation.* Available online at the Library of Economics and Liberty: http://www.econlib.org/library/Ricardo/ricP.html.

**Shikher, Serge.** 2011. "Capital, Technology, and Specialization in the Neoclassical Model." *Journal of International Economics* 83(2): 229–42.

**Simonovska, Ina, and Michael E. Waugh.** 2011. "The Elasticity of Trade: Estimates and Evidence." https://files.nyu.edu/mw134/public/uploads/56836/estimate_theta_paper.pdf.

**Tombe, Trevor.** 2011. "The Missing Food Problem." http://dl.dropbox.com/u/6874356/missing_food_rev1.pdf.

**Waugh, Michael E.** 2010. "International Trade and Income Differences." *American Economic Review* 100(5): 2093–2124.

**Weintraub, Roy E.** 2011. "Retrospectives: Lionel W. McKenzie and the Proof of the Existence of a Competitive Equilibrium." *Journal of Economic Perspectives* 25(2): 199–215.

Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 91–118

# Gains from Trade when Firms Matter

Marc J. Melitz and Daniel Trefler

The gains from long-distance international trade have been understood and exploited since prehistoric times. Our pre-urban ancestors were benefitting from long-distance trade in obsidian some 10,000 years ago; Plato's Academy was built on the profits of Athenian silver exports; and Rome was not built in a day partly because goods moved too slowly in the vast Roman trade network.

But whereas trade was once dominated by the movement of goods that could only be produced, harvested, or mined regionally, the international trade landscape is now dominated by two striking facts. The first is the rise of intra-industry trade—that is, two-way trade in similar products. Chinese consumers can now buy a midsize car from Toyota (Japan), Kia (Korea), General Motors (United States), and Chery (China). Ditto for consumers in Japan, Korea, and the United States. The second striking fact is that world trade is dominated by huge, innovative, and extraordinarily productive firms. For example, Intel is so large that it is the largest industrial employer in both Oregon and New Mexico and accounts for 20 percent of Costa Rica's exports. China's Foxconn infamously employs 450,000 workers in a *single* one of its many export-oriented electronics factories. These are big companies . . . and if you are reading this document on an Apple computer you know that there are other large companies, too.

The rising prominence of intra-industry trade and huge multinationals has transformed the way economists think about the gains from trade. In the past, we focused on gains that stemmed either from endowment differences (wheat for

■ *Marc J. Melitz is the David A. Wells Professor of Political Economy, Harvard University, Cambridge, Massachusetts. Daniel Trefler is the J. Douglas and Ruth Grant Canada Research Chair in Competitiveness and Prosperity, Rotman School of Management, University of Toronto, Toronto, Ontario, Canada.*

http://dx.doi.org/10.1257/jep.26.2.91. doi=10.1257/jep.26.2.91

iron ore) or inter-industry comparative advantage (David Ricardo's classic example of cloth for port). Today, we focus on three sources of gains from trade: 1) love-of-variety gains associated with intra-industry trade; 2) allocative efficiency gains associated with shifting labor and capital out of small, less-productive firms and into large, more-productive firms; and 3) productive efficiency gains associated with trade-induced innovation.

Back in the 1980s, a "New Trade Theory" was developed that focused on intra-industry trade in differentiated goods produced subject to increasing returns to scale. This theory centered on an elegant tension: consumers love variety and are willing to pay a premium for a desired product, but as the market fragments into niche products, producers struggle to attain the volumes needed to recoup their product development costs. International trade creates a larger marketplace, which means that each firm can operate at a larger scale and hence more firms can survive. The result reads like an advertisement for free trade: lower prices, more varieties. Paul Krugman earned the Nobel Prize in 2008 in large part for his work highlighting how economies of scale and product differentiation lead to intra-industry trade, just as in our example above of midsize cars. See Krugman (1979, 1980), Helpman and Krugman (1985), and Helpman (2011, chap. 4) for a review of love-of-variety gains from trade.

More recently, a second source of gains from trade has emerged from the research of Melitz (2003) and Bernard, Eaton, Jensen, and Kortum (2003). This is the firm-level "reallocation" effect that arises when there is firm heterogeneity. By firm heterogeneity we mean that even within narrowly defined industries some firms are much larger and more profitable than others because, for example, they are much more productive. Globalization generates both winners and losers among firms within an industry and these effects are magnified by heterogeneity. Better-performing firms thrive and expand into foreign markets, while worse-performing firms contract and even shut down in the face of foreign competition. This generates a new source of gains from trade: as production is concentrated towards better-performing firms, the overall efficiency of the industry improves. In this way, globalization raises average efficiency *within an industry*. Why is it that only the better-performing firms grow? Globalization expands markets but also increases competition in those markets. This competition effect dominates for the worse-performing firms while the increased market access dominates for the better-performing firms. Also, a firm's international expansion—whether by exporting, by offshore outsourcing of intermediate components and assembly, or by building plants abroad (multinationals)—entails some up-front fixed costs; and only the best-performing firms have the sales volumes needed to justify these fixed costs.

Our third source of gains from trade comes from the positive impacts of larger markets on *innovation*. New productivity-enhancing products and processes require up-front development costs. Trade integration, by expanding the size of the market, encourages firms to pony up these development dollars, and this in turn raises productivity. Theories of innovation-based gains from trade with homogeneous firms were developed by Grossman and Helpman (1991) and are supported by country-level evidence (Helpman 2004, chap. 5.6). At the firm level, there is a strong relationship between exporting and innovation. For example, Intel and Apple are major patent holders, and Foxconn holds 40 percent of all Chinese patents filed in the United States (Eberhardt, Helmers, and Yu 2011). Of course, this correlation between exporting and innovation does not provide evidence for causality and lacks a framing theory featuring heterogeneous firms. Recently, however, there has been a great deal of theoretical and empirical progress. Lileeva and Trefler (2010) show theoretically and empirically how the market-expanding effects of international integration causally encourage firms to innovate. Verhoogen (2008), Bustos (2011a, b), and Aw, Roberts, and Xu (2011) assess other interesting channels through which trade promotes firm-level innovation. Note that this third source of gains deals with *within-firm* efficiency; in contrast, the second source of gains above deals with *between-firm* or allocative efficiency.

This paper reviews these three sources of gains from trade both theoretically and empirically. Our empirical evidence will be centered on the experience of Canada following its closer economic integration in 1989 with the United States—the largest example of bilateral intra-industry trade in the world—but we will also describe evidence for other countries.

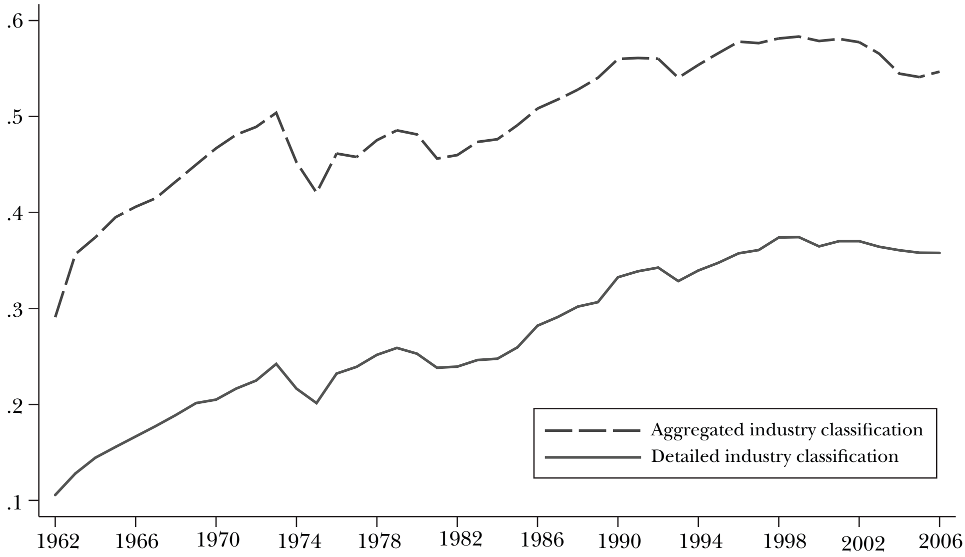
The related literature is huge. Here we focus on firms that expand internationally via exporting as in Melitz (2003) and Bernard, Eaton, Jensen, and Kortum (2003). Another related research topic analyzes how firm boundaries evolve across borders as multinational firms look abroad to "outsource" key parts of their production chain. The interested reader is directed to surveys by Antràs and Rossi-Hansberg (2009) and Helpman (2011, chap. 6).

## Gains from Love of Variety (Economies of Scale and Product Variety)

Our first source of gains from trade is intimately related to intra-industry trade. To measure intra-industry trade, one needs to start with a classification system that assigns trade flows to a particular "industry." One can then categorize trade flows as either intra-industry (two-way trade within the same industry classification code) or inter-industry (imports and exports in separate industry codes). The United Nations uses the Standard International Trade Classification, or SITC, to categorize world trade flows. In its most detailed form, the SITC contains 1,161 separate industry codes (that can be consistently traced back over time), but these industries are often aggregated into a smaller subset of industries.

Figure 1 shows the time trend for the share of intra-industry trade according to this most detailed classification, and a more aggregated version with only 59 industry codes. Mechanically, the share of intra-industry trade rises with the level of aggregation for the industrial classification system (after all, with a single aggregate industry code, all trade would be "intra" to this aggregated industry). However, the time trends for the two series are very similar: intra-industry trade grew rapidly from 1962 to the mid 1990s, before stabilizing at a substantially higher level. As countries

*Figure 1*
**World Share of Intra-Industry Trade 1962–2006**



*Source:* Data from Brühlhart (2009). We thank Marius Brülhart for generously sharing his data.
*Notes:* Figure 1 shows the time trend for the share of intra-industry trade according to the most detailed Standard International Trade Classification (1,161 separate industry codes) and a more aggregated version with only 59 industry codes.

industrialize, they tend to experience a higher share of intra-industry trade because they tend to produce and export differentiated manufactured goods that are similar to other brands of goods that are imported. However, some of the countries with the highest shares of intra-industry trade in 2000 were newly industrializing nations, such as the Czech Republic (77 percent), the Slovak Republic (76 percent), Mexico (73 percent), and Hungary (72 percent); for comparison, the United States had a 69 percent share of intra-industry trade in 2000 (*OECD Economic Outlook* 2002, chap. 6; based on the 59-industry level of aggregation). Most recently, China's share of intra-industry trade has risen above the 50 percent mark.

Why might a country both export and import goods that are similar? As a starting point, consider world trade in automobiles. Consumers in a car-producing country are not limited to buying the car models that are produced domestically: many of those consumers choose to buy models that are produced elsewhere and imported. The extent of this product differentiation is then limited by high fixed start-up costs for a new brand and by the related economies of scale.

We now highlight how the combination of product differentiation and economies of scale generates intra-industry trade using a theoretical example. Notice that this source of gains from trade provides a rationale for trade between two identical countries, which provides a stark contrast with the gains from inter-industry trade that arise from exploiting differences across countries, such as differences in technology (Ricardo) or differences in factor supplies (Heckscher–Ohlin).

In our theoretical example, two identical countries produce differentiated widget varieties subject to the same constant-returns-to-scale technology. Assume that one worker can produce one widget, but that production of any new variety of widgets requires four workers to cover fixed overhead costs: this implies decreasing average costs of production as the fixed cost is spread over an increasing number of output units (hence the economies of scale). Also, to be specific, suppose that both countries have a fixed supply of 12 workers. If they do not trade, then each country can produce: A) 8 units of one variety, or B) 2 units each of two different varieties.

Allowing countries to trade leads to a new possibility that is better than what either country can achieve on its own. Suppose that each country produces 8 units of one variety and exports 4 of these units to the other country. Consumers are now consuming 4 units of the home variety and 4 units of the foreign variety. This is preferred to either of the no-trade production plans above. Compared to choice B, there is the same number of varieties (2 varieties), but more of each variety (four versus two). Compared to choice A, there is the same number of units (8 units), but more varieties (two versus one). Thus, trade expands the set of consumer choices and eases the tradeoff between consumption units and product variety. Economic integration allows production of each individual variety to be consolidated for the whole integrated market; given increasing returns to scale, this reduces average production costs. At the same time, product variety increases because consumers can buy varieties produced anywhere in the integrated market.

One of the most salient real-world examples of economic integration between similar countries occurred between the United States and Canada. This integration started with the signing of the North American Auto Pact in 1964. Before then, most car models were produced in the United States for U.S. consumers and in Canada for Canadian consumers. High tariffs on auto trade made it uneconomical to export most car models across the border. Because the Canadian auto market was roughly one-tenth the size of the U.S. market, this implied substantial scale disadvantages for production in the Canadian market: labor productivity there was about 30 percent below the U.S. level. The U.S. market was large enough that assembly lines could be dedicated to one particular car model, while Canadian assembly lines had to switch across models, involving costly downtime and reconfiguration costs, and also had to hold substantially higher inventory levels.

The 1964 Pact established a free trade area for autos that allowed manufacturers to consolidate the production of particular car models in one country and export that model to consumers in the other country. For example, General Motors cut in half the number of models assembled in Canada. However, total production of autos in Canada increased as the remaining models produced in Canada supplied the U.S. market as well as the Canadian one. Canadian automotive exports to the United States increased from $16 million in 1962 to $2.4 billion in 1968. That same year, U.S. automotive exports to Canada were valued at $2.9 billion: intra-industry trade in action. Today, $85 billion worth of automotive products cross the

U.S.–Canada border each year—roughly half in each direction. The productivity gains associated with this consolidation were also substantial: by the early 1970s, the Canadian auto industry's 30 percent labor productivity shortfall relative to its U.S. counterpart had disappeared.

Later, this transformation of the automotive industry was extended to include Mexico. In 1989, Volkswagen consolidated its North American operations in Mexico, shutting down its plant in Pennsylvania. This process continued with the implementation of the North American Free Trade Agreement between the United States, Canada, and Mexico. In 1994, Volkswagen started producing the new Beetle for the entire North American market in that same Mexican plant.

This consolidation in response to closer economic integration with the United States was not limited to the auto industry. Following the implementation of the Canada–U.S. Free Trade Agreement in 1989, each Canadian manufacturing industry experienced a dramatic reduction in its product offerings, concentrating on a smaller number of products (Baldwin, Beckstead, and Caves 2002; Baldwin, Caves, and Gu 2005; Baldwin and Gu 2009; Bernard, Redding, and Schott 2011). Baldwin, Caves, and Gu (2005) also report that the decrease in product offerings was accompanied by substantial increases in production runs for individual products. This process is even evident in the Canadian wine industry, an industry that exclusively produced low-end wines that could not possibly compete with Californian giants such as Gallo. In response to the Agreement, Canadian manufactures dramatically reduced the number of varietals produced and focused on the varietals used to produce ice wine. The industry is now healthier than ever (Beamish and Celly 2003).

Another prominent example of economic integration began in 1957, when the major countries of Western Europe established a free trade area in manufactured goods: the European Economic Community, or EEC. Many politicians offered an old-fashioned Ricardian prediction that German manufacturers would eradicate their European competitors. The facts did not treat such predictions kindly: trade within the EEC grew twice as fast as world trade during the 1960s, and intra-industry trade as a share of EEC trade more than doubled from 1960 to 1990. The benefits of the original European Community agreement were about 1 percent of GDP for the largest economies and about 3 percent of GDP for mid-sized economies such as Belgium (Harrison, Rutherford, and Wooton 1989). (These numbers capture more than just pure love-of-variety gains.) Economic integration has continued in Europe as more countries have joined the free trade area that is today called the European Union and as a subset of EU countries adopted the euro as a common currency in 1999. Euro-zone members have experienced strong trade growth, especially intra-industry trade growth, relative to non-EU countries and even relative to EU countries that have not adopted the euro.

A substantial portion of the increased trade that comes with economic integration also delivers increased product variety to consumers. Balistreri, Hilberrry, and Rutherford (2011) show that the worldwide elimination of all trade barriers would raise the number of varieties available by about 3 percent, lower manufacturing prices by a similar amount, and raise world welfare by 2 percent. Most of these gains would accrue to developing countries such as China. Broda and Weinstein (2006) estimate that the number of products available to U.S. consumers through imports tripled between 1972 and 2001, resulting in welfare gains to U.S. consumers equivalent to a 2.6 percent rise in U.S. GDP. Feenstra (2010, table 2.1) examines how worldwide welfare would change if all countries went from autarky to their 1996 levels of trade. He estimates that the welfare gains from increased varieties are comparable to a 12.5 percent rise in world GDP. While the exact magnitudes of the gains from increased variety differ across studies due to differences in what exactly is being modeled, the clear message is that the gains have been very large for developed countries and continue to be large for developing countries.

Trade expands product variety both in final goods (which benefits consumers) as well as in specialized production inputs (which benefits the firms that use those inputs). Ethier (1982) showed that there is a close parallel between these two. Instead of the love-of-variety that accrues to consumers, firms benefit from the increased productivity derived from an increased range of available production inputs. Recent firm-level research has confirmed this product variety benefit for firms that import intermediate inputs. Using Hungarian data, Halpern, Koren, and Szeidl (2005) show that importing many varieties of foreign inputs increases firm productivity by 12 percent. Using Indonesian data, Amiti and Konings (2007) show that a 10 percentage point fall in input tariffs leads to a productivity gain of 12 percent for firms that import their inputs. Kasahara and Rodrigue (2008), Kasahara and Lapham (2012), Topalova and Khandelwal (2011), and Goldberg, Khandelwal, Pavcnik, and Topalova (2010) show similarly large gains for Chile and India. In the context of the Canada–U.S. Free Trade Agreement, Lileeva and Trefler (2010) find that the fall in Canadian tariffs on inputs that Canadian firms purchased from the United States resulted in a 0.5 percent rise in Canadian manufacturing productivity. The Canadian impacts are not nearly as large as effects from developing countries, which suggests that access to a variety of inputs is an important ingredient in the process of economic development.

More variety means more competition, and more competition forces firms to lower their markups and prices. We see evidence of this after the Turkish and Cote d'Ivoire trade liberalizations of 1985 (Levinsohn 1993; Harrison 1994) and in Belgium during the 1994–2004 period of increased integration (Abraham, Konings, Vanormelingen 2009; De Loecker 2011). On the other hand, there was no evidence of falling markups in Mexico after the trade liberalization of the early 1980s (Tybout and Westbrook 1995). We turn next to the gains associated with reallocation of resources across heterogeneous firms within an industry.

## Gains from Reallocation at the Firm Level

By the mid-1980s, a large body of theoretical work demonstrated that freer trade could affect productivity by forcing firms to move up or down their average cost curves. Much of the follow-on empirical work assumed that firms were identical, and it made a variety of assumptions that allowed inferences to be drawn from industry-level data (for example, Harris 1984). We now know that the heterogeneity of firms even within narrowly defined industries is a central feature of the data that cannot be ignored (for example, Bernard, Jensen, Redding, and Schott 2007).

Our second source of gains from trade is the result of shifting resources away from less-productive firms and towards more-productive firms. To analyze gains from reallocation of trade between firms, we need a model of trade with heterogeneous firms—that is, in which performance varies across different firms. We can then capture how firms with different characteristics respond differently to trade. Consider the case from the previous example where opening to trade leads to a transition from production plan A in which each country produces two varieties to production plan B in which each country produces one variety. In the real world, those varieties are associated with the firms that produce them. Opening up to trade therefore implies that one of the two firms in each country shuts down, while the remaining firm expands production from 2 units to 8 units. But what are the factors explaining which firms expand and which ones exit?

### Monopolistic Competitors with Heterogeneous Costs

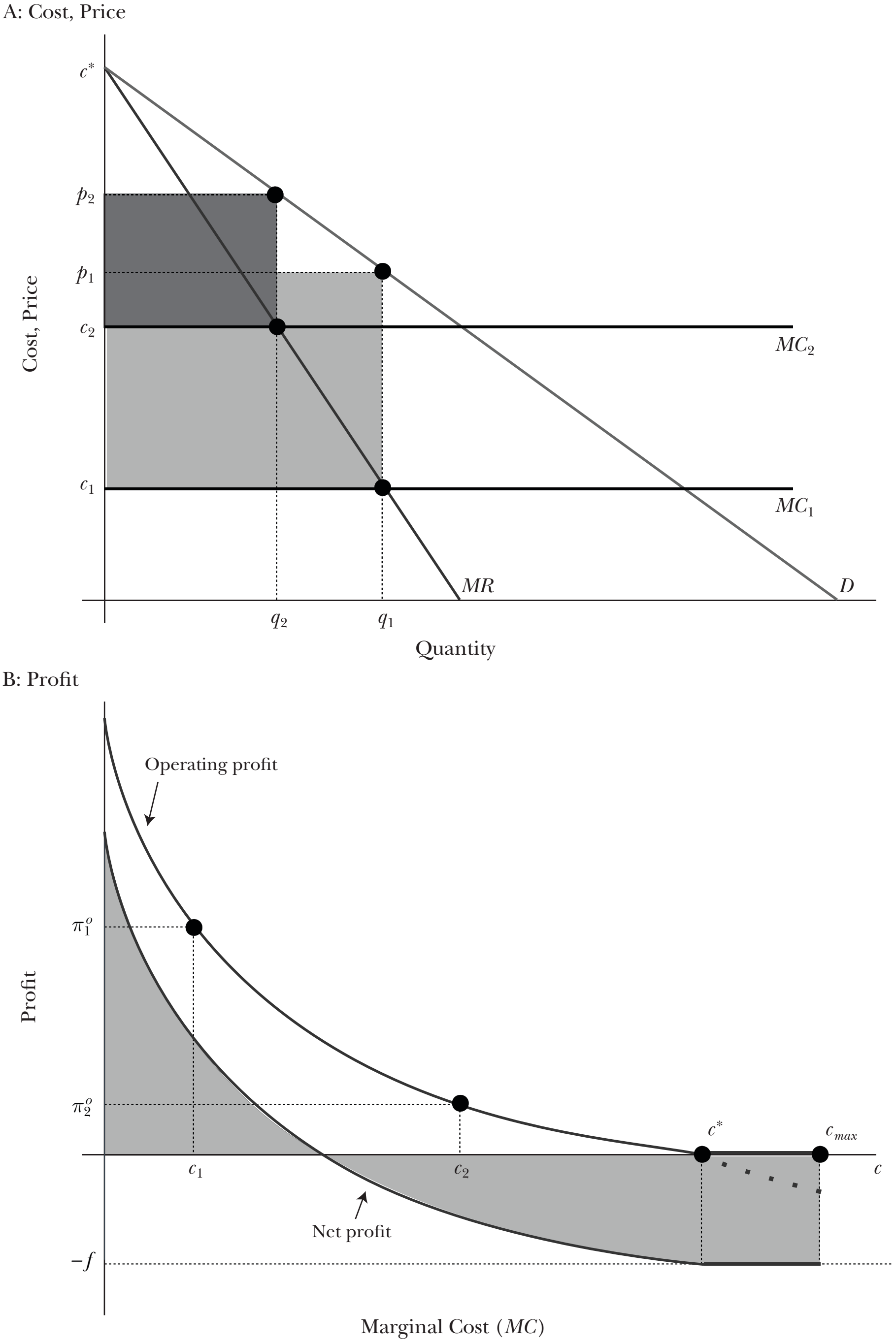
Melitz and Ottaviano (2008) develop a model of trade that allows for differences across firms; we use a simplified version of that model for the discussion here. Consider a monopolistically competitive industry in which many firms compete by offering different products that are relatively close substitutes for one another—at least as compared to products in other industries. For simplicity, we assume that each firm produces a single product, that demand for all products is symmetric, and that firms differ only with respect to productivity. Specifically, firms differ only with respect to their marginal costs of production $c_i$, where $i$ indexes firms. (A number of authors have developed related models that allow firms to produce multiple products: for example, Eckel and Neary 2010; Bernard, Redding, and Schott 2011; and Mayer, Melitz, and Ottaviano 2011. Also, demand need not be symmetric: there can be product-quality differences across firms. Such product-quality differences lead to very similar predictions for firm performance as the ones we now discuss for cost differences.)

Panel A of Figure 2 illustrates the price and quantity choices for two monopolistically competitive firms. Both firms face the same downward-sloping *residual* demand curve: residual demand is demand as perceived by the firm, and thus depends on the behavior of other competing firms in the market.[1] On the production side,

[1] The equation for the demand facing a firm that is used in what follows is $Q = S[(1/n) - b(\overline{p} - p)]$, where $Q$ is the quantity of output demanded, $S$ is the total output of the industry, $n$ is the number of firms in the industry, $b > 0$ is a constant term representing the responsiveness of a firm's sales to its price, $p$ is the price charged by the firm itself, and $\overline{p}$ is the average price charged by its competitors. This demand equation may be given the following intuitive justification: If all firms charge the same price, each will have a market share $1/n$. A firm charging more than the average of other firms will have a smaller market share, whereas a firm charging less will have a larger share.

*Figure 2*

**Performance Differences across Firms**



*Source:* Authors.

marginal costs for firm 1 are shown as lower than those for firm 2. In panel A, firm 1 has a lower marginal cost ($c_1$) than firm 2 ($c_2$). We also assume that economies of scale exist because of a fixed cost that a firm must incur to develop a product and set up its initial production.

In this setting, each firm maximizes profit by choosing an output level $q$ that equalizes marginal cost and marginal revenue. Firm 1 chooses a higher output level than firm 2 ($q_1 > q_2$), associated with a lower price ($p_1 < p_2$). Firm 1 also sets a higher markup than firm 2: $p_1 - c_1 > p_2 - c_2$; this is a consequence of the marginal revenue curve being steeper than the demand curve. Thus, firm 1 earns a higher operating profit than firm 2: $\pi_1^o > \pi_2^o$, as represented by the shaded areas in panel A of Figure 2. We assume that all firms face the same set-up cost $f$, so firm 1 also earns higher net profits (subtracting the fixed cost $f$ for all firms). Of course, differences in fixed costs would not affect marginal costs and thus would not affect firm decisions concerning price and output. We can thus summarize the relevant performance differences that result from marginal cost differences across firms in the following way. Compared to a firm with higher marginal cost, a firm with a lower marginal cost will: 1) set a lower price but at a higher markup over marginal cost, 2) produce more output, and 3) earn higher profits.

Panel B in Figure 2 shows how firm profit varies with its marginal cost $c_i$. Both operating and net profit will be decreasing functions of marginal cost, while the difference between the two is the fixed set-up cost $f$. Going back to panel A, we see that a firm can earn a positive operating profit so long as its marginal cost is below the intercept of the demand curve on the vertical axis. Let $c^*$ denote this cost cutoff. A firm with a marginal cost $c_i$ above this cutoff is effectively "priced out" of the market and would earn negative operating profits if it were to produce any output (represented by the dotted segment for operating profit in panel A). Such a firm would choose to shut down and not produce (earning zero operating profit but incurring a net profit loss $-f$ due to the fixed cost). Why would such a firm enter in the first place? Clearly, it would not if it knew about its high cost $c_i$ prior both to entry and to paying the fixed cost $f$.

We assume that entrants face some randomness about their future production cost $c_i$. This randomness disappears only *after* the set-up cost $f$ is paid and is sunk. Thus some firms will regret their entry decision, as their net profit is negative (they cannot recover the sunk cost $f$). This is the case for firm 2 in panel B; even though its operating profit is positive, it does not cover the sunk cost $f$. On the other hand, some firms discover that their production cost $c_i$ is very low and earn a high (and positive) net profit.

Firms consider all these possible outcomes, captured by the net profit curve in panel B when they make their entry decision. Firms anticipate that there is a range of lower costs where net profits are positive (shaded area to the left above the horizontal axis), and another range of higher costs where net profits are negative (shaded area to the right below the horizontal axis). In the long-run equilibrium, firms enter until their *expected* net profit across all potential cost levels $c_i$ is driven to zero. If every cost level $c_i$ from 0 to $c_{\max}$ is equally likely, then this equilibrium is reached when the two shaded areas are equal.[2] Panel B of Figure 2 summarizes the industry equilibrium for a given market size. It shows which range of firms survive and produce (with cost $c_i$ below $c^*$) and how their profits will vary with their cost levels $c_i$.

## What Changes When Economies Integrate?

How will the situation faced by these heterogeneous firms alter when economies integrate into a single larger market? A larger market can support a larger number of firms than a smaller market, which implies more competition in the larger market. Increased competition—absent any increase in market size—leads to an inward shift of each firm's residual demand curve. On the other hand, holding competition fixed, a larger market rotates out the residual demand curves for all firms. Putting these two effects of increased competition and greater market size together gives us the combined effect of international trade on the residual demand curve perceived by firms. This change is depicted in panel A of Figure 3, as the shift from demand curve $D$ to $D'$. The residual demand curve shifts in from the perspective of the smaller firms with lower output levels that operate on the higher part of the demand curve: here, the effect of tougher competition dominates. However, from the perspective of the larger firms that operate on the lower part of the demand curve, the residual demand curve has shifted out: here, the effect of the larger market size dominates.

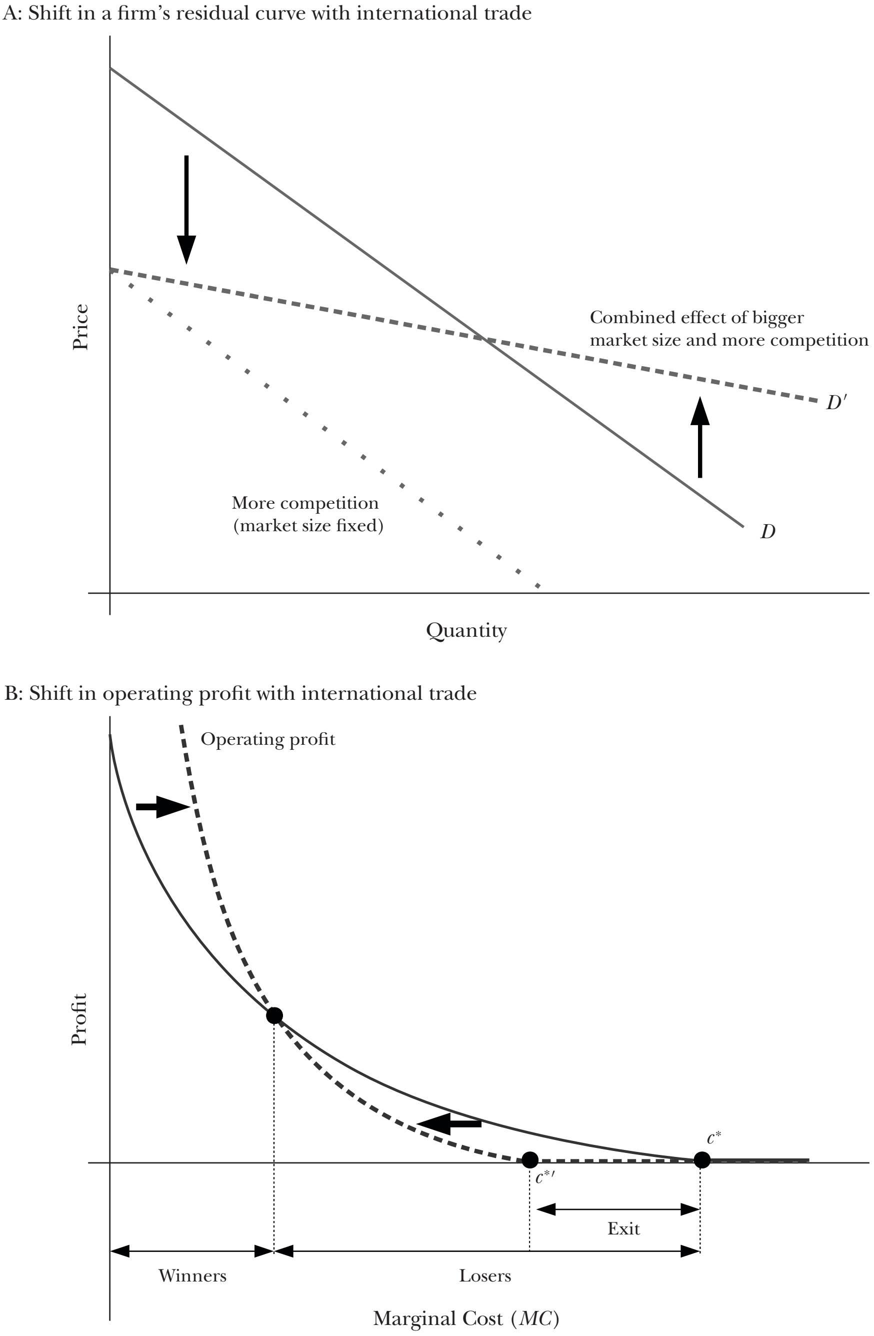
Panel B of Figure 3 shows the consequences of this demand change for the operating profits of firms with different cost levels $c_i$. The decrease in demand for the smaller firms translates into a new lower cost cutoff $c^{*\prime}$: Firms with the highest cost levels (above $c^{*\prime}$) cannot survive the decrease in demand and are forced to exit. On the other hand, the flatter demand curve is advantageous to firms with the lowest cost levels: they can adapt to the increased competition by lowering their markup (and hence their price) and gaining some additional market share. (Recall that the high-cost firms are already setting low markups, and cannot lower their prices to induce positive demand, as this would mean pricing below their marginal cost of production.) Thus, the best-performing firms with the lowest cost levels $c_i$ now earn increased operating and net profits. Panel B of Figure 3 illustrates how increased market size generates both winners and losers amongst firms in an industry. Low-cost firms thrive and increase their profits and market shares, high-cost firms contract, and the highest cost firms exit.

In this model, economic integration through market expansion does not directly affect firm productivity. Nevertheless, it generates an overall increase in aggregate productivity as market shares are reallocated from the low-productivity firms with high marginal costs to the high-productivity ones with low marginal costs.

[2] In contrast, when there is no uncertainty about marginal cost because all firms share the same cost $c$, then entry drives the realized net profit to zero for all firms. With firm heterogeneity, expected net profit is zero, but realized profits will vary as shown in panel B of Figure 2.

*Figure 3*

**Winners and Losers from Market Integration**



A: Shift in a firm's residual curve with international trade

B: Shift in operating profit with international trade

Price
Combined effect of bigger market size and more competition
D′
More competition (market size fixed)
D
Quantity
Operating profit
Profit
c*
c*′
Exit
Winners
Losers
Marginal Cost (MC)

*Source:* Authors.

**Trade Costs, Export Decisions, and Trade Liberalization**

The discussion to this point has implicitly modeled economic integration as a change in market size from a closed economy with no trade all the way to a single combined market with no trade barriers. In reality, initial trade costs are rarely so high as to block all trade prior to liberalization, and liberalization reduces trading costs without fully eliminating them. In a number of ways, this kind of partial trade liberalization has a very similar effect to the simpler case of full integration. With partial trade liberalization, the better-performing firms expand, the worse-performing ones contract, and the worst performing ones exit. This generates the same type of reallocation effect previously described and leads to a rise in aggregate productivity.[3]

However, adding trade costs also allows us to analyze an additional issue: whether firms choose to export. With trade costs, exporting is profitable only for a subset of better-performing firms. Some firms do not export, and instead only serve domestic consumers. We now extend our theoretical model to incorporate trade costs and firms' export decisions. For this purpose, we can no longer analyze a single market: instead, we need to look at firms' decisions in both the domestic and export markets jointly. For simplicity, we consider a special case where both countries are symmetric, so that demand conditions in both the domestic and export markets will be identical.

Assume that a firm must incur an additional trade cost $t$ for each unit of output that it sells to customers across the border. As a result of this trade cost, each firm will set a different price in its export market relative to its domestic market, which will lead to different quantities sold in each market, and ultimately to different profit levels earned in each market. Because we are assuming that each firm's marginal cost is constant and does not vary with production levels, the decisions regarding pricing and quantity sold in each market can be separated: a decision regarding the domestic market will have no effect on the profitability of different decisions for the export market.

Consider the case of firms located in Home. Their situation regarding their domestic (Home) market is exactly as was illustrated in Figure 2, except that all the outcomes such as price, output, and profit relate to the domestic market only. Now consider the export (Foreign) market. The firms face the same demand curve in Foreign as they do in Home (the two countries are identical). The only difference is that each firm's marginal cost in the export market is shifted up by the trade cost $t$. What are the effects of the trade cost on the firms' decisions regarding the export market? A higher marginal cost induces a firm to raise its price, which leads to a lower output quantity sold and to lower profits (as highlighted in Figure 2).

[3] The more general version of this model analyzed by Melitz and Ottaviano (2008) allows for multiple countries of different sizes and for arbitrary trade costs between any country pair (though the trade costs are proportional to production costs instead of per output unit as in the current version). That paper shows more formally that the effects of multilateral liberalization (all countries proportionally reduce trade costs) are very similar to the case of full economic integration that leads to a single larger market.

We also know that if marginal cost is raised above a threshold level $c^*$, then a firm cannot profitably operate in that market. Thus, when there are trade costs, some firms will find it profitable to operate in the domestic market but not in the export market because the trade cost pushes their marginal cost for that market above the threshold $c^*$.

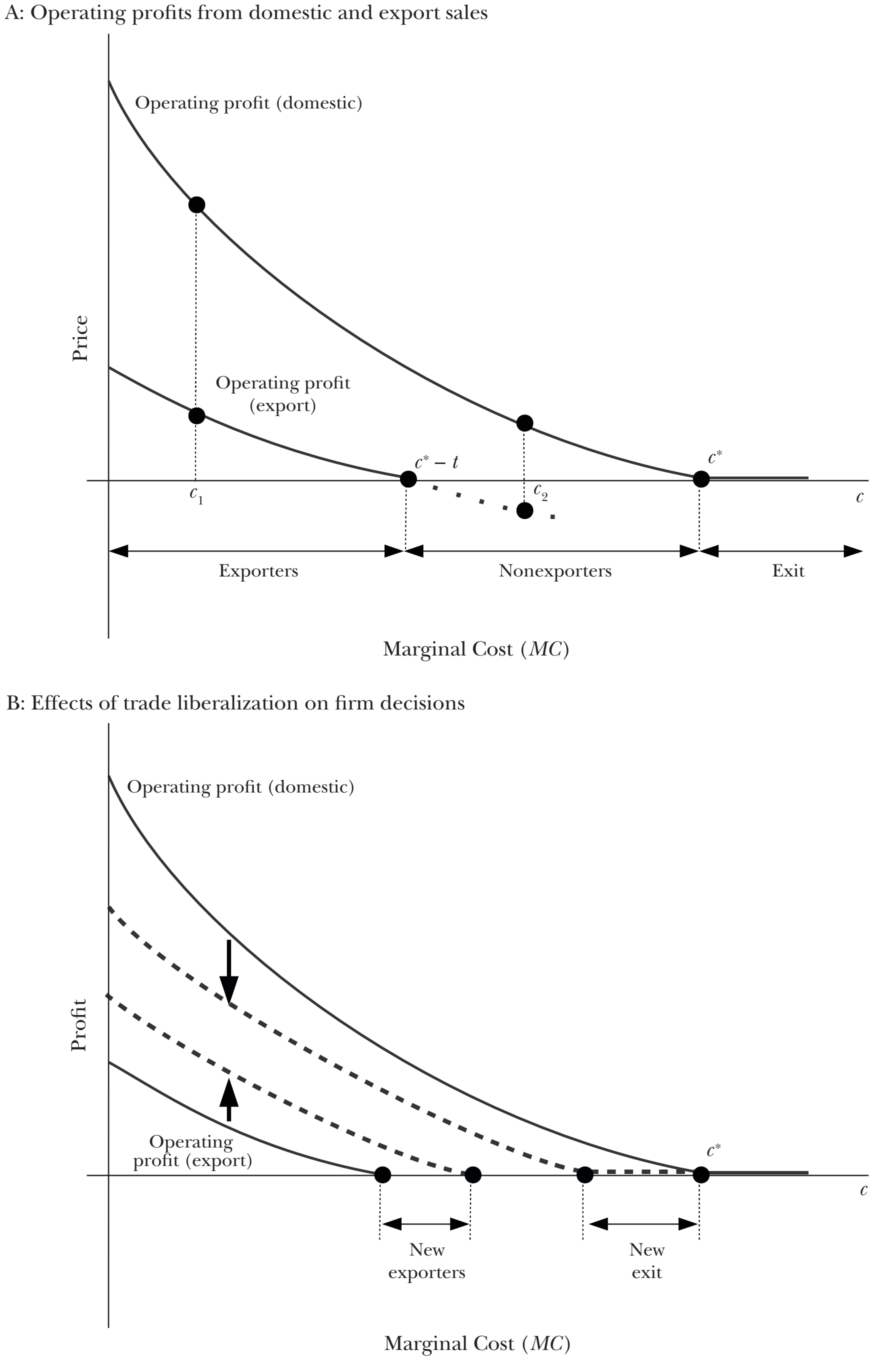
Figure 4 helps to visualize the production and export decisions for all firms based on their marginal cost $c_i$. Panel A of Figure 4 separates a firm's operating profit into a portion earned from domestic sales, and a portion earned from export sales. (Both portions are functions of a firm's marginal cost $c_i$ as in Figure 2.) Because the only difference between the domestic and export markets is the additional per-unit trade cost $t$, the horizontal distance between the two curves is equal to the trade cost $t$. Firm 1 earns positive operating profits from sales in both the domestic and export markets: it will export and reach consumers in both markets. This will be the case for all firms with cost below $c^* - t$. On the other hand, firm 2 only earns positive operating profits from sales in the domestic market—and thus chooses not to export. Any firm with cost above $c^* - t$ will be in this same situation and therefore will not export: those firms only serve their domestic market. As before, the worst performing firms with cost above $c^*$ cannot profitably operate at all (even in their domestic market) and therefore exit.

Panel B of Figure 4 summarizes the effects of trade liberalization—a reduction in the trade cost $t$—for those firm decisions. The figure shows the same two operating profit curves from panel A both before and after (dashed curves) trade liberalization. The operating profit for the domestic market shifts down due to the increase in competition (which shifts the residual demand curve for the domestic market inward as explained earlier). Some of the higher-cost firms that used to produce for domestic consumption no longer earn a positive operating profit after trade liberalization and exit. On the other hand, the operating profit for the export market shifts up due to the lower trade cost. (Increased competition in the export market tends to reduce operating profits there, but this effect is dominated by the direct effect of the trade cost reduction.) A key empirical prediction is that some firms start exporting. Specifically, among the firms that did not export prior to trade liberalization, only the most productive of these start exporting.

**Evidence on Gains from Inter-Firm Reallocations**

In many ways, the Canada–U.S. Free Trade Agreement is a useful natural experiment for considering the effects of trade integration. The policy experiment is clearly defined: it dealt only with market integration and was not part of a larger package of macroeconomic reforms that often accompany trade liberalization. The enactment of the agreement was largely unanticipated: a Canadian general election was fought on the issue one month before the agreement was to be signed into law and pollsters unanimously predicted that Canada's ruling party—along with the free trade agreement—would be defeated (Brander 1991; Thompson 1993). Thus, evidence about the extent of aggregate productivity changes as a result of reallocations among heterogeneous firms can be sought by looking at the distribution of

*Figure 4*
**Export Decision and Trade Liberalization**



A: Operating profits from domestic and export sales

B: Effects of trade liberalization on firm decisions

*Source:* Authors.

productivity across Canadian manufacturing plants before and after the agreement, at entrants before and after the agreement, and at the productivity distribution of exporters and nonexporters.

The agreement came into effect on January 1, 1989. Panel A of Figure 5 shows the distribution of labor productivity as measured by value-added per employee across Canadian manufacturing plants both before the agreement in 1988 and in 1996, when there had been time for firm adjustments to occur. For example, the 1996 curve summarizes the productivity distribution of all 35,000 Canadian manufacturing plants in that year. Clearly, the distribution of firms shifted rightward: between 1988 and 1996, the share of low-productivity plants in manufacturing declined and the share of high-productivity plants rose.

The horizontal axis is based on a measure of the log of labor productivity. However, to ensure that dispersion is driven by within-industry rather than between-industry differences in labor productivity, we scale each plant's log productivity by subtracting from it the median log productivity of the plant's four-digit SIC industry. Thus, the median plant in each industry has a score of zero on the horizontal axis. The vertical axis shows the share of plants with the indicated level of productivity. These frequencies are weighted by plant employment; otherwise, tiny plants that account for only a tiny fraction of total employment would dominate the figure.

To get a sense of the degree of productivity dispersion, consider the horizontal axis of Figure 5 and suppose that log productivity at plant A is one unit higher than at plant B. This is equivalent to saying that plant A is three times more productive than plant B. If A is four units higher than B, then A is 50 times more productive than B.[4]

Obviously, labor productivity as shown in Figure 5 is not an identical concept to the horizontal lines showing levels of marginal cost in the theoretical discussion. When marginal costs are low, we typically expect productivity to be high. Therefore, the inverse of marginal costs ($1/c$) is often proxied in empirical work by productivity.

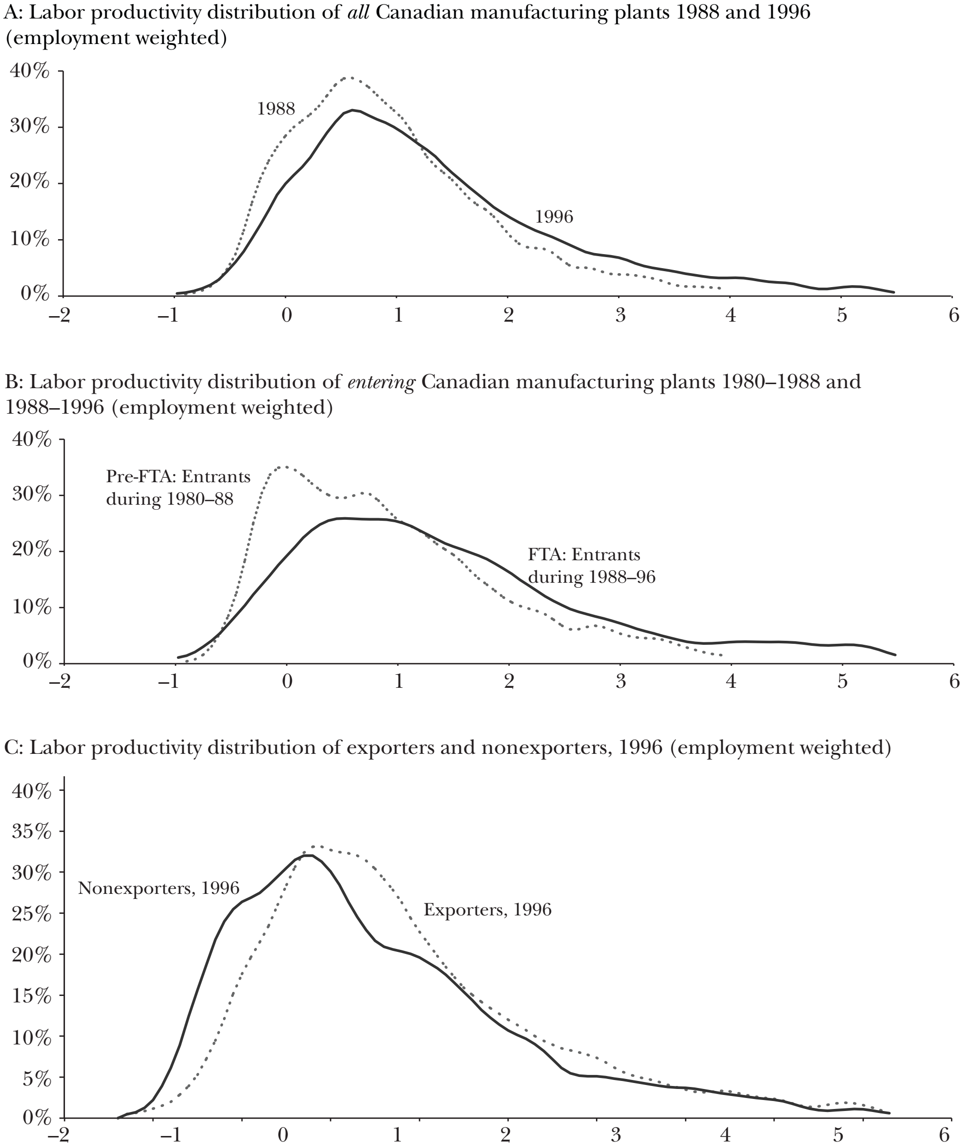
The productivity heterogeneity shown for Canadian manufacturing firms in Figure 5 is a pervasive feature of all economies including, for example, the United States (Bernard, Eaton, Jensen, and Kortum 2003), many European economies (Mayer and Ottaviano 2008; Bartelsman, Hatiwanger, and Scarpetta 2009), as well as China and India (Hsieh and Klenow 2009). Wagner (2007) surveys related studies from countries all around the world, reporting similar patterns regarding widespread firm heterogeneity within industries. Thus, the lessons derived from this example are not specific to the Canadian experience.

What caused the change from 1988 to 1996 in the productivity distribution of Canadian manufacturing firms? It was largely due to the reallocation mechanisms across plants described above. The first of these mechanisms that we examine is

[4] Let $\varphi_A$ and $\varphi_B$ be productivities of $A$ and $B$ and suppose that they are 1 unit apart i.e., $\ln(\varphi_A) - \ln(\varphi_B) = 1$. From the property of logs, $\varphi_A/\varphi_B = e^1 = 2.7 \approx 3$. For a difference of 4 units, $e^4 \approx 50$. On a more technical level, the figure is constructed starting with "standardized" log productivities (see the formula in Lileeva 2008, p. 369), which we then multiply by a single scale factor to transform the standardized log productivities into log productivities that are directly comparable with log productivities reported in studies from other countries.

*Figure 5*

**Distribution of Productivity across Canadian Manufacturing Plants before and after the Canada–U.S. Free Trade Agreement**



A: Labor productivity distribution of *all* Canadian manufacturing plants 1988 and 1996 (employment weighted)

B: Labor productivity distribution of *entering* Canadian manufacturing plants 1980–1988 and 1988–1996 (employment weighted)

C: Labor productivity distribution of exporters and nonexporters, 1996 (employment weighted)

*Source:* Authors' calculations using data from the Canadian Annual Survey of Manufactures.
*Notes:* The horizontal axes are based on a measure of the log of labor productivity. However, to ensure that dispersion is driven by within-industry rather than between-industry differences in labor productivity, we scale each plant's log productivity by subtracting from it the median log productivity of the plant's four-digit SIC industry. Thus, the median plant in each industry has a score of zero on the horizontal axis. The vertical axis shows the share of plants with the indicated level of productivity. These frequencies are weighted by plant employment; otherwise, tiny plants that account for only a tiny fraction of total employment would dominate the figure.

the fall in the survival threshold of marginal cost—that is, the leftward shift of $c^*$ in panel B of Figure 4. The empirical counterpart to a fall in $c^*$ is a rise in the break-even level of productivity. One can examine this mechanism by looking either at exit rates or at entry rates. Since a plant may not exit until it has completed the multiyear process of depreciating its fixed capital, it is best to look at entry rates, which adjust more quickly to shocks such as a free trade agreement. Panel B of Figure 5 displays the productivity levels of new entrants to Canadian manufacturing both for the pre-agreement period (1980–1988) and for the period immediately after the agreement came into force (1988–1996). There was a striking decline in the entry rates of plants with productivity near or below the median. To use a sports analogy, in the earlier period many low-productivity plants made the cut and joined the team while in the later period a number of such low-productivity plants no longer made the cut.[5]

The pattern here is confirmed by more formal econometric analysis (specifically, binary-outcome regressions of exit on plant characteristics as well as industry and time controls). For example, Baggs (2005) and Baldwin and Gu (2006) show that the free trade agreement tariff cuts raised exit by a large amount. Lileeva (2008) estimates that the free trade agreement tariff cuts raised exit rates by as much as 16 percent, with all of the increase involving the exit of nonexporters. Bernard, Jensen, and Schott (2006) find similar results for U.S. plants faced with U.S. tariff reductions.

### Trade Costs and the Export Decision

A central prediction of the theory is that in the presence of trade costs, only low-cost, high-productivity firms export. Panel C of Figure 5 shows the distribution of Canadian plants separately for exporters and nonexporters. Clearly, the distribution for exporters is to the right of that for nonexporters. On average, Canadian exporters are 40 percent more productive than nonexporters in the same industry (Baldwin and Gu 2003). Since the seminal work of Bernard and Jensen (1995), a huge body of research covering dozens of countries has found this same pattern of higher productivity for exporters relative to nonexporters.[6]

[5] Bernard and Jensen (1999), Trefler (2004), Lileeva (2008), and Lileeva and Trefler (2010) all point out that one must look not just at pre-Free Trade Agreement *levels* (as in Figure 5), but also at pre-FTA *trends*. All of the FTA results reported here hold with pre-FTA controls for both levels and trends. For example, variants of some of the panels in Figure 5 with pre-FTA trend controls appear in Lileeva (2008).

[6] In examining panel C of Figure 5, a critical reader may wonder why there are so many highly productive nonexporters and whether this contradicts the theory. A simple but prominent example may help to explain. Highly productive auto parts plants often cluster around a giant auto assembly plant—Ford, General Motors, and Honda all have major auto assembly plants near Toronto, Canada, that are surrounded by parts suppliers. These parts suppliers are highly productive, but do not directly export. This pattern is clearly not a challenge to the theory: these highly productive plants are supplying parts that are assembled into the autos that are ultimately exported to the United States: highly productive parts suppliers are "indirect" exporters. Of course, there are surely other factors outside the scope of our model that explain why some very productive firms do not export—and conversely—why some low-productivity firms export.

A much more demanding prediction of the theory deals with who will *start exporting in response to falling trade costs.* Panel B of Figure 4 shows that those who start exporting *will* be among the most productive of those who never exported before. To test this prediction, Lileeva and Trefler (2010) examined a sample of over 5,000 Canadian manufacturing plants that had never exported prior to the Canada–U.S. free trade agreement. A very large percentage of these plants (40 percent) started exporting after the agreement came into force on January 1, 1989. Lileeva and Trefler examine whether these plants started exporting because of the U.S. tariff cuts and, more importantly, whether those that started exporting because of the tariff cuts were more productive than nonexporters. To this end, Lileeva and Trefler divide up their sample into quartiles of the 1988 distribution of labor productivity (with the quartiles defined separately for each industry to net out industry characteristics). Only 20 percent of the plants in the bottom quartile of labor productivity started exporting because of the tariff cuts, compared to nearly 60 percent of the plants from the top quartile of labor productivity. (These estimates are from a probit regression in which the dependent variable is 1 if the plant started exporting and 0 if the plant remained a nonexporter. The key independent variable is a plant-specific measure of the change in the U.S. tariff. This measure is described below.) The key conclusion is that, among firms that did not export before trade liberalization, the most productive of these were three times more likely to start exporting in response to the U.S. tariff cuts. This is as predicted in panel B of Figure 4.[7]

### Quantifying the Gains from Trade Due to Reallocation across Heterogeneous Plants

In the wake of the Canada–U.S. free trade agreement, Canadian manufacturing productivity rose sharply. We have shown that part of this productivity gain was due to the reallocation mechanisms highlighted by the theory. But how important were these in quantitative terms for productivity growth and overall welfare?

In the conventional approach to estimating the gains from trade, the focus is on measuring welfare, or more specifically, on the income a society would be willing to pay for lower tariffs. These "compensating variation" gains are typically derived from models that 1) make a large number of parametric assumptions (assuming very specific functional forms for preferences that determine the extent of product differentiation, as well as for the utility derived from love-of-variety), and 2) make use of parameter estimates about which we are highly uncertain. In short, a lot of uncertainty surrounds welfare-gain estimates. In the heterogeneous-firms literature, the focus has shifted to estimating productivity gains rather than welfare. The last two decades have seen major improvements in our ability to estimate productivity gains, both because of the creation of

[7] On a related note, profits play a key role in all the mechanisms of our model. Baggs and Brander (2006) confirm that profits move in the expected directions. In particular, they find that falling Canadian tariffs are associated with declining Canadian profits, especially for import-competing firms, while falling U.S. tariffs are associated with increasing Canadian profits, especially for export-oriented firms.

high-quality, firm-level longitudinal data and because of methodological developments aimed at exploiting these data. Thus, although productivity gains are not the same as welfare gains, we have much greater confidence in our estimates of the productivity gains associated with freer trade.

The productivity gains associated with the reallocation of market shares across firms following the Canada–U.S. free trade agreement are usefully broken into two components. First, *the fall in the U.S. tariffs* allowed Canadian plants to export more. This shifted the composition of output towards high-productivity exporters and away from low-productivity nonexporters. Lileeva and Trefler (2010) estimate that the fall in U.S. tariffs causally raised Canadian manufacturing productivity by 4.1 percent via this export-composition channel. Second, *the fall in the Canadian tariffs* led to a shift in domestic market shares—exporters gained market share at the expense of nonexporters. In the extreme, many nonexporters simply went out of business. Trefler (2004) calculates that this selection effect increased overall Canadian manufacturing productivity by 4.3 percent.[8]

Putting these numbers together, we see that the reallocation and selection effects induced by the free trade agreement generated a productivity increase of 8.4 percent (4.1 + 4.3) for Canadian manufacturing. This represents a massive productivity increase in just a short time—especially when one considers that this productivity gain did not come from productivity improvements at the plant level: it only came from the shifting of market shares from less- to more-productive plants.

Canada is not the only country to have experienced such a substantial productivity boost from reallocations driven by trade liberalization. Bernard and Jensen (2004) find that almost half of all U.S. manufacturing productivity growth during 1983–1992 is explained by the reallocation of resources towards exporters. Pavcnik (2002) studies the response of the Chilean manufacturing sector to a massive trade liberalization episode that took place from 1979 to 1986. She finds that two-thirds of the ensuing 19 percent increase in productivity (another example of a massive increase in aggregate productivity) was generated by composition changes within industries due to a reallocation of market shares towards more-efficient producers. Surveys by Greenaway and Kneller (2007) and Wagner (2007) summarize the connections between trade liberalization and aggregate productivity—including this reallocation effect across heterogeneous firms—for a wide range of studies and countries.

[8] Specifically, Trefler (2004) regressed labor productivity growth in the period after the free trade agreement (relative to the pre–agreement period) on U.S. and Canadian tariff cuts mandated by the agreement. He then showed that the *Canadian tariff cuts* raised productivity at the industry level, but not at the plant level. This means that the gains in productivity were coming from selection, rather than from improvements at the plant level. Using this approach, he finds that the free trade agreement raised Canadian manufacturing labor productivity by 5.8 percent of which 4.3 percent was due to the exit associated with the Canadian tariff cuts.

## Gains from Rising Within-Plant Productivity

In this section, we move from this "between-plant" effect in which productive plants expand at the expense of less-productive plants to our third source of gains from trade: a "within-plant" effect in which trade raises productivity of individual plants by raising the returns to innovation.

At least as far back as Schmookler (1954), economists have known that the larger the market, the more profitable it is for firms to invest in productivity-enhancing activities. Firms in large markets have the large sales volumes needed to justify incurring the high fixed costs of innovation. The U.S. tariff cuts that were part of the U.S.–Canada free trade agreement greatly increased the size of the market faced by Canadian firms. It should therefore have encouraged Canadian firms to increase their exporting *and* to increase their investments in productivity-enhancing technologies. We start here with a short extension to the theoretical model that captures how larger markets generate incentives for some firms to innovate, and then turn to empirical evidence.

### A Theory of Market Size and Firm Innovation

Suppose that an innovation process requires an up-front fixed cost $f_I$ and in return generates a *reduction* in marginal cost $\Delta c_I$. That is, innovation reduces marginal cost from $c$ to $c - \Delta c_I$. A firm that produces $q$ units of output and engages in innovation will lower its production costs by $q \times \Delta c_I$. The firm will weigh this cost saving against the fixed innovation cost $f_I$, and innovate if $q \times \Delta c_I > f_I$ or

$$q > \frac{f_I}{\Delta c_I}.$$

In words, only firms with large volumes $q$ (that is, those with initially lower levels of marginal cost) will find it profitable to innovate. What happens to this firm-level innovation decision when trade is liberalized? Lower trade costs increase an exporter's sales in the export market, and thus increase the exporter's overall output level $q$. For some exporters, this increase in output will tip the balance in favor of innovating. For some nonexporters, trade liberalization will tip the balance in favor of exporting *and* innovating.

### Evidence on Within-Firm Productivity Growth and Trade

For evidence on the link from growth of trade to within-firm productivity, we turn again to Canada's experience with the free trade agreement. Lileeva and Trefler (2010) look at their sample of 5,000 Canadian manufacturing plants that did not export prior to 1988 and divide these plants into those that started exporting after the passage of the free trade agreement and those that did not. In the raw data, the labor productivity of those that started to export rose 29 percent more than for nonexporters; starting to export was highly correlated with within-plant productivity growth. Of course, this 29 percent number does not take into account a serious problem of reverse causality: does exporting lead to increased productivity or does increased productivity lead to exporting?

To answer this question, one needs an instrument for exporting: that is, one needs an event that causes a firm to export but that does not directly affect its productivity growth. As Lileeva and Trefler (2010) show, "plant-specific" tariff cuts fit the bill as an instrument. Consider a single Canadian plant called Lumberjack and the many products it produces. Empirically, products are defined very narrowly, at the six-digit level of the Harmonized System product classification, so that there are thousands of products in manufacturing. For each product produced by Lumberjack, one can calculate the U.S. tariff cut. Averaging these tariff cuts across all of Lumberjack's products yields a plant-specific tariff cut. This plant-specific tariff cut has enormous power in predicting whether a Canadian plant starts exporting and how much it exports. The tariff cut also has no direct impact either theoretically or empirically on a plant's productivity growth. It is thus a novel and valid instrument.

Lileeva and Trefler (2010) actually do something fancier than instrumental variables—they estimate the local average treatment effect (LATE). This is the effect on productivity of starting to export *for those plants that started exporting because of the tariff cuts.* Thus, unlike previous studies of the causal impacts of exporting on productivity, their work only uses information drawn from plants that were likely to be affected by the free trade agreement. Using their plant-specific tariff instrument and the local average treatment effect estimator, they establish that the free trade agreement *caused* the productivity of new exporters to rise by 15.3 percent. Since this 15.3 percent rise occurred in plants that accounted for 23 percent of Canadian manufacturing output, the 15.3 percent rise in labor productivity of new exporters raised *overall* Canadian manufacturing productivity by 3.5 percent ($3.5 = 15.3 \times 0.23$; see Table 2 below).

Lileeva and Trefler (2010) then trace the sources of this productivity gain back to investments in productivity. The authors examine advanced manufacturing technologies and rates of innovation at these plants. Table 1 presents the results. Consider the first row, which deals with management techniques essentially associated with lean manufacturing. In the period immediately after implementation of the free trade agreement, 16 percent of new exporters adopted such techniques, 10 percentage points more than nonexporters did. The final column, which reports local average treatment effect (LATE) estimates, shows that 7 of the 10 percentage points was attributable to the effects of increased exporting resulting from the U.S. tariff cuts. As shown in Table 1, similar results hold for the adoption of other technologies and for innovation.

These results break with the discussion of Bernard, Jensen, Redding, and Schott (2007) in this journal, who correctly argue that most careful studies show exporting *does not* raise productivity. Over the years, however, a few careful studies have found otherwise, as in Canada (Baldwin and Gu 2003, 2004; Lileeva 2008), in nine African countries (Van Biesebrock 2005), and in Slovenia (De Loecker 2007). López (2005) provides an overall survey.

*Table 1*
**Innovation Response to Free Trade Agreement by New Exporters**

| | *Raw adoption and innovation rates* | | | *LATE* |
|---|---|---|---|---|
| | *New exporters* | *Nonexporters* | *Difference* | *Difference* |
| Manufacturing information systems | 16% | 6% | 10% | 7% |
| Inspection and communications | 18% | 10% | 8% | 8% |
| Any product or process innovation | 30% | 20% | 10% | 8% |
| Any product innovation | 26% | 14% | 12% | 7% |

*Source:* Lileeva and Trefler (2010).
*Note:* "LATE" is the local average treatment effect.

What has recently buttressed the minority view that a rise in exporting can lead to a rise in productivity is a spate of papers isolating the causal mechanisms through which exporting affects productivity. We have already seen the market-size mechanism of Lileeva and Trefler (2010). Bustos (2011a, b) develops a related model of scale-biased technology choice, which she takes to Argentinean firm-level data for the 1992–1996 period. Bustos (2011b, table 6) shows that firms that began exporting between 1992 and 1996 also increased their technology spending. Bustos (2011a) shows that technology spending increased most in sectors that experienced improved access to Brazilian product markets (through the tariff cuts in the Mercosur regional trade agreement). In a series of studies of Taiwanese electronics exporters, Aw, Roberts, and Winston (2007) and Aw, Roberts, and Xu (2008, 2011) show empirically that there is a complex dynamic interplay between the decisions to export and conduct research and development, with today's decisions about one affecting tomorrow's decisions about the other—and both affecting future productivity. Extending this dynamic methodology to general equilibrium, Shen (2011) also finds strong complementarities between exporting and productivity-enhancing investments among Spanish firms. Bloom, Draca, and Van Reenen (2011) show that import competition from China has led to increases in R&D, patenting, information technology, and total factor productivity among more technologically advanced European firms. Atkeson and Burstein (2010) are the only ones to examine exporting and investing in productivity within a full-blown general equilibrium analysis. They find the general equilibrium feedbacks are important.

## A New Dimension of Heterogeneity

In our theoretical model above, firms below a certain productivity threshold should not be exporting. Yet in the empirical work reviewed above, we saw that many low-productivity Canadian plants started exporting in response to U.S. tariff cuts. There is a second puzzle that we have not yet noted: Lileeva and Trefler (2010, table 3) report that the plants that gained most from starting to export (both in terms of productivity gains and increased innovation) were primarily plants that initially had low productivity. That is, among plants that started to export, the benefit was greatest for the least-productive plants.

To see why, consider a firm that is just indifferent between investing and not investing. From the earlier equation, indifference means that $q = f_I / \Delta c_I$, where $\Delta c_I$ is the reduction in marginal cost or the increase in productivity. Rearranging ($\Delta c_I = f_I / q$) and noting that sales $q$ are increasing in initial productivity, we arrive at a simple conclusion. *Among the set of firms that are just indifferent between innovating and not innovating*, the less-productive, low-$q$ firms must expect larger productivity gains $\Delta c_I$ from innovation. Lileeva and Trefler's (2010) results strongly confirm this prediction.

## Conclusions

Recent research into the welfare gains from intra-industry trade have focused on three sources of gains: 1) gains from increased variety and economies of scale, 2) productivity gains at the industry level from shifting resources away from low-productivity firms and towards high-productivity firms, and 3) productivity gains at the firm level from innovating for a larger market. Each of these mechanisms have proven to be highly important empirically in the context of the exhaustively studied Canada–U.S. free trade agreement, and also appear important in many other less-studied contexts. Indeed, Balistreri, Hillberry, and Rutherford (2011) show that adding firm heterogeneity to standard computable equilibrium models of trade raises the gains from trade liberalization by a multiple of four. Empirical confirmation of the gains from trade predicted by models with heterogeneous firms represents one of the truly significant advances in the field of international economics.

We summarize the causal effects of the free trade agreement on overall Canadian manufacturing productivity in Table 2. As the last row shows, Canadian manufacturing labor productivity rose by 13.8 percent. The idea that a single government policy could raise productivity by such a large amount and in such a short time-span is truly remarkable.

In writing this review, we have focused on the net gains from trade. Yet the model we have developed highlights how intra-industry trade will generate both winners and losers. For example, in the context of the Canada–U.S. free trade agreement, Trefler (2004) shows that 12 percent of the workers in low-productivity firms lost their jobs. Recent research suggests that American workers are similarly struggling in response to the Chinese import surge (Liu and Trefler 2011; Autor, Dorn, and Hanson 2011). Clearly, this suggests an important role for policies that provide an adequate safety net and transitional assistance for those affected workers. The blow to those workers could be cushioned by policies that impede the reallocation process across firms. However, such policies—as opposed to policies that provide some form of direct assistance to the affected workers—would also entail a substantial long-run cost. After all, it is precisely this reallocation process that generates some of the long-run gains that we have described. In addition, policies that seek to impede the

*Table 2*
**Overall Effect of Free Trade Agreement on Canadian Manufacturing Productivity**

| | |
|---|---|
| Selection/reallocation (between plants) | |
| Growth of exporters (most-productive plants) | 4.1% |
| Contraction and exit of least-productive plants | 4.3% |
| Within-plant growth | |
| New exporters invest in raising productivity | 3.5% |
| Existing exporters invest in raising productivity | 1.4% |
| Improved access to U.S. intermediate inputs | 0.5% |
| **Total** | **13.8%** |

*Source:* Trefler (2004) and Lileeva and Trefler (2010).

reallocation process by making firm contractions and expansions costlier would also reduce the potential gains to firm innovation and hence lead to less innovation and further depress the potential long-run gains from trade. Nonetheless, it is important to remember that there are winners and losers from trade liberalization not just among firms, but also among their employees.

■ *Trefler gratefully acknowledges funding from the Social Sciences and Humanities Research Council of Canada (SSHRC) and the Canadian Institute for Advanced Research (CIFAR). His research would not have been possible without the tremendous support of John Baldwin and Alla Lileeva. We thank the editors (David Autor, Chad Jones, John List, and Timothy Taylor) for their invaluable comments.*

## References

**Abraham, Filip, Jozef Konings, and Stijn Vanormelingen.** 2009. "The Effect of Globalization on Union Bargaining and Price-Cost Margins of Firms." *Review of World Economics (Weltwirtschaftliches Archiv)* 145(1): 13–36.

**Amiti, Mary, and Jozef Konings.** 2007. "Trade Liberalization, Intermediate Inputs, and Productivity: Evidence from Indonesia." *American Economic Review* 97(5): 1611–38.

**Antràs, Pol., and Esteban Rossi-Hansberg.** 2009. "Organizations and Trade." *Annual Review of Economics* 1(1): 43–64.

**Atkeson, Andrew, and Ariel Tomás Burstein.** 2010. "Innovation, Firm Dynamics, and International Trade." *Journal of Political Economy* 118(3): 433–84.

**Autor, David H., David Dorn, and Gordon H. Hanson.** 2011. "The China Syndrome: Local

Labor Market Effects of Import Competition in the United States." August. http://economics.mit.edu/files/6613.

**Aw, Bee Yan, Mark J. Roberts, and Tor Winston.** 2007. "Export Market Participation, Investments in R&D and Worker Training, and the Evolution of Firm Productivity." *World Economy* 30(1): 83–104.

**Aw, Bee Yan, Mark J. Roberts, and Daniel Yi Xu.** 2008. "R&D Investments, Exporting, and the Evolution of Firm Productivity." *American Economic Review* 98(2): 451–56.

**Aw, Bee Yan, Mark J. Roberts, and Daniel Yi Xu.** 2011. "R&D Investment, Exporting, and Productivity Dynamics." *American Economic Review* 101(4): 1312–44.

**Baggs, Jen.** 2005. "Firm Survival and Exit in Response to Trade Liberalization." *Canadian Journal of Economics* 38(4): 1364–83.

**Baggs, Jen, and James Brander.** 2006. "Trade Liberalization, Profitability, and Financial Leverage." *Journal of International Business Studies* 37(2): 196–211.

**Baldwin, John R., Desmond Beckstead, and Richard Caves.** 2002. "Changes in the Diversification of Canadian Manufacturing Firms (1973–1997): A Move to Specialization." Statistics Canada, Analytical Studies Branch Research Paper 2002179.

**Baldwin, John R., Richard E. Caves, and Wulong Gu.** 2005. "Responses to Trade Liberalization: Changes in Product Diversification in Foreign- and Domestic-Controlled Plants." Chap. 10 in *Governance, Multinationals and Growth,* edited by Lorraine Eden and Wendy Dobson. Cheltenham, UK: Edward Elgar Publishing.

**Baldwin, John R., and Wulong Gu.** 2003. "Export-Market Participation and Productivity Performance in Canadian Manufacturing." *Canadian Journal of Economics* 36(3): 634–57.

**Baldwin, John R., and Wulong Gu.** 2004. "Trade Liberalization: Export-Market Participation, Productivity Growth and Innovation." *Oxford Review of Economic Policy* 20(3): 372–92.

**Baldwin, John R., and Wulong Gu.** 2006. "Plant Turnover and Productivity Growth in Canadian Manufacturing." *Industrial and Corporate Change* 15(3): 417–65.

**Baldwin, John R., and Wulong Gu.** 2009. "The Impact of Trade on Plant Scale, Production-Run Length and Diversification." Chap. 15 in *Producer Dynamics: New Evidence from Micro Data,* edited by Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts. University of Chicago Press.

**Balistreri, Edward J., Russell H. Hillberry, and Thomas F. Rutherford.** 2011. "Structural Estimation and Solution of International Trade Models with Heterogeneous Firms." *Journal of International Economics* 83(2): 95–108.

**Bartelsman Eric J., John C. Haltiwanger, and Stefano Scarpetta.** 2009. "Cross-Country Differences in Productivity: The Role of Allocation and Selection." NBER Working Paper 15490.

**Beamish, Paul W., and Nikhil Celly.** 2003. *Vincor and the New World of Wine.* HBR Case Study, *Harvard Business Review.*

**Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum.** 2003. "Plants and Productivity in International Trade." *American Economic Review* 93(4): 1268–90.

**Bernard, Andrew B., and J. Bradford Jensen.** 1995. "Exporters, Jobs, and Wages in U.S. Manufacturing: 1976–1987." *Brookings Papers on Economic Activity* Microeconomics: 67–112.

**Bernard, Andrew B., and J. Bradford Jensen.** 1999. "Exceptional Exporter Performance: Cause, Effect, or Both?" *Journal of International Economics* 47(1): 1–25.

**Bernard, Andrew B., and J. Bradford Jensen.** 2004. "Exporting and Productivity in the U.S." *Oxford Review of Economic Policy* 20(3): 343–57.

**Bernard, Andrew B., J. Bradford Jensen, Stephen J. Redding, and Peter K. Schott.** 2007. "Firms in International Trade." *Journal of Economic Perspectives* 21(3): 105–130.

**Bernard, Andrew B., J. Bradford Jensen, and Peter K. Schott.** 2006. "Trade Costs, Firms and Productivity." *Journal of Monetary Economics* 53(5): 917–37.

**Bernard, Andrew B., Stephen J. Redding, and Peter K. Schott.** 2011. "Multi-Product Firms and Trade Liberalization." *Quarterly Journal of Economics* 126(3): 1271–1318.

**Bloom, Nicholas, Mirko Draca, and John Van Reenen.** 2011. "Trade Induced Technical Change? The Impact of Chinese Imports on Innovation, IT and Productivity." NBER Working Paper 16717.

**Brander, James A.** 1991. "Election Polls, Free Trade, and the Stock Market: Evidence from the 1988 Canadian General Election." *Canadian Journal of Economics* 24(4): 827–43.

**Broda, Christian, and David E. Weinstein.** 2006. "Globalization and the Gains from Variety." *Quarterly Journal of Economics* 121(2): 541–85.

**Brülhart, Marius.** 2009. "An Account of Global Intra-industry Trade, 1962–2006." *World Economy* 32(3): 401–459.

**Bustos, Paula.** 2011a. "Trade Liberalization, Exports, and Technology Upgrading: Evidence on the Impact of MERCOSUR on Argentinian Firms." *American Economic Review* 101(1): 304–40.

**Bustos, Paula.** 2011b. "The Impact of Trade Liberalization on Skill Upgrading: Evidence from Argentina." http://www.crei.cat/people/bustos/Trade_Skill_PaulaBustos.pdf.

**De Loecker, Jan.** 2007. "Do Exports Generate

Higher Productivity? Evidence from Slovenia." *Journal of International Economics* 73(1): 69–98.

**De Loecker, Jan.** 2011. "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity." *Econometrica* 79(5): 1407–51.

**Eberhardt, Markus, Christian Helmers, and Zhihong Yu.** 2011. "Is the Dragon Learning to Fly? An Analysis of the Chinese Patent Explosion." CSAE Working Paper WPS/2011-15.

**Eckel, Carsten, and J. Peter Neary.** 2010. "Multi-Product Firms and Flexible Manufacturing in the Global Economy." *Review of Economic Studies* 77(1): 188–217.

**Ethier, Wilfred.** 1982. "National and International Returns to Scale in the Modern Theory of International Trade." *American Economic Review* 72(3): 389–405.

**Feenstra, Robert C.** 2010. *Product Variety and the Gains from International Trade.* Cambridge MA: MIT Press.

**Goldberg, Pinelopi Koujianou, Amit Kumar Khandelwal, Nina Pavcnik, and Petia Topalova.** 2010. "Imported Intermediate Inputs and Domestic Product Growth: Evidence from India." *Quarterly Journal of Economics* 125(4): 1727 –67.

**Greenaway, David, and Richard Kneller.** 2007. "Firm Heterogeneity, Exporting and Foreign Direct Investment." *Economic Journal* 117(517): 134–61.

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

**Halpern, László, Miklós Koren, and Adam Szeidl.** 2005. "Imports and Productivity." CEPR Discussion Paper 5139.

**Harris, Richard.** 1984. "Applied General Equilibrium Analysis of Small Open Economies with Scale Economies and Imperfect Competition." *American Economic Review* 74(5): 1016–32.

**Harrison, Ann E.** 1994. "Productivity, Imperfect Competition and Trade Reform: Theory and Evidence." *Journal of International Economics* 36(1–2): 53–73.

**Harrison, Glenn W., Thomas F. Rutherford, and Ian Wooton.** 1989. "The Economic Impact of the European Community." *American Economic Review* 79(2): 288–94.

**Helpman, Elhanan.** 2004. *The Mystery of Economic Growth.* Cambridge, MA: Belknap Press of Harvard University Press.

**Helpman, Elhanan.** 2011. *Understanding Global Trade.* Cambridge, MA: Harvard University Press.

**Helpman, Elhanan, and Paul R. Krugman.** 1985. *Market Structure and Foreign Trade.* MIT Press.

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

**Kasahara, Hiroyuki, and Beverly Lapham.** 2012. "Productivity and the Decision to Import and Export: Theory and Evidence." February 13. http://faculty.arts.ubc.ca/hkasahara/working papers/productivity.pdf.

**Kasahara, Hiroyuki, and Joel Rodrigue.** 2008. "Does the Use of Imported Intermediates Increase Productivity? Plant-level Evidence." *Journal of Development Economics* 87(1): 106–118.

**Krugman, Paul R.** 1979. "Increasing Returns, Monopolistic Competition, and International Trade." *Journal of International Economics* 9(4): 469–79.

**Krugman, Paul R.** 1980. "Scale Economies, Product Differentiation, and the Pattern of Trade." *American Economic Review* 70(5): 950–59.

**Levinsohn, James.** 1993. "Testing the Imports-as-Market-Discipline Hypothesis." *Journal of International Economics* 35(1–2): 1–22.

**Lileeva, Alla.** 2008. "Trade Evidence and Productivity Dynamics: Evidence from Canada." *Canadian Journal of Economics* 41(2): 360–90.

**Lileeva, Alla, and Daniel Trefler.** 2010. "Improved Access to Foreign Markets Raises Plant-level Productivity . . . For Some Plants." *Quarterly Journal of Economics* 125(3): 1051–99.

**Liu, Runjuan, and Daniel Trefler.** 2011. "A Sorted Tale of Globalization: White Collar Jobs and the Rise of Service Offshoring." NBER Working Paper 17559.

**López, Ricardo A.** 2005. "Trade and Growth: Reconciling the Macroeconomic and Microeconomic Evidence." *Journal of Economic Surveys* 19(4): 623–48.

**Mayer, Thierry, Marc J. Melitz, and Gianmarco I. P. Ottaviano.** 2011. "Market Size, Competition, and the Product Mix of Exporters." NBER Working Paper 16959.

**Mayer, Thierry, and Gianmarco I. P. Ottaviano.** 2008. "The Happy Few: The Internationalisation of European Firms: New Facts Based on Firm-Level Evidence." *Intereconomics: Review of European Economic Policy* 43(3): 135–48.

**Melitz, Marc J.** 2003. "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71(6): 1695–1725.

**Melitz, Marc J., and Gianmarco I. P. Ottaviano.** 2008. "Market Size, Trade, and Productivity." *Review of Economic Studies* 75(1): 295–316.

**OECD.** 2002. *OECD Economic Outlook.* June 2002, no. 71.

**Pavcnik, Nina.** 2002. "Trade Liberalization, Exit, and Productivity Improvements: Evidence from Chilean Plants." *Review of Economic Studies* 69(1): 245–76.

**Schmookler, Jacob.** 1954. "The Level of Inventive Activity." *Review of Economics and Statistics* 36(2): 183–190.

**Shen, Leilei.** 2011. "Product Restructuring, Exports, Investment, and Growth Dynamics." September 24. http://individual.utoronto.ca/lshen/dynamic.pdf.

**Thompson, Aileen J.** 1993. "The Anticipated Sectoral Adjustment to the Canada–United States Free Trade Agreement: An Event Study Analysis." *Canadian Journal of Economics* 26(2): 253–71.

**Topalova, Petia, and Amit Khandelwal.** 2011. "Trade Liberalization and Firm Productivity: The Case of India." *Review of Economics and Statistics* 93(3): 995–1009.

**Tybout, James R., and M. Daniel Westbrook.** 1995. "Trade Liberalization and the Dimensions of Efficiency Change in Mexican Manufacturing Industries." *Journal of International Economics* 39(1–2): 53–78.

**Trefler, Daniel.** 2004. "The Long and Short of the Canada–U.S. Free Trade Agreement." *American Economic Review* 94(4): 870–95.

**Van Biesebroeck, Johannes.** 2005. "Exporting Raises Productivity in Sub-Saharan African Manufacturing Firms." *Journal of International Economics* 67(2): 373–91.

**Verhoogen, Eric.** 2008. "Trade, Quality Upgrading and Wage Inequality in the Mexican Manufacturing Sector." *Quarterly Journal of Economics* 123(2): 489–530.

**Wagner, Joachim.** 2007. "Exports and Productivity: A Survey of the Evidence from Firm-level Data." *World Economy* 30(1): 60–82.

Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 119–140

# Globalization and U.S. Wages: Modifying Classic Theory to Explain Recent Facts

Jonathan Haskel, Robert Z. Lawrence, Edward E. Leamer, and Matthew J. Slaughter

The last 10 to 15 years have seen dramatic changes in globalization and technology juxtaposed with dramatic changes in U.S. earnings: Have the former been driving the latter? Back in the mid-1990s, most research found that the effect of trade on U.S. wages was relatively minor; for example, this was the tenor of a three-paper symposium on "Income Inequality and Trade" in the Summer 1995 issue of this journal. But since the mid-1990s, globalization has taken some new forms: China, for example, was barely on the global economic map in 1995. The patterns of inequality and wages have taken new forms, too. Thus, earlier conclusions are being reconsidered. For example, Nobel laureate Paul Krugman had found in his earlier research on trade and wages, like most others, that the impact of trade on U.S. earnings had been small. But in Krugman (2008), he now conjectures that past might not be prologue: "It's no longer safe to assert that trade's impact on the income distribution in wealthy countries is fairly minor. There's a good case that it is big, and getting bigger."

This paper seeks to review how globalization might explain the recent trends in real and relative wages in the United States. We begin with an overview of what is

■ *Jonathan Haskel is Professor of Economics, Imperial College Business School, Imperial College London, England. Robert Z. Lawrence is Albert L. Williams Professor of International Trade and Investment, Kennedy School of Government, Harvard University, Cambridge, Massachusetts. Edward E. Leamer is Chauncey J. Medberry Chair in Management, Professor of Economics and Professor of Statistics, University of California at Los Angeles, Los Angeles, California. Matthew J. Slaughter is Signal Companies' Professor of Management, Tuck School of Business, Dartmouth College, Hanover, New Hampshire. Their e-mail addresses are ⟨j.haskel@imperial.ac.uk⟩, ⟨robert_lawrence@Harvard.Edu⟩, ⟨edward.leamer@anderson.ucla.edu⟩, and ⟨matthew.j.slaughter@dartmouth.edu⟩.*

http://dx.doi.org/10.1257/jep.26.2.119. doi=10.1257/jep.26.2.119

new during the last 10–15 years in globalization, productivity, and patterns of U.S. earnings. To preview our results, we then work through four main findings: First, there is only mixed evidence that trade in goods, intermediates, and services has been raising inequality between more- and less-skilled workers. Second, it is more possible, although far from proven, that globalization has been boosting the real and relative earnings of superstars. The usual trade-in-goods mechanisms probably have not done this. But other globalization channels—such as the combination of greater tradability of services and larger market sizes abroad—may be playing an important role. Third, seeing this possible role requires expanding standard Heckscher–Ohlin trade models, partly by adding insights of more recent research with heterogeneous firms and workers. Finally, our expanded trade framework offers new insights on the sobering fact of pervasive real-income declines for the large majority of Americans in the past decade. We believe that the connections between globalization, technology, and wages have become much more important during the last 10–15 years.

## New Patterns in Globalization and Wages

The forces of economic globalization have been building since soon after the end of World War II. But the context and patterns of globalization and U.S. wages have evolved in important ways since the mid-1990s. We begin by reviewing what we see as the main changes.

### Five Changes Affecting Globalization and Technology

First, political barriers to trade have been declining. At the multilateral level, the Uruguay Round, in many ways the most comprehensive trade agreement ever, was implemented largely in the decade after its 1994 closing. At the national level, a number of far-reaching unilateral, bilateral, and regional liberalizations have been implemented since the mid-1990s as well, including the North American Free Trade Agreement and China's accession to the World Trade Organization in December 2001. At the industry level, the Information Technology Agreement was signed in 1996, whereby 70 countries representing about 97 percent of world trade in information technology products agreed to eliminate duties on certain information technology products.

Second, natural barriers to trade are declining, especially as a result of the information technology revolution surrounding the Internet. Since Netscape's initial public offering in August 1995, connectivity and communication facilitated by information technology and the Internet have driven marginal transmission costs of voice and data to near zero. This change has reduced the costs of trading goods, and for international trade and investment in services, vastly expanded the scope of what services are tradable.

Third, the U.S. economy has seen a dramatic acceleration in aggregate labor productivity growth since the mid-1990s. The U.S. Bureau of Labor Statistics reports that nonfarm business sector output per hour growth accelerated from 1.4 percent per year over 1973–1995 to 2.5 percent per year over 1996–2009 (Bureau of Labor Statistics data series #PRS85006092, as reported on 9/1/11 at ⟨http://www.bls.gov⟩). A large literature has analyzed this faster U.S. productivity growth and has found a central role for the production and use of information technology hardware (for example, Jorgenson, Ho, and Stiroh in the Winter 2008 issue of this journal, and the references therein)—which, remember, is the one industry in the past generation that implemented a global free trade agreement.

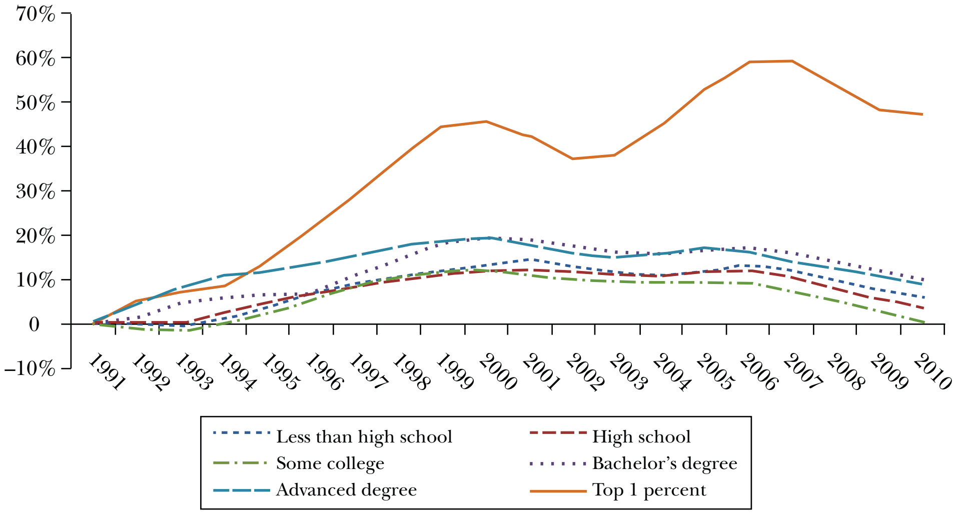
Fourth, GDP growth has accelerated worldwide since the mid-1990s—in particular, in middle- and low-income countries such as Brazil, Russia, India, and China. From 1990 through 2008, annual growth in U.S. gross domestic product averaged 2.7 percent—in contrast to 1990–2008 annual averages of 3.4 percent for the overall world, 4.6 percent for emerging and developing countries as a whole, 6.3 percent in India, and a remarkable 9.9 percent in China (calculated from International Monetary Fund 2008, tables A1–A4.)

Finally, these first four factors have helped to propel a surge in flows of international trade and investment, both worldwide and into and out of the United States. Much of this surge has come from middle- and low-income countries. By 2005, U.S. imports from non-oil developing countries surpassed the value of imports from industrial countries. In addition, U.S. prices of manufactured imports from developing countries declined dramatically. Here again, China stands out: its share of global exports rose from only about 3 percent in 2001 to about 11 percent today, such that it is now the world's largest exporting country. This surge in trade has involved intermediates as well as final products, and services as well as goods (for example, Feenstra 1998; Blinder 2006; Jensen 2011). For the U.S. economy, this surge in trade was far larger for imports than for exports, with resulting historic multilateral trade deficits for the United States peaking at over 5.3 percent of GDP in 2006.

### Three Changes in the Patterns of U.S. Wages

In general, U.S. wages have moved in an upward trajectory over time, with a pattern of rising inequality since the 1970s. However, these patterns have taken on different shapes in the last 15 years or so (for example, Autor 2010a, b; Goldin and Katz 2008; Piketty and Saez 2006; Saez 2012). The key patterns are visible in Figure 1, showing patterns of earnings from 1991 to 2010 for five education groups and also for the top 1 percent of U.S. earners. Of course, these patterns are affected by cyclical factors, with 2000 near the top of a strong business cycle and 2010 one year from the bottom of a severe downturn, but the patterns are nonetheless revealing (and are qualitatively the same if the data end in 2007 before the financial crisis). Figure 1 shows cumulative percentage changes relative to 1991 in mean real (that is, adjusted for price inflation) money earnings for working adults (aged 25 and above) by educational cohort in terms of the highest level of education attained, which is an easily available (if basic) measure of worker skills. The figure also shows cumulative percentage changes relative to 1991 in mean real income (excluding

*Figure 1*
**Changes in U.S. Real Income, Working Adults, by Education and for Top 1 Percent**



*Sources:* The nominal wage data in Figure 1 come from the U.S. Bureau of the Census, Table P-18—Educational Attainment, People 25 Years Old and Over by Mean Income and Sex, 1991 to 2010. Nominal wages are converted into 2010 real wages using the CPI-U-RS index of consumer prices from the United States Bureau of Labor Statistics. The real-income data for the top 1 percent of tax filers comes from Saez (2012, supplemental table A4), where each year's nominal income is deflated using the same price index.
*Notes:* Figure 1 shows cumulative percentage changes relative to 1991 in mean real (that is, adjusted for price inflation) money earnings for working adults (aged 25 and above) by educational cohort in terms of the highest level of education attained. All percentage changes along the *y*-axis are actually log changes (which approximate percentage changes), smoothed to three-year moving averages to eliminate occasional annual volatility. There are important measurement differences between these two wage sources—Census Table P-18 and Saez (2012, supplemental table A4). One is the units of observation: Table P-18 measures income for individual workers; Saez measures income for tax units, which can contain more than one worker because they can consist of, for example, an individual, a head of household with children dependents, or a couple with children dependents. That said, the basic income patterns in Figure 1 are robust to measurement issues.

capital gains) for the top 1 percent of all tax units filing returns to the U.S. Internal Revenue Service.

Figure 1 contains three key messages about U.S. earnings. First, in the second half of the 1990s, all groups of workers by education status experienced strong increases in real income; but after about 2000, all these groups of workers experienced *declines* in real income, such that over the full 1991–2010 period, growth in real earnings was very weak. Post-2000, all five educational groups shown suffered falls in average real money incomes, and over the full 20 years, average real income grew less than 10 percent for all five groups. This picture of poor real earnings performance improves only slightly when factoring in changes in the roughly 18 percent of total labor compensation accounted for by nonmonetary benefits (including life insurance, health insurance, stock and stock-option grants). How could falling real incomes for so many American workers coexist with ongoing U.S. GDP and productivity growth during the 2000s? Part of the answer was sharply higher earnings by capital. Corporate profits rose strongly over the 2000s—as they had in the late 1990s, too. As a share of GDP, U.S. corporate profits reached 12.4 percent in 2010—the highest percentage ever recorded in the roughly 60 years the U.S. government has tracked this item. Of course, this high level was in part a result of high unemployment and low labor earnings in the aftermath of the Great Recession.

Second, many of the standard measures of income inequality that focus on the very broad middle of the distribution rose very modestly, if at all, since the mid-1990s. For example, the ratio of the median annual earnings of college graduates to high school graduates stood at 1.69 in 1999 and 1.71 in 2009 (the similar ratio for mean earnings was unchanged at 1.79). The ratio of the earnings of the median worker to the earnings of the worker at the 10th percentile of the overall income distribution actually declined during this time.[1]

Third, the income of the highest-earning workers has risen dramatically both in absolute terms and relative to all others. Average real income of this top 1 percent of IRS tax filers rose from $534,264 in 1991 to a peak of $1,003,791 in 2007 and was still $857,477 in 2010. The share of U.S. income (again, excluding capital gains) accounted for by this top 1 percent rose from just 7.7 percent in 1973 to 13.5 percent in 1995 and 16.5 percent in 2000; this share then rose further, to 18.3 percent in 2007—although it has declined since then in the wealth meltdown of the Great Recession. High-income earners tend to be highly educated, but this linkage is not perfect: for example, Bill Gates is a college dropout. We, like many others, will refer to this small group of highly skilled, highly compensated workers as *superstars* (Rosen 1981).[2]

## Relating Trade, Technology, and Wages

### Old Frameworks Applied to the New Facts

A conceptual framework should fit the circumstances. From the mid-1970s to the mid-1990s, rising levels of U.S. wage inequality took the form of a pervasive, economywide increase in returns to skills that were easily identified by education (for example, college versus high school) or occupation (for example, white- versus blue-collar). It was thus not surprising that much analysis of income inequality used models that assumed two homogenous types of labor: skilled and unskilled. Similarly, many labor economists used a one-product model that focused on technology

[1] The 50–10 statistic comes from Figure 3 of Autor, Katz, and Kearney (2008). The education-based ratio is based on the authors' own calculations using the data in the note to Figure 1.

[2] Existing research documenting and/or examining the causes of rising superstar earnings include Gabaix and Landier (2008), Gordon and Dew-Becker (2007), Kaplan and Rauh (2010), Lemieux (2006), Piketty and Saez (2003, 2006), and Saez (2012).

innovations boosting demand for skilled workers. Many trade economists used a two-product model with differing factor intensities and with perfect labor mobility across the two sectors.

Given the pervasive shifts in relative wages, it was natural to base empirical analysis on the intuition of the classic Stolper–Samuelson theorem that rising U.S. returns to skill were driven by rising prices of skill-intensive products relative to unskilled-intensive products. Some authors looked for Stolper–Samuelson effects with small general equilibrium simulation models (for example, Krugman 1995; Cline 1997), and some examined observed prices directly (see summary in Slaughter 2000). Others looked for the labor supplies embodied in trade flows (for example, Borjas, Freeman, and Katz 1997). Most studies found some link from trade to rising inequality, but with a few exceptions, the magnitude was not large. Cline's (1997) comprehensive survey argued that "a reasonable estimate based on the literature would be that international influences contributed about 20 percent of the rising wage inequality in the 1980s."

During the more recent period of what seems to be accelerating global exposure of the U.S. labor market, one might expect these effects to be even stronger: surely trade must anchor the returns of the homogenous lower-skilled categories of labor? But the uneven performance of skilled workers, with some wage declines and superstar increases, suggest that models based on two types of labor cannot capture what is occurring. In response, labor economists like Autor, Levy, and Murnane (2003) have developed a more sophisticated theory of skill-biased technological change in which computers and other innovations in information technology complement highly skilled nonroutine cognitive tasks, substitute for moderately skilled routine tasks, and have little effect on less-skilled manual tasks. The result is downward pressure on wages and employment opportunities on moderately skilled workers, such that inequality between them and their less-skilled counterparts no longer rises. Autor (2010b) discusses this "polarization" of the U.S. labor market.[3]

Efforts to apply the simple two-factor Stolper–Samuelson framework to recent data have run into various problems. For example, using U.S. factor inputs at the most disaggregated level for which skill measures are available, the factor content of U.S. imports from developing countries is *not* especially intensive in unskilled labor (Edwards and Lawrence 2010). A large share of U.S. manufactured imports from developing countries is in skill-intensive industries such as computers and electronics. Indeed, Mishel, Burnstein, and Shierholz (2009) estimate that the education mix of the net factor content of U.S. trade in recent years is very similar to that of the labor force overall (of course, this might to some extent reflect a measurement problem whereby the imports of Chinese unskilled-intensive hours of iPad assembly are classified by the

[3] Goldin and Katz (2007) and Yellen (2006) suggest that the globalization of production has similar properties in inducing polarization of wage distribution. They argue that suppliers of personal services at the low end escape downward pressures from trade because these services must be provided locally. But those at the top are rewarded by trade, while those in the middle are hurt. A related issue may be that returns to skills may be changing outside of conventional educational measures—for example, the returns to noncognitive skills discussed in Heckman, Stixrud, and Urzua (2006).

U.S.-measured skill-intensive final good). Bivens (2007) analyzed a simulation model that simply *assumes* all developing-country imports are unskilled-intensive and that all goods imported to the U.S. economy are also made domestically—thus clearly leading to an upward-biased estimate of how trade might affect inequality—yet he found that increased U.S. trade with developing countries boosted the U.S. skill premium by only about 2 percent between 1995 and 2006.

Thus, a number of trade-based studies of U.S. wages have, perhaps surprisingly, not found much connection between surging U.S. imports from low-wage countries and recent U.S. wage trends when analyzed by the traditional Stolper–Samuelson trade logic.

**Let's Be More Specific: Newer Trade Frameworks**

The standard Heckscher–Ohlin model with mobile workers between industries implies wages are due to general returns to skill. In settings with different types of firms and workers, international trade can also affect the returns to worker attributes that are more "specific" to the worker-employer match. Research on worker mobility has long found that human capital is partly specific to industries and occupations (for example, Jacobson, LaLonde, and Sullivan 1993; Neal 1995; Kambourov and Manovskii 2009). The field of international trade has seen a surge of research developing and analyzing a richer set of interactions among firms and workers of different types. For example, if workers are at least partly immobile across industries, then freer trade often boosts earnings of workers specific to export industries while lowering earnings of their specific counterparts in import-competing industries (an intuition first developed in classic papers such as Jones 1965; Mussa 1974; and Neary 1978). Alternatively, if autarchic product or labor markets are not perfectly competitive, trade has a pro-competitive effect. For example, unionized workers may be forced to accept lower wages if freer trade makes the demand curves faced by firms more sensitive to price (Rodrik 1997) or affects rents to be shared (Lawrence and Lawrence 1985).

In many new trade models, heterogeneous firms and workers interact in previously unexplored ways. For example, in the Melitz (2003) model (described elsewhere in this symposium) reductions in trade costs boost profitability in the most productive (and thus exporting) firms. This raises profit inequality across firms—but there is no wage inequality in the basic version of the model because workers are assumed to be identical. In other heterogeneous-firm models, wage inequality does arise by assuming some sort of link from profits to wages. Examples here include notions of fairness (Egger and Kreickemeier 2009); rent sharing (Amiti and Davis 2008); and incentives to search for quality workers (Helpman, Istkhoki, and Redding 2010), reduce worker shirking (Davis and Harrigan 2007), or upgrade skills (Verhoogen 2008). Other theories focus on the process by which firms match with heterogeneous workers who span a continuum of skills. Here, opening to trade can alter the process by which workers sort into firms and, through this, impact earnings related to skill. Sometimes, trade can have wage effects that resemble classic Stolper–Samuelson linkages. But wage outcomes in these heterogeneous firms and workers settings can be quite different and can potentially describe recent U.S. wage trends with stagnant earnings for both less- and moderately-skilled workers and rising superstar earnings (Blanchard and Wilmann 2011; Costinot and Vogel 2010; Manasse and Turrini 2001).

### Applying New Frameworks to the Data

Countries have experienced a wide variety of wage changes after trade liberalization, as surveyed by Goldberg and Pavcnik (2007). New theories of trade with heterogeneous firms and workers, by allowing explanations of inequality that reflect more than the returns to broad skill categories, offer some possibility of explaining these patterns. However, testing these new theories against particular episodes—such as the recent U.S. experience—requires different empirical approaches from Stolper–Samuelson analyses. After all, Stolper–Samuelson analyses are based on the general-equilibrium Heckscher–Ohlin trade framework in which there is no relationship between a worker's wages and the trade (or lack thereof) in that worker's industry. In contrast, various specific-factors theories must be tested by linking wages to firm and/or industry characteristics. Indeed, in the Heckscher–Ohlin model, wages by skill are the *same* in all industries, so any observed correlation between wages and industry features merely signals the types of workers employed in an industry with those characteristics: industries with large import volumes likely employ unskilled workers doing mundane tasks.

Recent empirical research surveyed comprehensively by Harrison, McLaren, and McMillan (2010) has examined the effect of trade on wages at the level of firms, occupations, regions, and industries. Some U.S. studies link data on trade and other variables to individual worker data. Ebenstein, Harrison, McMillan, and Phillips (2009) find no effect of import competition at the level of industry wages, but they find that workers displaced from manufacturing earn 3–9 percent less if reemployed in other sectors. Autor, Dorn, and Hanson (2011) find that Chinese manufacturing imports did not reduce wages within manufacturing, but did depress local wages more generally by 2 percent over 17 years. McLaren and Haboyan (2010) reach qualitatively different conclusions about NAFTA's impact, finding no impact on local wages but downward pressure on industry wages. Liu and Trefler (2008) find that outsourcing of traded services has reduced U.S. industry earnings, but that these effects are "tiny;" Liu and Trefler (2011) examine the effect of traded services on U.S. occupational switching.

One way to read these studies is that they examine labor-market adjustments, such as labor-force participation, unemployment, and occupational change, about which the classic Heckscher–Ohlin model is silent. As an empirical issue, these may well be the most important short- to medium-run adjustment margins to globalization, rather than wages. Thus the Heckscher–Ohlin model might be best suited for examining longer-run wage outcomes. Our view is that a more-complete accounting of wage outcomes in the overall U.S. economy needs at least a model that integrates both general and specific returns. In the next section of this paper, we offer such an approach rooted in the classic Heckscher–Ohlin trade model.

## A Heckscher–Ohlin Trade Model with a Richer Wage Structure

### The Basic Heckscher–Ohlin Framework

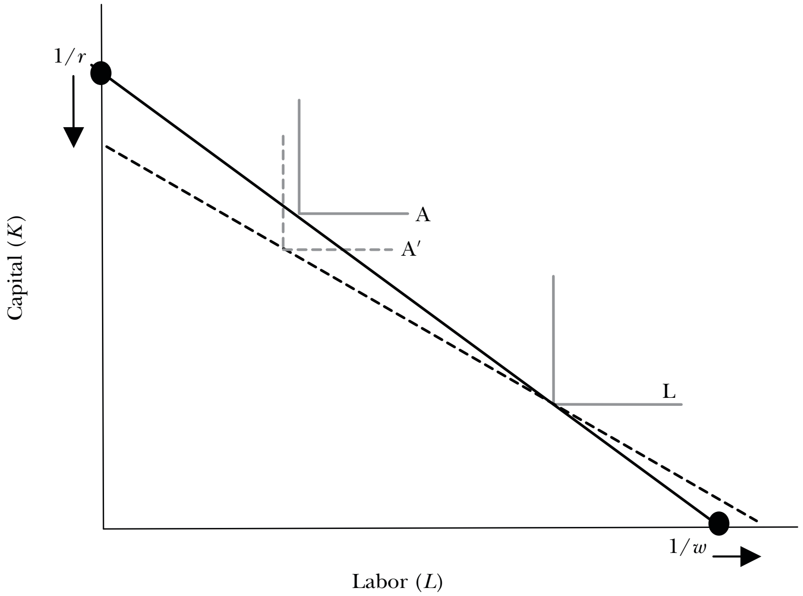
In the standard one-sector model with skilled and unskilled workers, relative demand for labor depends only on relative wages within the sector. There is only one margin of adjustment following a shock to relative labor demand or supply: namely, a shift in relative wages, the size of which depends on the factor elasticity of substitution. However, any university dean knows that wages of finance professors seem to be determined not by conditions inside the education industry, but elsewhere. A key feature of the Heckscher–Ohlin framework is precisely this: the industry is not the market. Multiple products mean more margins of adjustment to industry shocks besides relative wages—for example, relative outputs can change, too. Counter to much of the firm- and individual-level work set out above, wages depend on conditions in the market as a whole, and not just in the particular worker-firm match.

Figure 2 describes a world where the industry is decidedly not the market: instead, the whole economy is the market. The right-angle shapes are unit-value isoquants, showing the quantities of capital $K$ (physical or human) and labor $L$ required to produce 1 unit of value—say, \$1 of value—of the capital-intensive (A) and labor-intensive (L) goods at prevailing exogenous goods prices and technologies. Unit-value isoquants in which inputs are always used in fixed proportions to produce a unit of output and there is zero elasticity of factor substitution (this is called "Leontief" technology) are a simplifying assumption. No key results depend on this assumption (Leamer 1995). The location of each isoquant depends on *both* technology and goods prices. The straight lines are unit-cost lines, showing the costs of $K$ and $L$, $r$ (the capital cost) and $w$ (the wage), such that total costs are \$1: that is, $1 = wL + rK$. (Actually we see the reciprocals of $r$ and of $w$, rather than $r$ and $w$ directly.) $K$ and $L$ are assumed to be mobile across industries, which is why the industry is not the market. Wages are determined such that profits are zero in both industries (or else factors would move between industries); this is indicated by the heavy straight line. The reciprocals of $r$ and $w$ ($1/r$ and $1/w$) are the heavy dot intercepts of this line.

How do capital costs $r$ and wages $w$ change? Consider anything that, at these initial factor prices, makes the capital-intensive industry more able to produce \$1 worth of output using fewer inputs: for example, a rise in its output price or a technological change favoring that industry (that is to say, lowering unit costs in that industry at initial $r$ and $w$). Either of these effects would shift A to A′ towards the origin, since less $K$ and $L$ are now required to make a capital-intensive good of value \$1. As the diagram shows, the only way to restore equilibrium is for the unit-cost line to flatten (the dotted line). Thus, $w$ must fall and $r$ must rise to restore zero profit equilibrium. This result embodies the Stolper–Samuelson intuition: changes in product prices or production technology that raise the profitability of a sector, at initial wages, tend to raise the wages of factors employed intensively in that sector.

As discussed earlier, this elegant Heckscher–Ohlin model does not seem well-suited to explain many of the recent wage developments; for example, the highly

*Figure 2*
**A Basic Heckscher–Ohlin Model**



*Notes:* The axes show quantities of labor and capital. There are two industries (A and L), the upper left using capital intensively relative to the lower right. The right-angle shaped lines are unit-value isoquants, showing combinations of labor and capital needed to produce one dollar's worth of the two goods. The downward-sloping lines are isocost lines showing the combination of capital costs and wages such that there is full employment of factors at given technologies and product market prices. The dotted isocost line corresponds to a rise in the relative price of the capital-intensive goods. The heavy dot intercepts on the *x*- and *y*-axes show the reciprocal of the equilibrium factor prices of *L* and *K*, so a movement on the axis towards the origin is an increase in factor prices.

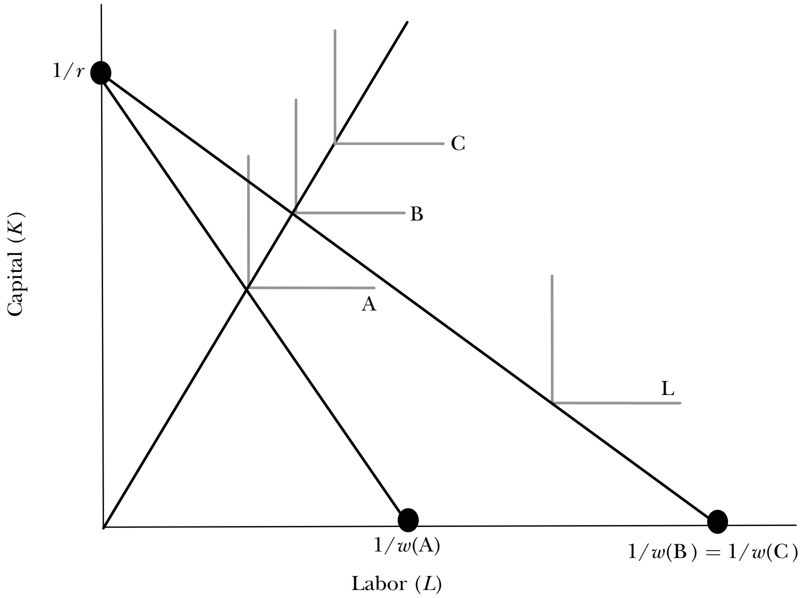
skilled, highly-paid workers whose "superstar" earnings have risen so dramatically. Also, its assumption of homogeneous firms does not allow consideration of worker returns specific to particular firms or industries, or specific to noncognitive or nonroutine skills.

### A Richer Heckscher–Ohlin Framework

Following Leamer (1995, 2012), consider extending the basic Heckscher–Ohlin model to allow capital and heterogeneous labor with varying amounts of "talent." This approach also allows capital–talent complementarity: that is, talented workers are more productive when working with capital, whereas they are no more productive in unskilled tasks.

Figure 3 presents this richer model. It shows four unit-value isoquants: three for workers with respective talents A, B, and C in the capital-intensive sector and another in a labor-intensive sector, L, where talent is assumed not to affect productivity. The diagram also shows the single *r* (common since capital is assumed to be

*Figure 3*
**A Richer Heckscher–Ohlin Model**

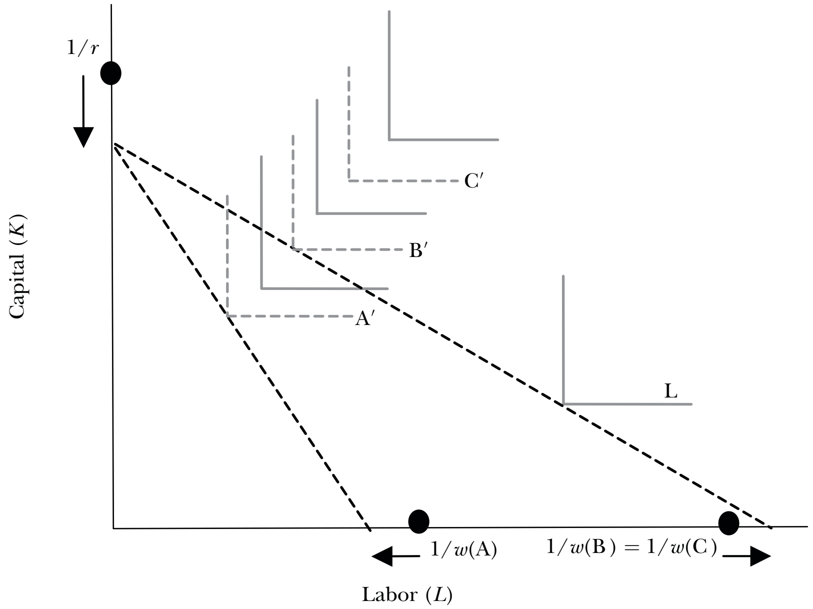


*Notes:* In Figure 2, there is only one worker type. Here, there are three workers types: highly talented to less talented (A to C). The highly talented are assumed more productive in the capital-intensive industry. All talent types are assumed equally productive in the labor-intensive industry. The downward-sloping isocost line show the combination of capital costs and wages so there is full employment of all talent types and types B and C are indifferent between the industries they choose. See also the notes to Figure 2.

mobile across sectors) and a set of wages that maintains full employment. The most talented type-A workers are more productive, by assumption, when renting capital. Thus they can command, at the given *r*, a high wage consistent with the position of the dot on the labor axis, 1/*w*(A) (determined in turn by the tangency with the unit-value isoquant A). Now consider type-B workers. As drawn, with wages *w*(B), they are the marginal talent level: just indifferent between working in the *K*-intensive industry—where their talents at *w*(B) make them sufficiently productive to be profitably employed there but insufficiently productive to profitably command *w*(A)—or working in the *L*-intensive industry. Finally, the type C's earn *w*(C) = *w*(B), the same as type Bs, but only if they work in the *L*-intensive sector (L).

What does inequality look like in this economy? First, wage inequality stems from talent–capital complementarity: type-A workers pay the same *r* as others but are more productive with capital and so can command a higher wage consistent with zero profits (shown by the intercept on the *L*-axis lying closer to the origin). Untalented type-C workers, insufficiently complementary, optimally work in the *L*-intensive sector. In sum, the market sorts heterogeneous worker types and determines a competitive talent premium where no rents are shared.

*Figure 4*
**Changes in the Richer Heckscher–Ohlin Model**



*Notes:* The dotted "right-angle" shapes are unit-value isoquants following a rise in relative prices or technical change in the capital-intensive industry. The dotted isocost lines show the combination of capital costs and wages so there is full employment of all talent types and types B and C are indifferent between the industries they choose after the increase in output prices in the capital-intensive industry. The arrows on the axes originating from the heavy dots show the rise in capital prices and wages of type A's and the fall in wages of type B's and type C's. See also the notes to Figure 3.

Second, if talent is unobservable to the econometrician, then inequality will have a "within-group" feature: $w$(A) versus $w$(B) for observably identical workers in the same industry.

Third, and related, the model can explain the "fractal" nature of inequality: that is, within-group wage inequality in successively narrower and narrower defined groups. There is inequality between all labor in the economy. There is also inequality between all MBA graduates, some of whom are unobservably talented and work with a lot of capital and command high wages while some are unobservably not so talented and are paid the same as other MBAs working with less capital.

How is inequality affected by a price rise or technical advance favoring the $K$-intensive sector? One might suppose a straightforward answer, namely a rise in wages of all talented workers. Not so.

The reason is set out in Figure 4, where the talent-specific, $K$–intensive unit-value isoquants have shifted and are now the dotted A′, B′, and C′ following a rise in that sector's price (or its improved technical opportunity)—for simplicity holding all other product prices fixed.

As in the traditional case, *r* rises.[4] And as in the traditional case, with *r* rising, wages in the *L*-intensive sector must fall to restore profitability. That means *w*(C) falls. But as well as lowering wages in the L sector, it also reduces wages for the type-B workers, even though they are in the *K*-sector and even though their market demand has risen. Type-B workers are insufficiently talented to command higher wages in the face of the increased price of capital with which they have to work. Type-C workers lose along with the type-B because their wages match those of the type-B workers. The gainers are the type-A workers. They experience the negative effect of higher capital rental charges but this is completely offset by the favorable productivity effect: in the diagram, their intercept point on the *L*-axis moves to the left. (Formally, what drives the rise in wages of the A-workers is the Jones (1965) "amplification," the fact that the percentage increase in capital costs is less than the percentage increase in the price of the capital-intensive good, which means that even after paying higher capital costs, there is more left over to pay workers.)

Thus, it is *not* the case that the wages of *all* workers fall in response to a relative increase in the price of the capital-intensive good, as in Figure 2. There are winners and losers. The winners are the most talented workers matched with, or sorted into, the industry where their talent matters most: where they are most effective in operating the expensive capital. The losers are workers with less talent, even if they are working in the capital-intensive sector or in the labor-intensive sector where their talent does not help them.

If one regards the type-A workers as the most skilled, like those with advanced degrees and/or special skills, type-B workers represent the moderately skilled, perhaps ranging from those with nonprofessional degrees to those with some college, and type-C workers represent the less-skilled, like those with only a high school degree or less education, then our model can explain Figure 1: rising wages for those few at the very top and falling relative and stagnating real wages for all others. To flesh this out more we need to be more specific on what we mean by capital and talent.

### Applying This Richer Heckscher-Ohlin Framework to the Recent U.S. Experience: Capital and Talent in the Modern Global Economy

Does the model help us understand recent U.S. wages? Since the key is the talent–capital complementarity, along with price and/or technology changes in particular sectors, we focus on the possible different interpretations of "capital" and "talent" in the U.S. economy.

From this perspective, one initial puzzle might be how this enriched Heckscher–Ohlin model applies to talented workers at Goldman Sachs, Microsoft, Disney, Facebook, and Google, who don't seem to work with much physical capital. But suppose all labor in Figures 3 and 4 are actors, and the capital with which

[4] To see this graphically observe that one cannot draw a new isocost line with its origin at the previous value of $1/r$ (the large dot on the vertical) that is tangent to both the new unit value isoquant B′ *and also* L. The isocost line must flatten, which implies that *r* must rise and *w*(B) fall.

they work is *intellectual or intangible* capital—movie scripts, special effects, software, scenery, and directorial and editorial talent.[5] Talented actors with good scripts are potentially very profitable and thus have unit-value isoquants A. Less-talented actors are not quite talented enough to command $w$(A) but can earn a lower wage $w$(B) working with the same ratio of intangible capital, or they can work in an intangible-capital–extensive sector along with the untalented (that industry might be movies with poor scripts/scenery/special effects, or nonmovie industries that also use intangible capital but not as intensively, like the cinemas that rent the movies, or industries also using software but not as intensively).

Now globalize the movie industry, such that previously domestic-only actors can now potentially command global audiences. This globalization could arise from a number of forces discussed earlier: foreign GDP growth that stimulates demand for entertainment in newly emerging middle classes, or governments removing restrictions on imports of U.S. movies, or the information technology revolution reducing the costs of cross-border digital distribution of films. Whatever the causes, the result is a shift in A and B to A′ and B′, leaving L the same. In turn, it is *not* the case that *all* actors earn higher real incomes. The most talented actors become superstars, now earning stratospheric wages. The less-talented actors earn less, even if they remain in the movie industry. They earn less because that is the only way that they can now be profitably employed in the movie industry at their talent levels. The other parties who earn more are, of course, the owners of "capital."

Thus, this model seems to have a number of attractively accurate predictions. The stars in the Harry Potter films earn more. The owner of the Harry Potter "capital," author J. K. Rowling who owns the script copyrights, earns more. The movie industry expands. But actors not in *Harry Potter* earn less, because at their talent levels they have to take lower wages to accommodate the increased costs of paying copyright-holders. And less-talented actors also face lower wages if they don't work in movies. Likewise, star computer programmers earn more in the expanded software industry; if they also own the intellectual and reputational capital, like Mark Zuckerberg, Sergei Brin, and Larry Page, they earn the capital rents as well.

A similar logic can apply to highly educated occupations, such as bankers and lawyers, if the talented ones among them are complementary with capital in financial and legal services. What is the capital here? One interpretation is that $L$ is hours and $K$ is human capital, so that a high $K/L$ production technology is one with intensive use of banking or legal human capital per hour. This might reflect complexity of the task at hand, and so a rise in the output price of the industry is an increase in the price of complex financial and legal services. Thus, assume that

[5] Much of this capital comes close to sounding like other workers. What matters formally is that the ownership of the capital be distinct from individuals, so that one can distinguish between the earnings of capital and the earnings of workers. For example, the copyright to the script is owned by the movie studio, as opposed to the actor, so that $r$ is the rental price paid by the studio of scripts and $w$ is actor wages. Of course, if the script is owned by the actor as well, then all returns accrue to the actor. The increasing importance of intangible capital in the United States is documented by Corrado, Hulten, and Sichel (2005, 2009).

talented lawyers working on complicated cases are more productive than untalented lawyers on similar cases and also than if they were processing routine legal administration. Then the model predicts a rise in wages for the most talented and a rise in the return to human capital when legal services globalize. Here, human capital might be embodied in workers or it might be a law firm's contacts and know-how among its partners. But such an increase in these capital costs lowers wages for the less talented (like paralegals) and hence within-industry inequality rises. A further possibility is that the productivity of top lawyers is further enhanced by working with other top lawyers, in which case the unit-value isoquant for A types shifts in again, raising $w$(A) yet more. Of course, to understand fractal inequality, one would have to assume further that even talented lawyers or bankers are imperfect substitutes for each other; that is to say, criminal lawyers are more productive when matched to criminal cases than property lawyers.

What of chief executive officers and top managers, such as Jack Welch or Steve Jobs? Suppose now that capital is reputational or organizational capital at the firm. Thus, a high capital/labor ratio firm has a high reputation, very efficient supply chains, or well-structured hierarchies. Think of Apple, for example, which has relatively little physical capital but very high design and reputational capital as well as organizational capital in managing multicountry production. Consider now a rise in relative prices for such high-capital firms—one perhaps triggered by global trade liberalization, or fast growth in emerging markets, boosting demand for their services. Indeed, a rise in demand for reputation or organizational capital may well be a consequence of globalization, where these intangible assets (like supply-chain management) can be spread across borders via multinational firms (Markusen 2002; Spence 2011). The Internet revolution, bringing a rise in anonymous remote transactions, might increase the return to reputation. The model predicts a rise both in highly talented wages and a fall in lesser-talented wages, even if working for the same firms, and a rise in the returns to good capital: in this case, company reputation or organizational capability. This interaction gives an extra dimension to the outsourcing literature, in which workers suffer because unskilled tasks can be outsourced. Here that finding remains true, but in addition, able managers gain from outsourcing because they can apply their scarce talents to managing the outsourcing process.

Note from our above examples that "globalization" should be conceived of quite broadly. The traditional trade mechanisms revolve around changes in the prices of tradable products—often in response to changes in trade policy. Here, globalization means something broader: any change that raises profits in the capital-intensive sector at current product prices, factor prices, and technology. In this broader meaning of the term, a rise in globalization still could be triggered by traditional mechanisms such as trade policy. But it could be triggered by many other mechanisms as well—especially for the widening span of tradable services (as described in Jensen 2011) that are often fostered by information technology innovations (as documented by Fort 2012): for example, rising global demand for American banking services, consulting services, movies, and sports triggered by rising global GDP.

### What about Technological Change in This Richer Heckscher–Ohlin Model?

Cheaper communications—and what we have in mind here is the Internet—increase the scope for outsourcing and/or increase market size. More efficient semiconductors make computer capital more efficient. Both changes potentially shift unit-value isoquants in Figure 4. But the message of this model is that inequality might rise or fall, depending on the interactions between the technical change and the worker types. To illustrate, suppose the *K*-intensive industry is music, and talents refer to musicians. To modify Leamer (1995) slightly, the question is whether the technical change improves microphones or mixing desks. A better microphone improves the relative productivity of the most musical and thus shifts in type A's unit-value isoquant and raises the talent premium as above. A better mixing desk renders production of studio-quality music within the reach of even the most talentless. This might squeeze the gap between A and B; then wages of type A's fall, and wage inequality might fall.

A recent current of literature suggests that computers might not just affect the productivity of the skilled versus the unskilled, but also the productivity of those performing nonroutine activities (for example, Goos and Manning 2007; Autor 2010b). These papers mostly look at the consequent effects on employment. In general, if we relabel the industries in Figure 3 and 4 as nonroutine and routine, and assume that computers make it relatively cheaper to perform routine tasks such that the price of the routine industry falls, then employment in routine tasks falls and the effects on wages depend crucially on the effect on talent and the industry concerned.

All this illustrates a general point: namely, that inequality in information-rich societies looks totally different from that in the past (Leamer 1995). To illustrate this, consider physically strong workers, valuable workers in their day in rural societies with no machinery to perform heavy tasks. With the advent of manufacturing and more recently computers, cheaper capital removes most of their comparative advantage; these "talented" workers work alongside other worker types, and wages are equal.[6]

But now computers don't just do routine tasks, but carry information all over the world. This now raises the potential return to physically strong workers in entertainment services such as NFL football. Such workers leave manufacturing to play NFL football and endorse consumer brands, where their talent is complementary with the global market and reputation capital in the entertainment industry. Workers segregate and inequality rises, but not necessarily along educational lines; in this example, football players need not be the most educated.

[6] In terms of the diagram, an economy with little intangible capital in it (it might be a closed economy, or a developing economy, or an economy with mostly routine production technologies such as 60 years ago) has a high price of intangible capital and thus a relatively "flat" unit cost line. Thus there are no or very few A-type workers with enough talent to work in the capital-intensive industry. All workers of all talent types work together in the labor-intensive industry and are paid the same; we have a very low-inequality country but with talented workers matched into basic industries. An alternative view is that such intangible capital was mostly unimportant in earlier days of industrialization when tangible capital was most important such that there was not a separate industry intensive in the use of intangible capital—and thus there were not separable industries like A, B, and C. Beaudry and Green (2003) present a model where modern economies use capital and skilled labor intensively, a combination that tends to lower unskilled wages.

### Relating Heterogeneity in This Richer Heckscher–Ohlin Model to Heterogeneity Elsewhere

Suppose now that heterogeneity is not in workers, but rather in firms. This kind of model is considered by Melitz (2003), Helpman, Istkhoki, and Redding (2010), and others. Suppose the *K*-intensive industry has the opportunity to export and that some firms are more "talented," which in this context means more productive, in that industry. Such firms export and their talent, perhaps being superior owner-managers, earns higher returns than that of owner-managers of lesser firms. Thus, within the *K*-intensive sector there emerges profit inequality, in the sense of inequality of returns to the scarce factor that is correlated with observables such as exporting status.

The original Melitz (2003) model had no wage inequality, because workers were the same ability, but it did have profit inequality. Indeed, the aim of that model was to explain how trade led to productivity growth via the sorting of firms, as Melitz and Trefler explain in their paper in this volume.

Researchers have sought to add labor-market imperfections to this basic model. In Helpman, Istkhoki, and Redding (2010), workers differ by ability, and firms screen and bargain over quasi-rents with them. Larger firms screen workers more intensively and so employ higher ability mixes, workers who are able to bargain higher wages. As a result, within-group wage inequality emerges in a single sector using a particular *K*-intensity. Falling trade barriers create larger exporting firms, and so within-sector wage inequality rises.[7] This result essentially follows from matching followed by rent sharing. As they remark, general equilibrium effects from other sectors can arise as standard Stolper–Samuelson effects.[8]

In the model we have presented, there is no rent sharing (for simplicity) but there is sorting of workers among sectors, similar to Costinot and Vogel (2010). In that model, there is no capital, but rather there are many industries each employing one worker with a certain skill who performs a certain task. The key talent/complementarity (assignment) relationship is between tasks and production: high skilled are substitutable for low skilled, but high- and low-skilled workers in a skill-intensive industry produce more than such workers in a low-skill-intensive industry.

[7] Wage inequality in the Helpman, Istkhoki, and Redding (2010) model is driven by a number of factors. First, due to assumed matching problems, there are unemployed and employed and hence inequality for that reason. Second, within the employed, some are in exporting firms and some domestic, the former of which screen out low-ability types. So inequality results for that reason as well. Third, each firm is assumed to bargain a single wage for all its workers depending on the average expected ability level, so inequality is affected by the size of each firm. Furthermore, inequality is the same if all firms are purely domestic or all purely exporting, so the relation between inequality and exporting volume is hump-shaped such that inequality initially rises and then falls.

[8] However, it is worth noting that in a supplement to the paper (Helpman, Istkhoki, and Redding 2010, technical appendix, section 5.2) they extend their model to include a second sector, the nearest parallel to our model set out above. Like Davidson, Martin, and Matusz (1988), they consider a second sector that employs workers without search frictions. As they comment, changes in relative wages "will be determined by Heckscher–Ohlin forces, which directly affect between-group wage inequality, but have no effect on within-group inequality in the differentiated sector" (p. 20, section S5.2).

Opening to trade for a skill-abundant country allows it to specialize in skill-intensive tasks, which in turn raises skilled wages—similar to the analysis we have presented.

## Conclusions and Future Research

We hope that readers will take from our paper three main conclusions about the recent trends in U.S. real and relative incomes. First, to date there is little evidence that globalization through the classic channel of international trade in goods, intermediates, and services has been raising inequality between more-skilled and less-skilled workers. Second, there is at least suggestive evidence that globalization has been boosting the real and relative earnings of superstars. The usual trade mechanisms probably have not done this, but other globalization channels—in particular, the combination of greater tradability of services and larger market sizes abroad—may be playing an important role. Third, our analysis sheds new light on the sobering fact of pervasive real-income declines for the large majority of Americans in the past decade. These real-income declines may be part of the same globalization and innovation forces shaping returns to superstars and to capital.

These conclusions must be placed in the proper context, which is "there is so much more we need to know from future research." A good deal of recent empirical work investigates the effects of trade on the adjustment process of particular workers, occupations, and industries (which simple models ignore), and documents (the sometimes long-lasting) adverse effects. Our goal here, however, has been to advance some basic models describing the economywide evolution of, for example, widespread real-wage declines but rising earnings of superstars. Of course, future research will hopefully explore not only the experience of the United States but that of many other countries as well—both developed and developing.

For superstars, we do not yet fully understand product prices in sectors that employ superstars relatively intensively. This is both because existing industry data do not distinguish highly talented individuals well (if at all), and because many of the sectors in which we presume superstars are concentrated like finance, law, consulting, athletics, and entertainment do not have reliable data on product prices (or much else). Nor do we have good data on personal attributes that make individuals potential superstars. We suspect that for at least some of these superstar-intensive industries, globalization has played an important role in boosting demand for their services—both via the information technology revolution reducing their natural trade costs and thus boosting their tradability, and via fast economic growth around the world boosting demand for their services. But these conjectures await additional analysis.

With regard to the sobering falls in real income for the large majority of Americans, our framework does add some new insights. We agree with Autor (2010a) that explaining falling real income for so many American workers remains a daunting empirical challenge. Much research to date has focused on income inequality, not income levels. We argue that this focus should change, because the post-2000 real-income declines are pervasive, new, and troubling. Our enriched trade framework offers some possible explanations for how globalization and/or innovation can boost superstar real earnings yet reduce real earnings of so many others.

■ *For extremely helpful editorial suggestions and guidance we particularly thank JEP Editors David Autor and Timothy Taylor. In addition, for comments and discussions when writing this paper we also thank Gordon Hanson, Chad Jones, Donald Kohn, Frank Levy, John List, Steven Redding, and David Wessel.*

## References

**Amiti, Mary, and Donald R. Davis.** 2008. "Trade, Firms, and Wages: Theory and Evidence." NBER Working Paper 14106.

**Autor, David H.** 2010a. "Comment on 'A Quantitative Analysis of the Evolution of the U.S. Wage Distribution, 1970–2000.'" In *NBER Macroeconomics Annual 2009*, vol. 24, edited by D. Acemoglu, K. Rogoff, and M. Woodford, 277–99. University of Chicago Press and National Bureau of Economic Research.

**Autor, David H.** 2010b. "The Polarization of Job Opportunities in the U.S. Labor Market: Implications for Employment and Earnings." Washington, D.C.: The Brookings Institution Hamilton Project.

**Autor, David H., David Dorn, and Gordon H. Hanson.** 2011. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." August. http://economics.mit.edu/files/6613.

**Autor, David H., Lawrence F. Katz, and Melissa S. Kearney.** 2008. "Trends in U.S. Wage Inequality: Revising the Revisionists." *Review of Economics and Statistics* 90(2): 300–323.

**Autor, David, Frank Levy, and Richard Murnane.** 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118(4): 1279–1334.

**Beaudry, Paul, and David A. Green.** 2003. "Wages and Employment in the United States and Germany: What Explains the Differences?" *American Economic Review* 93(3): 573–602.

**Bivens, Josh.** 2007. "Globalization, American Wages and Inequality: Past, Present and Future." EPI Working Paper 279, Economic Policy Institute, Washington, D.C.

**Blanchard, Emily, and Gerald Wilmann.** 2011. "Trade, Education, and the Shrinking Middle Class." http://emilyblanchard.weebly.com/uploads/3/2/0/5/320517/multisector-oct19-2011.pdf.

**Blinder, Alan S.** 2006. "Offshoring: The Next Industrial Revolution?" *Foreign Affairs* 85(2): 113–28.

**Borjas, George J., Richard B. Freeman, and Lawrence F. Katz.** 1997. "How Much Do Immigration and Trade Affect Labor Market Outcomes?" *Brookings Papers on Economic Activity*, no. 1, pp. 1–90.

**Cline, William R.** 1997. *Trade and Income Distribution.* Washington, D.C.: Peterson Institute for International Economics.

**Corrado, Carol. A., Charles R. Hulten, and Daniel E. Sichel.** 2005. "Measuring Capital and Technology: An Expanded Framework." In *Measuring Capital in the New Economy,* edited by C. A. Corrado, J. C. Haltiwanger, and D. E. Sichel. University of Chicago Press.

**Corrado, Carol A., Charles R. Hulten, and Daniel E. Sichel.** 2009. "Intangible Capital and U.S. Economic Growth." *Review of Income and Wealth* 55(3): 661–85.

**Costinot, Arnaud, and Jonathan Vogel.** 2010. "Matching and Inequality in the World Economy." *Journal of Political Economy* 118(4): 747–786.

**Davidson, Carl, Lawrence Martin, and Steven Matusz.** 1988. "The Structure of Simple General Equilibrium Models with Frictional Unemployment." *Journal of Political Economy* 96(6): 1267–93.

**Davis, Donald. R., and James Harrigan.** 2007. "Good Jobs, Bad Jobs, and Trade Liberalization." NBER Working Paper 13139.

**Ebenstein, Avraham, Ann Harrison, Margaret McMillan, and Shannon Phillips.** 2009. "Estimating the Impact of Trade and Offshoring on American Workers using the Current Population Surveys." NBER Working Paper 15107.

**Edwards, Lawrence, and Robert Z. Lawrence.** 2010. *Rising Tide: Is Trade with Emerging Economies Good for the United States?* Washington, D.C.: Peterson Institute for International Economics.

**Egger, Hartmut, and Udo Kreickemeier.** 2009. "Firm Heterogeneity and the Labor Market Effects of Trade Liberalization. *International Economic Review* 50(1): 187–216.

**Feenstra, Robert C.** 1998. "The Integration of Trade and the Disintegration of Production." *Journal of Economic Perspectives* 12(4): 31–50.

**Fort, Teresa.** 2012. "Breaking Up Is Hard To Do: Why Firms Fragment Production across Locations." http://econweb.umd.edu/~fort/FORT_JMP.pdf.

**Gabaix, Xavier, and Augustin Landier.** 2008. "Why Has CEO Pay Increased So Much?" *Quarterly Journal of Economics* 123(1): 49–100.

**Goldberg, Pinelopi Koujianou, and Nina Pavcnik.** 2007. "Distributional Effects of Globalization in Developing Countries." *Journal of Economic Literature* 45(1): 39–82.

**Goldin, Claudia, and Lawrence F. Katz.** 2007. "Long-Run Changes in the Wage Structure: Narrowing, Widening, Polarizing." *Brookings Papers on Economic Activity*, no. 2, pp. 135–65.

**Goldin, Claudia, and Lawrence F. Katz.** 2008. *The Race between Education and Technology.* Cambridge: Harvard University Press.

**Goos, Maarten, and Alan Manning.** 2007. "Lousy Jobs and Lovely Jobs: The Rising Polarization of Work in Britain." *Review of Economics and Statistics* 89(1): 118–33.

**Gordon, Robert J., and Ian Dew-Becker.** 2007. "Unresolved Issues in the Rise of Income Inequality." *Brookings Papers on Economic Activity,* no. 2, pp. 169–192.

**Harrison, Ann, John McLaren, and Margaret S. McMillan.** 2010. "Recent Findings on Trade and Inequality." NBER Working Paper 16425, September.

**Heckman, James J., Jora Stixrud, and Sergio Urzua.** 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24(3): 411–82.

**Helpman, Elhanan, Oleg Itskhoki, and Steven. J. Redding.** 2010. "Inequality and Unemployment in a Global Economy." *Econometrica* 78(4): 1239–83.

**International Monetary Fund.** 2008. *World Economic Outlook: Financial Stress, Downturns, and Recoveries.* October 2008. Washington, DC: IMF.

**Jacobson, Louis S, Robert J. LaLonde, and Daniel Sullivan.** 1993. "Earnings Losses of Displaced Workers." *American Economic Review* 83(4): 685–709.

**Jensen, J. Bradford.** 2011. *Global Trade in Services: Fear, Facts, and Offshoring*. Washington, D.C.: Peterson Institute for International Economics.

**Jones, Ronald W.** 1965. "The Structure of Simple General Equilibrium Models." *Journal of Political Economy* 73(6): 557–72.

**Jorgenson, Dale W., Mun S. Ho, and Kevin J. Stiroh.** 2008. "A Retrospective Look at the U.S. Productivity Growth Resurgence." *Journal of Economic Perspectives* 22(1): 3–24.

**Kambourov, Gueorgui, and Iourii Manovskii.** 2009. "Occupational Specificity of Human Capital." *International Economic Review* 50(1): 63–115.

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

**Krugman, Paul R.** 1995. "Growing World Trade: Causes and Consequences." *Brookings Papers on Economic Activity,* no. 1, pp. 327–77.

**Krugman, Paul R.** 2008. "Trade and Wages, Reconsidered." *Brookings Papers on Economic Activity*, no. 2, pp. 103–137.

**Lawrence, Robert Z., and Colin Lawrence.** 1985. "Manufacturing Wage Dispersion: An End Game Interpretation." *Brookings Papers on Economic Activity*, pp. 47–116.

**Leamer, Edward E.** 1995. "The Heckscher–Ohlin Model in Theory and Practice." The Frank D. Graham Memorial Lecture. Princeton Studies in International Finance, No. 77, February.

**Leamer, Edward E.** 2012. *The Craft of Economics: Lessons from the Heckscher–Ohlin Framework.* Ohlin Lectures. Cambridge: MIT Press.

**Lemieux, Thomas.** 2006. "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, and Rising Demand for Skill." *American Economic Review* 96(3): 461–98.

**Liu, Runjuan, and Daniel Trefler.** 2008. "Much Ado about Nothing: American Jobs and the Rise of Service Outsourcing to China and India." NBER Working Paper 14061.

**Liu, Runjuan, and Daniel Trefler.** 2011. "A Sorted Tale of Globalization: White Collar Jobs and the Rise of Service Offshoring." NBER Working Paper 17559.

**McLaren, John, and Shushanik Hakobyan.** 2010. "Looking for Local Labor Market Effects of NAFTA." NBER Working Papers 16535.

**Manasse, Paolo, and Alessandro Turrini.** 2001. "Trade, Wages, and 'Superstars.'" *Journal of International Economics* 54(1): 97–117.

**Markusen, James R.** 2002. *Multinational Firms and the Theory of International Trade*. Cambridge: MIT Press.

**Melitz, Marc J.** 2003. "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71(6): 1695–1725.

**Mishel, Lawrence, Jared Bernstein, and Heidi Shierholz.** 2009. *The State of Working America, 2008/2009*. Washington, DC: Economic Policy Institute.

**Mussa, Michael.** 1974. "Tariffs and the Distribution of Income: The Importance of Factor Specificity, Substitutability, and Intensity in the Short and Long Run." *Journal of Political Economy* 82(6): 1191–1204.

**Neal, Derek.** 1995. "Industry-Specific Human Capital: Evidence from Displaced Workers." *Journal of Labor Economics* 13(4): 653–77.

**Neary, J. Peter.** 1978. "Short-Run Capital Specificity and the Pure Theory of International Trade." *Economic Journal*, Royal Economic Society, 88(351): 488–510.

**Piketty, Thomas, and Emmanuel Saez.** 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118(1): 1–39.

**Piketty, Thomas, and Emmanuel Saez.** 2006. "The Evolution of Top Incomes: A Historical and International Perspective." *American Economic Review* 96(2): 200–206.

**Rodrik, Dani.** 1997. *Has Globalization Gone Too Far?* Washington, D.C.: Institute for International Economics.

**Rosen, Sherwin.** 1981. "The Economics of Superstars." *American Economic Review* 71(5): 845–58.

**Saez, Emmanuel.** 2012. "Striking It Richer: The Evolution of Top Incomes in the United States (Updated with 2009 and 2010 Estimates)." March 2. http://elsa.berkeley.edu/~saez/saez-UStopincomes-2010.pdf.

**Slaughter, Matthew J.** 2000. "What Are the Results of Product-Price Studies and What Can We Learn from Their Differences?" In *The Impact of International Trade on Wages*, edited by Robert C. Feenstra. National Bureau of Economic Research Conference Volume.

**Spence, Michael.** 2011. "Globalization and Unemployment." *Foreign Affairs*, July/August. http://www.foreignaffairs.com/articles/67874/michael-spence/globalization-and-unemployment.

**U. S. Bureau of the Census.** N.d. Table P-18—Educational Attainment, People 25 Years Old and Over by Mean Income and Sex, 1991 to 2010. Excel file. http://www.census.gov/hhes/www/income/data/historical/people/.

**Verhoogen, Eric A.** 2008. Trade, Quality Upgrading and Wage Inequality in the Mexican Manufacturing Sector. *Quarterly Journal of Economics* 123(2): 489–530.

**Yellen, Janet L.** 2006. "Economic Inequality in the United States." Speech to the Center for the Study of Democracy, 2006–2007 Economics of Governance Lecture, University of California, Irvine, November 6. At *Federal Reserve Bank of San Francisco* website: http://www.frbsf.org/news/speeches/2006/1106.html.

Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 141–166

# Why is the Teen Birth Rate in the United States So High and Why Does It Matter?

Melissa S. Kearney and Phillip B. Levine

Teens in the United States are far more likely to give birth than in any other industrialized country in the world. U.S. teens are two and a half times as likely to give birth as compared to teens in Canada, around four times as likely as teens in Germany or Norway, and almost 10 times as likely as teens in Switzerland. Among more developed countries, Russia has the next highest teen birth rate after the United States, but an American teenage girl is still around 25 percent more likely to give birth than her counterpart in Russia. Moreover, these statistics incorporate the almost 40 percent fall in the teen birth rate that the United States has experienced over the past two decades. Differences across U.S. states are quite dramatic as well. A teenage girl in Mississippi is four times more likely to give birth than a teenage girl in New Hampshire—and 15 times more likely to give birth as a teen compared to a teenage girl in Switzerland.

This paper has two overarching goals: understanding why the teen birth rate is so high in the United States and understanding why it matters. Thus, we begin by examining multiple sources of data to put current rates of teen childbearing into the perspective of cross-country comparisons and recent historical context. We examine teen birth rates alongside pregnancy, abortion, and "shotgun" marriage rates as well as the antecedent behaviors of sexual activity and contraceptive use. We seek insights as to why the rate of teen childbearing is so unusually high in the United States as a whole, and in some U.S. states in particular. We argue that explanations that economists have tended to study are unable to account for any sizable

■ *Melissa S. Kearney is Associate Professor of Economics, University of Maryland, College Park, Maryland. Phillip B. Levine is Katherine Coman and A. Barton Hepburn Professor of Economics, Wellesley College, Wellesley, Massachusetts. Both authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their e-mail addresses are ⟨kearney@econ.bsos.umd.edu⟩ and ⟨plevine@wellesley.edu⟩.*

http://dx.doi.org/10.1257/jep.26.2.141. doi=10.1257/jep.26.2.141

share of the variation in teen childbearing rates across place. We describe some recent empirical work demonstrating that variation in income inequality across U.S. states and developed countries can explain a sizable share of the geographic variation in teen childbearing. To the extent that income inequality is associated with a lack of economic opportunity and heightened social marginalization for those at the bottom of the distribution, this empirical finding is potentially consistent with the ideas that other social scientists have been promoting for decades but which have been largely untested with large data sets and standard econometric methods.

Our reading of the totality of evidence leads us to conclude that being on a low economic trajectory in life leads many teenage girls to have children while they are young and unmarried and that poor outcomes seen later in life (relative to teens who do not have children) are simply the continuation of the original low economic trajectory. That is, teen childbearing is explained by the low economic trajectory but is not an additional cause of later difficulties in life. Surprisingly, teen birth itself does not appear to have much direct economic consequence.

Moreover, no silver bullet such as expanding access to contraception or abstinence education will solve this particular social problem. Our view is that teen childbearing is so high in the United States because of underlying social and economic problems. It reflects a decision among a set of girls to "drop-out" of the economic mainstream; they choose nonmarital motherhood at a young age instead of investing in their own economic progress because they feel they have little chance of advancement. This thesis suggests that to address teen childbearing in America will require addressing some difficult social problems: in particular, the perceived and actual lack of economic opportunity among those at the bottom of the economic ladder.

## Documenting Patterns

We begin by describing the sources of data that researchers use to examine teen fertility, along with the sexual activity and contraceptive use that lead to different fertility outcomes. We detail the patterns in these behaviors over time and across locations. Finally, we examine differences in teen fertility and its antecedent behavior by demographic group, including race and ethnicity.

### Data Sources

Researchers focusing on teen fertility in the United States have at least five main sources of data at their disposal. Perhaps the most important are the natality data from the Vital Statistics system, which contain all of the information from birth certificates. The main strength of these data is their universal nature and large sample size; data exist for virtually every live U.S. birth. However, these data are limited to the information on a birth certificate; that is, we know the mother's age and education, race/ethnicity, marital status, birth weight, and a limited number of additional pieces of information, but little else.

The National Surveys of Family Growth (NSFG) provide a wealth of data on the sexual activity, contraceptive use, and pregnancy histories of a national sample of 7,000 to 10,000 women of childbearing age (15 to 44). These surveys were completed in 1982, 1988, 1995, and 2002 (earlier surveys in 1973 and 1976 only included married women, limiting their value for present purposes). Beginning in 2006, the survey design changed from one larger survey every several years to a smaller annual survey; data from 2006–2008 are currently available. The survey oversamples minorities and teens, but is otherwise nationally representative. For an analysis of teen fertility, the sample size of teens in the survey contemporaneously is not that large. On the other hand, pregnancy histories can be used to generate teen fertility outcomes for virtually all women in the sample (excluding those who are, say, only 15 years old on the survey date).

Data from the Youth Risky Behavior Surveillance (YRBS) system provide useful information on the activities of school-age teens. These data are collected biannually for students who respond to the survey at school. Students are asked about their sexual activity and contraceptive use. Respondents are typically between 14 and 18, which is not quite the same age range as a traditional measure of teen fertility—typically ages 15 to 19—but is close enough to draw useful inferences. The sample sizes are reasonably large—in the thousands per year. One disadvantage is that not all states participate in the program in every survey year. Another disadvantage is that the sample does not include high school dropouts, which is a group that may have a higher risk of teen pregnancy.

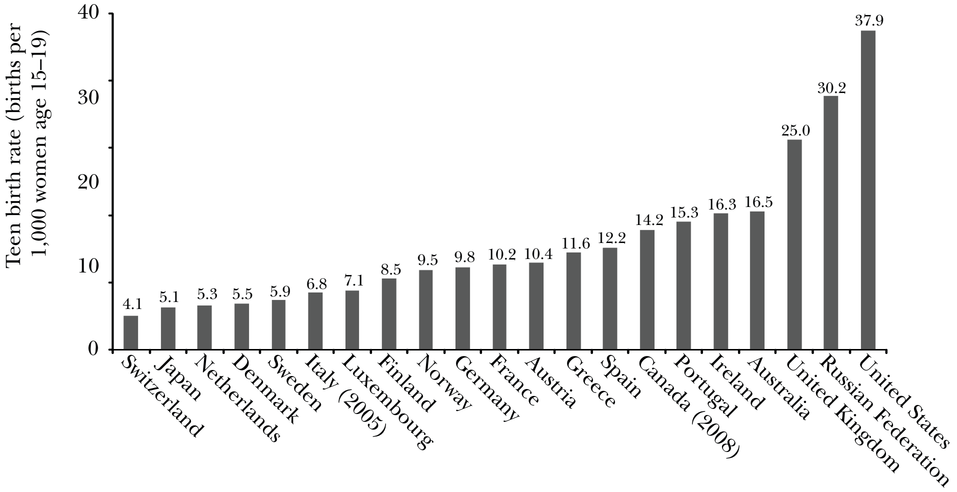
Two additional sources of data are at our disposal.[1] The Guttmacher Institute collects and reports aggregate data on abortions and also aggregate numbers for pregnancy and fertility. The Family and Fertility Survey (FFS) is a dataset that is much like the National Surveys of Family Growth, but it includes a survey like this from several more developed countries. In most countries, the data are available for the early to mid 1990s (the 1995 National Survey of Family Growth is the American contribution to the FFS). In this paper, we use these five datasets.

### Geographic and Time-Series Variation in Teen Fertility Rates

Figure 1 and Table 1 display the substantial geographic variation in the teen birth rate across countries and across U.S. states, respectively, in 2009. Typical developed countries have a teen birth rate in the range of 5 to 15 births per 1,000 women between the ages of 15 and 44. The U.S. teen pregnancy rate is 37.9 in 2009 (although it fell to 34.3 in 2010). Some New England states with teen birth rates under 20 are fairly close to other developed countries and are comparable to that in other English-speaking countries like Australia, Ireland, and the United Kingdom. Some southern states with teen birth rates over 60 are extreme outliers.

[1] The National Longitudinal Survey of Adolescent Health (often called "adhealth") includes similar data, but we have not found this to be a useful source of data for our purposes. The focus of these data is the interrelationships between respondents. This data source is less frequently used for more traditional cross-sectional evidence.

*Figure 1*
**International Comparison of Teen Birth Rates, 2009**



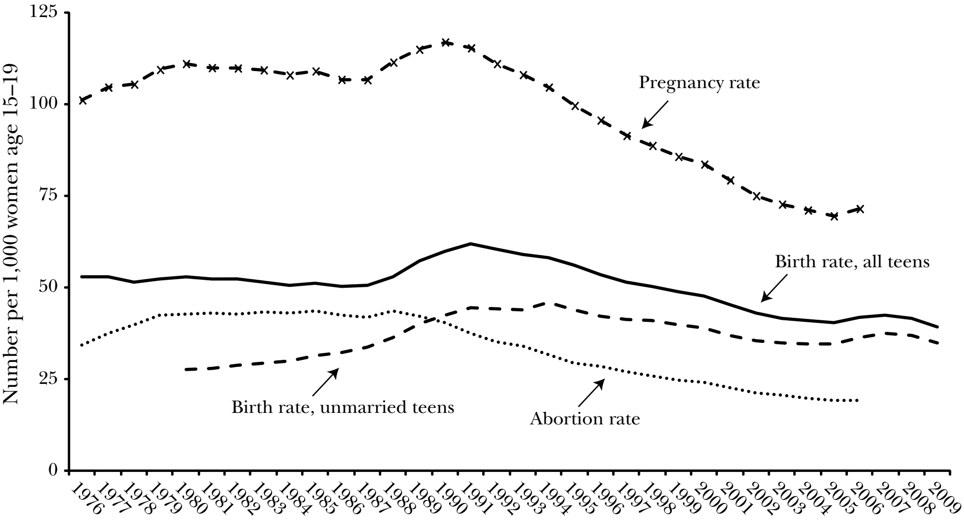
*Sources:* UNECE Statistical Database and *United Nations Demographic Yearbook, 2009–2010.*

*Table 1*
**Teen Birth Rates (Birth per 1,000 Females Age 15–19), 2009**

| *State* | *Teen birth rate* | *State* | *Teen birth rate* | *State* | *Teen birth rate* |
|---|---|---|---|---|---|
| *Low* | | *Moderate* | | *High* | |
| New Hampshire | 16.4 | Michigan | 32.7 | North Carolina | 44.9 |
| Vermont | 17.4 | Oregon | 33.1 | Wyoming | 45.0 |
| Massachusetts | 19.6 | Nebraska | 34.6 | Nevada | 47.4 |
| Connecticut | 21.0 | Delaware | 35.3 | Washington, DC | 47.7 |
| New Jersey | 22.7 | Idaho | 35.9 | Georgia | 47.7 |
| Minnesota | 24.3 | Illinois | 36.1 | South Carolina | 49.1 |
| Maine | 24.4 | California | 36.6 | West Virginia | 49.8 |
| New York | 24.4 | South Dakota | 38.4 | Arizona | 50.6 |
| Rhode Island | 26.8 | Colorado | 38.5 | Tennessee | 50.6 |
| North Dakota | 27.9 | Montana | 38.5 | Alabama | 50.7 |
| Pennsylvania | 29.3 | Ohio | 38.9 | Kentucky | 51.3 |
| Wisconsin | 29.4 | Florida | 39.0 | Louisiana | 52.7 |
| Utah | 30.7 | Hawaii | 40.9 | Arkansas | 59.2 |
| Virginia | 31.0 | Missouri | 41.6 | Oklahoma | 60.1 |
| Maryland | 31.3 | Indiana | 42.5 | Texas | 60.7 |
| Washington | 31.9 | Kansas | 43.8 | New Mexico | 63.9 |
| Iowa | 32.1 | Alaska | 44.5 | Mississippi | 64.2 |

*Source:* Martin et al. (2011).

*Figure 2*
**Trends in the Teen Pregnancy, Abortion, and Birth Rate**



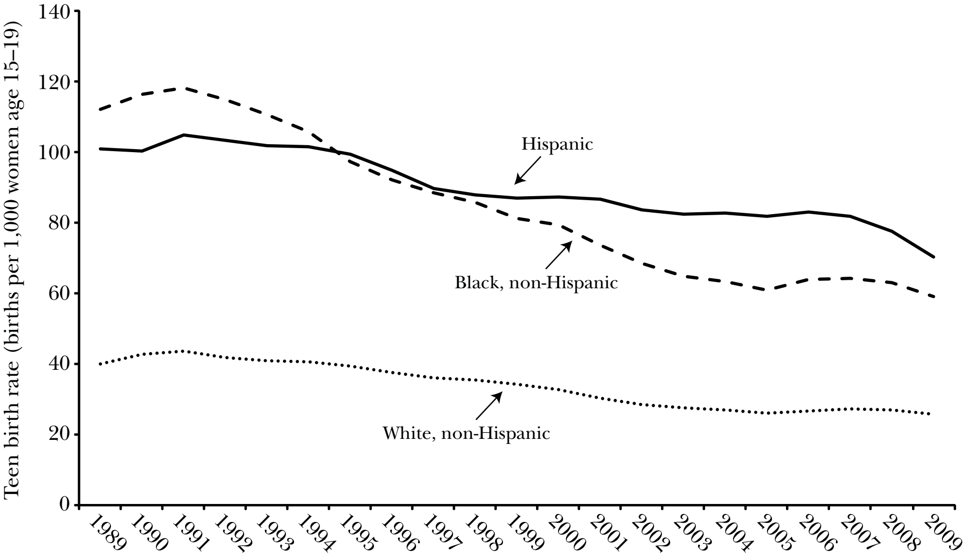
*Sources:* Martin, Hamilton, Sutton, Ventura, Mathews, and Osterman (2010), Hamilton, Martin, and Ventura (2010), and Guttmacher Institute (2010).

Figure 2 shows the teen birthrate since 1976, using Vital Statistics data.[2] The teen birth rate holds roughly constant through the late 1970s and most of the 1980s between 50 and 55 births per 1,000 women between the ages of 15 and 19. A large blip developed in the time series beginning in the late 1980s, and the teen birth rate rose to a level of around 60 births per 1,000 teenage women in the early 1990s. It has been generally declining since then. The teen birth rate was 37.9 per thousand in 2009, down from the peak of 61.8 in 1991. Figure 2 also shows that the composition of teen births has shifted dramatically towards unmarried women. The birth rate among unmarried teens used to be considerably lower than that for all teens, but in 2009, 87 percent of teen births were to unmarried mothers (Martin et al. 2011).

Trends in teen births can be driven by changes in the likelihood of a pregnancy or changes in the likelihood of aborting a pregnancy once it occurs (we assume miscarriage rates are roughly stable over time). Figure 2 also displays trends in the pregnancy and abortion rate, again as measured per 1,000 women age 15 to 19. Pregnancies and abortions were roughly flat during the period in which teen births were flat through the late 1980s. During this period, roughly 10 percent of teens got pregnant and 4 percent had an abortion each year. The spike in teen births in the early 1990s was driven almost entirely by an increase in the pregnancy rate; almost 12 percent of teens got pregnant at the peak in 1990. Clearly, the recent decline in teen births is not attributable to greater use of abortion; instead, it is the result of

[2] Although the data go back further than that, we focus on the period after the 1960s and early 1970s, which saw dramatic legal changes in access to abortion and contraception and the sexual revolution.

*Figure 3*
**Trends in the Teen Birth Rate by Race/Ethnicity**



*Source:* Martin et al. (2011).

fewer teens getting pregnant. More recently, 7 percent of teens got pregnant and 2 percent had an abortion each year.

Fertility outcomes for teens differ dramatically by race and ethnicity. Figure 3 displays teen birth rates for Hispanics, non-Hispanic blacks, and non-Hispanic whites beginning in 1989 (the first year in which race/ethnicity were separately identified). Teen birth rates for white, non-Hispanic women have been considerably lower than the other groups over the entire period, falling in the range of 25 to 40 or so births per 1,000 women as opposed to rates that are two to three times that for the other groups. Although all groups have experienced a substantial decline in teen birth rates since the early 1990s, the decline for black, non-Hispanic teens has been particularly notable. For this group, the teen birth rate fell by half from 118.2 to 59 between 1991 and 2009. Presently, Hispanic teens have the highest teen birth rate at a level of 70.1 per thousand, which is nonetheless a sizable decline from a rate of 104.6 per thousand in 1991. For all race/ethnic groups, the declines in teen births are driven entirely by declines in pregnancies, not increases in abortion (Guttmacher Institute 2010).

### Trends in Fertility Outcomes by Age 20

All of the statistics reported in the preceding section are based on the behavior of teens in a given year. A related, but not identical, issue is the behavior of women over all of their teen years. For purposes of illustration, consider 100 teenage girls between the ages of 15 and 19, with 20 girls at each particular age and 5 percent give birth in a given year. It could be that only 5 percent of teens *ever* give birth as a teenager—say one teen at each age (15, 16, 17, 18, and 19)—if the same mothers give birth year after year. Alternatively, if the only teens who give birth do so at age 19, then if 5 percent of teens give birth per year, it would imply that 25 percent of the 19 year-olds give birth, and 25 percent of women would end up giving birth by the time they reach age 20. In this section, and in subsequent analyses, we use individual-level data from the National Surveys of Family Growth to construct cumulative statistics by birth cohort.

Roughly 20 percent of women who have turned age 20 over the past 30 or so years have given birth as a teen. This statistic has fluctuated similarly to what we see with point-in-time statistics in Figure 2, with a spike for the cohorts hitting age 20 in the early 1990s and declining in recent years. The comparison of a 5 percent annual teen birth rate to a typical rate of teen childbearing for a birth cohort of around 20 percent suggests that roughly 20 percent of teen births are not first births. For the most recent cohorts, the likelihood of giving birth as a teen has fallen to 17 percent. Again, mothers who have never married have become more prominent among those giving birth as teens.

We can also divide these data by the level of education obtained by the mother of the teenager. Although the exact composition of teens across maternal education categories varies by birth cohort, as an approximation 30 percent of teen mothers have mothers who dropped out of high school, 40 percent have mothers who are high school graduates, and 30 percent have mothers who attended college. Daughters of women who have dropped out of high school have children as teens at a rate in the vicinity of 33 percent. Daughters of women who have attended some college do so at around one-third this rate. However, the middle group—that is, daughters of women who are high school graduates but have not attended any college—is the group driving the rise and fall in the overall teen birth rate. This middle group saw by far the sharpest rise in the probability of giving birth by age 20 from 19 percent in 1990 to 29 percent in 1998, before falling back to 16 percent by 2006.

The National Surveys of Family Growth data also allows us to investigate how each pregnancy was resolved. We focus on those teen pregnancies that began when the women were unmarried and track whether they led to a nonmarital birth, a marital birth, a miscarriage, or an abortion. The dramatic change here is that so-called "shotgun marriages" (meaning those that take place after the pregnancy but before the birth) have fallen throughout the period as nonmarital births have risen.[3] In the 1970s, nearly 40 percent of all nonmarital pregnancies resulted in a shotgun marriage. Now that rate has fallen under 10 percent. Conversely, nearly 40 percent of nonmarital pregnancies resulted in a nonmarital birth in 1976, but since the mid 1990s that has been in the range of 65–70 percent of all teen pregnancies. With fewer abortions and fewer shotgun marriages among pregnant, unmarried women over the past few decades, a substantial majority of unmarried pregnant teens now give birth outside of marriage.

[3] Akerlof, Yellen, and Katz (1996) show that these trends for all women (not just teens) have been falling since the 1960s.

*Table 2*
**Rates of Pregnancy, Birth, and Abortion across Countries and States in the United States**

| | *Pregnancies (per 1,000)* | *Births (per 1,000)* | *Abortions (per 1,000)* | *% of pregnancies aborted* |
|---|---|---|---|---|
| Denmark (2003) | 24 | 5 | 15 | 63.2 |
| Germany (2003) | 23 | 12 | 7 | 31.1 |
| New Hampshire (2005) | 33 | 18 | 11 | 33.3 |
| United Kingdom (2003) | 59 | 27 | 23 | 38.8 |
| United States (2005) | 70 | 40 | 19 | 27.1 |
| Mississippi (2005) | 85 | 61 | 11 | 12.9 |

*Sources:* State data are from Guttmacher Institute (2010). International birth data are from the UNECE statistical database. International abortion data are from Sedgh, Henshaw, Singh, Bankole, and Drescher (2007).

The final category here is pregnancies not carried to term, which can include either an abortion or a miscarriage. Abortions are notoriously underreported in survey data, so by including all pregnancies that are not resolved by birth, we do not have to worry about separately identifying abortions and miscarriages. Under the assumption that the rate of miscarriages has been roughly constant over time, we can interpret trends in so-called uncompleted pregnancies as being driven by changes in the use of abortion. We see an increase in the rate of uncompleted pregnancies from 20–25 percent in 1976 and 1977 to about 30–35 percent in the mid 1980s, followed by a decline back to about 20–25 percent by the mid 1990s and relative stability in this outcome since then.

### Geographic Comparisons of Pregnancies and Abortions

How does the variation in teen birth rates across countries and U.S. states reflect differences in pregnancy rates and abortion rates?[4] We present some summary data in Table 2. For our international comparison, we highlight numbers from the United States, United Kingdom, Germany, and Denmark. This set is chosen because abortion data is available for all four of these countries, and because the countries span most of the range of country teen birth rates reported in Figure 1. We also report numbers from Mississippi and New Hampshire, the U.S. states with the highest and lowest teen birth rates.

The main finding that emerges from these data is that pregnancy rates across locations line up very closely with birth rates. Differences in pregnancy rates appear to be the primary driver of differences in birth rates. Nonetheless, there are some

[4] Pregnancy rates are reported by the Guttmacher Institute (2010) for the United States and for separate states. For the other countries, we take the ratio of the estimated number of pregnancies to the sum of abortions and births in the United States as an adjustment factor to account for miscarriage and apply that ratio to the sum of abortions and births in the other countries. This approach assumes a constant miscarriage rate across countries.

interesting patterns in abortion rates. For instance, the lower rate of abortion in Germany relative to Denmark means that in Germany more births result from fewer pregnancies. The United States also has a relatively low rate of abortion conditional on pregnancy, as compared to the other three countries. This pattern holds despite the fact that abortion laws are relatively more lenient in the United States, certainly relative to Germany and the United Kingdom (Levine 2004). We also see that the lowest teen birth rate U.S. states, like New Hampshire, have teen pregnancy and abortion rates that are comparable to many other developed countries. In Mississippi, in contrast, 8 percent of female teens become pregnant each year, and few have abortions.

## The Roles of Sexual Activity and Contraceptive Use

The earlier discussion (for example, of Figure 2 and Table 2) shows that the dramatic decline in teen births in the United States and the variation across countries is largely (although perhaps not exclusively) attributable to patterns in teen pregnancy, not use of abortion. This section explores to what extent the geographic variation in the likelihood of teen pregnancy is generated by a variation in teen sexual activity versus rates of contraceptive use among those who are sexually active.[5]

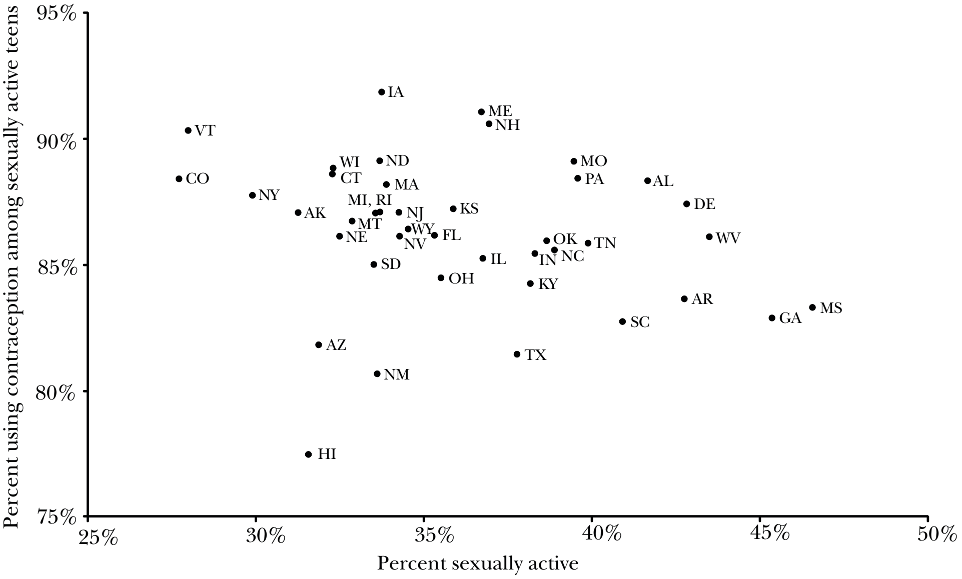
### Descriptive Statistics

Figure 4 displays a scatter plot across U.S. states of the percentage of teens that are sexually active—that is, who have engaged in sexual intercourse in the past three months—and the percentage of sexually active teens who used a form of contraception the last time they had sex. To calculate these statistics, we used data from the 2007 and 2009 Youth Risky Behavior Surveillance surveys, which contain information on these outcomes for 41 states. On average, 36 percent of teens are sexually active, but that statistic ranges from 28 percent in Colorado and Vermont to over 45 percent in Georgia and Mississippi. Contraceptive use is very high in these data; an average of 86 percent of teens that had sex in the last three months used some form of contraception at last intercourse. Again, a tremendous amount of variation across states exists, ranging from a high of over 90 percent in states like Maine, New Hampshire, and Vermont—which are among the lowest teen birth rate states—to 81 percent in states like Texas and New Mexico, which are among the highest teen birth rates states. Interestingly, the two statistics across states are negatively correlated, although only weakly ($r = -0.21$). These comparisons suggest the importance of both sexual activity and contraceptive use in driving variation in teen birth rates; we conduct a more formal decomposition of the relative importance of the two components subsequently.

[5] In a companion paper (Kearney and Levine 2012), we focus on the relative contributions of changes in sexual activity versus contraceptive use in understanding the downward trend in teen birth rates in the United States.

*Figure 4*

**Rates of Sexual Activity and Contraceptive Use among School-Aged (14–18) Girls**



*Source:* Authors calculations from the 2007 and 2009 Youth Risky Behavior Surveillance survey state microdata.

International comparisons suggest that contraceptive use among U.S. teens is lower than in other developed countries (although the data available for such an exercise is somewhat limited), and this more than offsets the lower rate of sexual activity among U.S. teens, leading ultimately to a higher birth rate for U.S. teens. The most recent data for all teens, reported in Darroch et al. (2001), is now 15 to 20 years old. At that time, teens in the United States were somewhat less likely to be sexually active than teens in other countries. In the United States, 58.7 percent of teens had intercourse in the past three months, whereas 62.2 percent, 63.9 percent, and 78.7 percent had sex in Great Britain, France, and Sweden, respectively. Contraceptive use was lower among U.S. teens, however, with 20 percent of teens failing to use any form of contraception at last intercourse, compared to 4 to 12 percent in these same three countries. More recent data from the early 2000s reported in Godeau, Gabhainn, Vignes, Ross, Boyce, and Todd (2008) and Santelli, Sandfort, and Orr (2008) focus explicitly on the behavior of 15 year-olds using data from the Youth Risky Behavior Surveillance in the United States and the Health Behaviour in School-Aged Children study in other countries. These data show that 18, 29, and 40 percent of 15 year-olds have ever had sex in France, Sweden, and England whereas our own calculation suggests that the rate in the United States is 23 percent, again suggesting that U.S. teens are toward the lower end of this distribution. In terms of contraception, Godeau, Gabhainn, Vignes, Ross, Boyce, and Todd (2008) report that 74, 83, and 90 percent of 15 year-olds in Sweden, England, and France used either the condom or pill at last sexual intercourse, as compared to only 66 percent of 15 year-old girls in the United States. Of course, the focus on 15 year-olds in this comparison is somewhat limiting.

Within the United States, teen sexual activity is trending downward and contraceptive use is trending upward. Based on our analysis of National Surveys of Family Growth and Youth Risky Behavior Surveillance data over roughly the past two decades, we find that the percentage of teens that are sexually active has fallen from around 40 to 33 percent since around 1990 (although the precise levels and specific years are somewhat different across datasets). The prevalence of contraceptive use at last intercourse has risen in each survey from around 80 percent to 85 percent. Patterns of sexual activity and contraceptive use across racial and ethnic groups are generally consistent with differences in levels and trends in fertility outcomes across these groups.

## Decomposition of Fertility Rates into Sexual Activity and Contraceptive Use

In this section, we undertake a decomposition exercise to determine the extent to which differences in the teen birth rate across regions can be attributed to teen sexual activity or contraceptive use.[6] We emphasize that such an exercise only serves to understand the mechanistic drivers of teen birth rates. A more substantive analysis regarding the underlying reason *why* teens behave differently across time and place requires more sophisticated methods that can enable us to draw causal conclusions about behavioral responses.

We take advantage of Youth Risky Behavior Surveillance (YRBS) data between 1991 and 2007 and the observed teen birth rates in the subsequent years (1992–2008) by state to relate sexual activity and contraceptive use to teen fertility.[7] Data from the YRBS reflect the behavior of high school age (14 to 18 year-old) females. If they become pregnant, the birth would occur nine months later, at which point the vast majority would fit into the 15-to-19 year-old category that is captured by traditional measures of teen fertility. So we link the YRBS survey year with the teen birth rate in the following year. (We are not able to use the 2009 YRBS, which is currently available but cannot be linked to 2010 teen birth rates, as these are not yet available by state.) We then estimate regression models of the state-year teen

[6] Santelli, Lindberg, Finer, and Singh (2007) focus on the relative contributions of the two types of behavioral changes to describe the trend towards lower teen fertility. They start with the rates of sexual activity in the past three months and the likelihood of using contraception at last intercourse, but they focus specifically on the method of contraception used. Then they make a series of assumptions based on the percentage of teens who would get pregnant if they used no contraception (which is a function of sexual activity and the likelihood of pregnancy in its absence) and the effect of various forms of contraception on reducing that percentage. Then they apply observed changes in rates of sexual activity and contraceptive use to predict how much each one would have reduced the pregnancy rate. Comparing these predictions with actual declines enables them to decompose the overall decline into the two components. The results of their analysis suggest that about three-quarters of the decline in teen birth rates between 1995 and 2002 are attributable to greater use of contraception. We take a totally different approach to this question that is far less reliant on assumptions.

[7] Using analogous methods, in Kearney and Levine (2012), we find that reduced sexual activity and increased contraceptive use can explain 35 percent and 65 percent, respectively, of the decline in teen childbearing.

birth rate as a function of measures of sexual activity and contraceptive use by state-year to determine how changes in these behaviors are linked to changes in teen birth rates.[8]

The results of regressions based on this analysis are presented in Table 3. The dependent variable in each model is the probability of giving birth as a teen in a year. The independent variables are share of teenagers in a state in a given year who have engaged in sexual activity in the previous three months along with measures of alternative contraceptive choices. The left panel focuses on the broader measure of whether any contraception was used at last intercourse, while the right panel of the table breaks up methods of contraception into specific forms. Estimates from these models indicate the effectiveness of alternative forms of contraception in reducing pregnancy as used in practice for this sample of teenagers and do not necessarily correspond to their reported efficacy when used as prescribed.

From the left panel of Table 3, the estimated coefficient on using contraception if sexually active is actually larger in absolute value than the estimated coefficient on being sexually active. This result, taken literally, does not make sense—since abstinence has to be the strongest form of contraception. This puzzle could be explained, for instance, by a lower rate of sexual activity among those who use contraception. However, the two estimated coefficients are not statistically distinguishable. These results indicate that a woman who reports engaging in sexual activity in the preceding three months has about a 16 percent chance of getting pregnant and giving birth. Using contraception can dramatically reduce the probability of getting pregnant, perhaps to the point of eliminating it entirely. The results in the right panel indicate that the pill is the most effective form of birth control. Condoms are found to be about three-quarters (.120/.156) as effective as the pill. Although we do not find that Depo Provera is particularly effective, this is probably attributable to the very low rates of its use in these data, suggesting that the measurement of its use may be error-ridden. Use of the withdrawal method is not found to be related to a reduction in teen fertility (although this could be attributable to other factors—perhaps those using withdrawal are having sex more often).

We apply the results in the left panel of the table to determine the relative contribution of reduced sexual activity and increased contraceptive use in explaining differences in teen birth rates across states and countries. Based on our earlier discussion, as a rough approximation, American teens are 10 percentage points less likely to be sexually active and 15 percentage points less likely to use contraception if they are sexually active. Our regression results indicate that the lower rate of sexual activity would reduce the teen birth rate by about 1.6 percentage points (10 × 0.162) and the lower rate of contraceptive use would increase the teen birth rate by 2.8 percentage points (15 × –0.186). On net, teen births would be

[8] We use the size of the teen population in each state-year cell as weights in these regressions. State-level data in the Youth Risky Behavior Surveillance (YRBS) data are not available in every state in every year, and some states do not ask sexual activity questions in some (or all) years. In total, our dataset for this exercise is comprised of 167 state-year pairs. We used YRBS microdata to construct the state-year aggregates; these data are not publicly available, but can be obtained from the Centers for Disease Control.

*Table 3*
**Mechanical Correlations with Teen Fertility**
*(standard errors in parentheses)*

| *Sexual activity and use of any contraception* | | *Sexual activity and specific forms of contraception* | |
|---|---|---|---|
| % Any sexual activity in past 3 months | 0.162<br>(0.017) | % Any sexual activity in past 3 months | 0.151<br>(0.019) |
| % Used any contraception if sexually active | –0.186<br>(0.030) | % Used pill if sexually active | –0.156<br>(0.028) |
| | | % Used condom if sexually active | –0.120<br>(0.023) |
| | | % Used Depo or other if sexually active | –0.018<br>(0.043) |
| | | % Used withdrawal if sexually active | 0.022<br>(0.062) |
| $R^2$ | 0.64 | $R^2$ | 0.71 |
| Number of states/years | 167 | Number of states/years | 167 |

*Source:* Authors using data from the Youth Risky Behavior Surveillance survey and Vital Statistics natility data.
*Notes:* We estimate regression models of the state-year teen birth rate as a function of measures of sexual activity and contraceptive use by state-year. The dependent variable in each model is the probability of giving birth as a teen in a year. The independent variables are share of teenagers in a state in a given year who have engaged in sexual activity in the previous three months along with measures of alternative contraceptive choices. All regressions are weighted by the population of women age 15 to 19 in each state-year. Withdrawal is counted as a form of contraception.

1.2 percentage points higher in the United States, which translates to a teen birth rate that is 12 points higher per 1,000 than in other countries. Based on the statistics reported in Figure 1, the observed difference is considerably greater than that.

Across states, the data suggest that it would not be out of the question for some states to have a 10 percentage point higher rate of teen sexual activity and a 10 percentage point lower rate of contraceptive use compared to other states. Applying the above approach, our estimates would generate corresponding teen birth rates that are 16 points higher per 1,000 for the higher sexual activity and 18 points higher for the lower rates of contraceptive use, totaling 34 points per 1,000. This is close to the roughly 40 point per 1,000 gap that exists between high and low teen birth rate states. It suggests that these differences are roughly equally attributable to differences in sexual activity and contraceptive use.

For some analysts, these results would lead immediately to a call for more abstinence education and/or increased accessibility of contraception.[9] However,

[9] Boonstra (2002, p. 8) provides an example: "If recent declines in teen childbearing are the result of fewer teens getting pregnant in the first place, the obvious next question is: why? Are fewer teens avoiding pregnancy by abstaining from sex, or are those who are having sex using contraception more

jumping from a mechanistic decomposition to a policy recommendation ignores the underlying causes of any changes in behavior. Why did contraceptive use rise in the United States? Why are sexually active teens in Europe more likely to use contraception? It is not obvious that teens in the United States have more information or better access now than they used to. Nor is it obvious that information or access is better for teens in Europe. If we randomly assigned some U.S. teens to have greater access to contraception, would it affect the rate of childbearing among these teens? Perhaps not. To understand more fully what is driving the patterns in teen fertility that we observe, we need to go further and understand why teens in some places or in some years are more or less likely to use contraception or to abstain from sex.

## Standard Models, Prescriptions, and Evidence

The standard economic model of childbearing considers an individual who maximizes utility over children and other consumption subject to a budget constraint (for example, Becker and Lewis 1973). Preferences are generally assumed to be fixed, and explanations have focused on differences in constraints, like policies making welfare more or less attractive, policies making abortion more or less readily available, and policies increasing access to low-cost contraception. Moffitt (1998, 2003) reviews the evidence on the link between welfare benefits and nonmarital childbearing, including teen childbearing. The general consensus is that more generous welfare benefits have a modest positive effect on rates of nonmarital childbearing. However, the lower rate of teen childbearing in Europe with its much more generous welfare system provides a counterexample and prima facie case against the hypothesis that social support is largely to blame for high rates of teen childbearing in the United States. It also appears that the redesign of welfare reform in 1996 had only minor effects on rates of teen childbearing, at best (for example, Kearney 2004; Grogger and Karoly 2005).

Levine (2004) reviews the evidence on the link between abortion policy and fertility outcomes and finds that restrictive abortion policies such as parental notification laws or mandatory delay periods are not associated with higher rates of teen childbearing. In Kearney and Levine (2009), we examine expanded access to Medicaid family planning services during the 1990s and early 2000s. We find that it led to statistically significant reductions in teen childbearing, on the order of a 4 percent reduction. But this effect is not sufficiently large such that one could conclude that limited access to free contraception is a primary driver of teen childbearing rates or that expanding access further would drive the rates down to European levels, for example.

successfully? Not surprisingly, the answer is: both. But deconstructing that answer is critical, because it goes to the heart of a number of relevant and timely public policy questions, among them the debate over public funding for abstinence-only education and for more-comprehensive approaches." Santelli, Linberg, Finer, and Singh (2007) express similar views.

In Kearney and Levine (2012), we econometrically investigate the effects of an extensive list of state-level policies on state-level teen birth rates using Vital Statistics birth data between 1981 and 2008. The list of policies we examine include: welfare benefit levels and welfare reform, expansions of Medicaid family planning services to non-Medicaid recipients; legal abortion restrictions; implementation of the State Children's Health Insurance Program (S-CHIP) (which provides health insurance to children and teens who do not otherwise qualify for Medicaid under traditional guidelines); a measure of federal abstinence education funding; and indicator variables for whether the state mandates sex education or contraception counseling. We also consider the role of state-level economic factors, including the state minimum wage and the unemployment rate. We include a rich set of demographic controls, including the racial/ethnic composition of the state female teenage population, and broader population characteristics, including the percent married, the percent aged 15–19, and the educational composition of the state population. We control for mean differences across states, shared year effects, and state-specific time trends. We estimate regressions for all births, and then separately for demographic subgroups defined by age, race/ethnicity, and marital status. We also examine abortion and pregnancy outcomes, though those analyses are limited by data and generally do not yield statistically precise estimates.

The results of this analysis are consistent with past research. We find that the Medicaid Section 1115 family planning waivers, which expanded eligibility for publicly funded family planning services, reduce teen fertility. Lower levels of welfare benefits also lead to lower rates of teen childbearing. Neither of these effects are sufficiently large to explain the extent of geographic variation we observe in the data or the dramatic drop in teen birth rates over the past 20 years. Other policy interventions, including abortion policies, sex education, abstinence education, and S-CHIP implementation are not found to have a statistically significant, causal impact on teen fertility.

In short, the standard policy prescriptions that are often advocated to alter rates of teen childbearing do not come close to explaining the extensive geographic variation that exists. To be sure, this evidence supports expanding access to free or subsidized contraception for teenagers, but it will not come close to a full solution for teen childbearing. Given that these policies cannot explain much cross-state variation, we are skeptical that a comparable analysis using similar policy variation across countries (difficult as it would be to construct these measures) would explain cross-country differences.

## The Economics of Marginalization and Hopelessness

Social scientists outside economics have been studying and debating the causes and consequences of early nonmarital childbearing for nearly half a century. An early report made famous by its lead author, Daniel Patrick Moynihan (U.S. Department of Labor 1965) drew attention to the issue of nonmarital childbearing among black families in the United States, when the rate was one in three. Moynihan emphasized the declining economic prospects of men as an important factor leading black women to have births outside marriage. The social theories of the psychologist Clark (1965), the "Culture of Poverty" explanation of the anthropologist Lewis (1969), the "social isolation" theory of sociologist Wilson (1987), and more recently, the ethnographic account of Edin and Kefalas (2005) all highlight how growing up in an environment where there is little chance of social and economic advancement can lead young women to have babies outside of marriage.[10]

A number of studies have documented the relationship between background disadvantage and rates of early childbearing (for example, Duncan and Hoffman 1990; An, Haveman, and Wolfe 1993; Lundberg and Plotnick 1995; Duncan, Yeung, Brooks-Gunn, and Smith 1998). Girls who grow up in poverty or in a single-parent household are nearly twice as likely to have a teen birth as girls without these background disadvantage factors. We described differences by level of maternal education earlier. In an examination of cohort rates of early childbearing, we find that the proportion of a female cohort born economically disadvantaged—as captured by being born to a teen mother, a single mother, or to a mother with a low level of education—is tightly linked to the subsequent rate of early childbearing in that cohort (Kearney and Levine 2010). But, strikingly, we find that state and year of birth fixed effects capture much of the variation. We interpret that finding as suggestive of the importance of some "cultural" dimension, which is largely unmodeled in the typical economics framework.

In recent work, we have sought to open this black box of fixed "cultural" differences across states and countries (Kearney and Levine 2011). We consider how the economic circumstances in a place affect the decisions of girls growing up economically disadvantaged. One of our goals is to operationalize notions like "marginalization" and "hopelessness," emphasized in the anthropological and ethnographic research mentioned earlier, in a parsimonious economic model. Our model of early nonmarital childbearing rests on the notion that young women with limited opportunities to advance socially and economically—either through human capital investments or the marriage market—will be relatively more likely to choose early nonmarital childbearing, as compared to other women. This choice is modeled as a utility maximization problem based on a trade-off between the current period satisfaction associated with a baby and the potential long-term cost associated with foregone economic opportunities. The intuition is that if girls perceive their chances at long-term economic success to be sufficiently low even if they do

[10] Some of the explanations for teen childbearing seem too universal to explain the striking differences in rates of early nonmarital childbearing across socioeconomic groups, over time, or across states or countries. For example, developmental psychologists have suggested that teen childbearing is attributable to teens' stage of cognitive development, arguing that they are not quite ready to make the types of decisions that would prevent a pregnancy (Brooks-Gunn and Furstenberg 1989; Hardy and Zabin 1991; Brooks-Gunn and Paikoff 1997). Behavioral economists O'Donohue and Rabin (1999) suggest that teens are "hyperbolic discounters" who place disproportionate weight on present happiness as compared to future well-being. But we doubt that the particularly high rate of teen childbearing among U.S. teens in certain states rather than others, or for U.S. teens as compared to their counterparts in Europe can be attributed to the more limited decision-making capacity—or more present-biased preferences—of the teenage brain at certain times or in certain places.

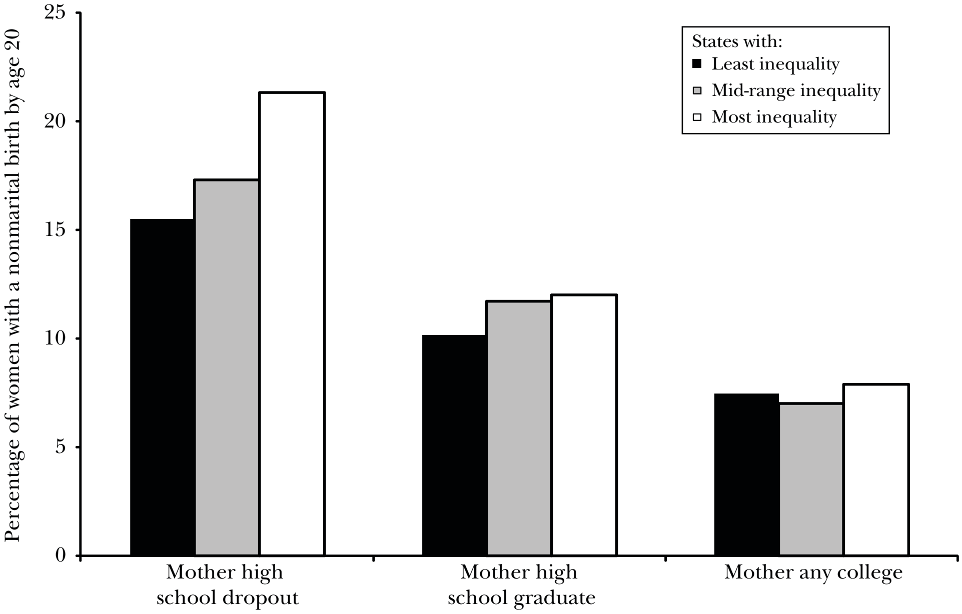
"play by the rules," then early childbearing is more likely to be chosen. We speculate that the combination of being poor and living in a more unequal (and less mobile) society contributes to a low perception of possible economic success and hence leads to choices that favor short-term satisfaction—in this case, the decision to have a baby when young and unmarried.

Our main empirical analysis examines whether women with low socioeconomic status are more responsive to differences in the level of income inequality in terms of their childbearing and marital outcomes. We use individual-level data from the National Survey of Family Growth to look across U.S. states. We also use individual-level data from the Fertility and Family Survey, conducted by the United Nations, to look across a set of roughly a dozen developed countries. An illustrative example of the results of our analysis is provided in Figure 5, focusing on cross-state variation in teen births within the United States. We use the level of education attained by the mother of each teen to separate them into different categories of socioeconomic status and then we divide the states into high-, middle-, and low-inequality categories based on the 50/10 ratio of household income (the ratio of the income at the median of the income distribution to the income at the 10th percentile of the income distribution). Among teens with high socioeconomic status whose mothers attended college, we observe no difference in the likelihood of giving birth as a teen across these states, despite the reasonably large number of these women who do so. Among teens with lower socioeconomic status, though, there is a clear pattern of teen fertility across inequality categories. Teens in the highest-inequality states are roughly 5 percentage points more likely to give birth as a teen than teens in the lowest-inequality states. We find the opposite pattern when we focus on abortions as a teen—much less-frequent abortions among teens with low socioeconomic status in high-inequality states—and no pattern like this when we repeat this exercise for sexual activity. We have also conducted a similar exercise at the county level and obtained similar results. We also obtain analogous results in a cross-country analysis.

One potential concern in an analysis like this is that income inequality at the state level may be capturing any other state fixed factor that happens to be correlated with it, suggesting that inequality may not be the causal determinant of teen fertility. Although it is impossible to rule out this alternative completely, we have experimented with including other conditions that could lead to economic "despair," such as poverty concentration, the incarceration rate, absolute levels of deprivation, and others. We also consider other potentially confounding factors that would be outside the proposed model, such as measures of state religiosity, political preferences, and a measure of social capital. None of these additional factors are found to alter the estimated relationship between inequality and teen fertility among women with low socioeconomic status.

Thus, we conclude that women with low socioeconomic status have more teen, nonmarital births when they live in higher-inequality locations, all else equal. The proximate mechanism driving this finding is that conditional on getting pregnant, more of these girls choose to carry their pregnancy to term. Indeed, our estimates suggest that income inequality can explain a sizable share of the geographic variation

*Figure 5*
**Rate of Nonmarital Childbearing by Age 20, by Mother's Level of Education and State Level of Income Inequality**



*Source:* Authors using data from the 1982, 1988, 1995, 2002, and 2006–2008 National Surveys of Family Growth.

observed in the teen childbearing rate, on the order of 10 to 50 percent. We believe these results are consistent with the large body of work in other social science disciplines arguing that social marginalization and hopelessness are to blame for young, nonmarital childbearing. To the extent that greater levels of inequality are associated with a heightened sense of economic despair and marginalization, our empirical findings support this claim. Certainly, additional research into this link is warranted. This explanation is one of the first that has the potential to explain any sizable share of the geographic variation in teen childbearing.

## Teen Childbearing as Cause or Effect?

A premise of most public discussions about teen fertility is that having a baby as a teen *leads* to inferior outcomes for the mother and the child. Indeed, women who give birth during their teenage years are more likely than other women to drop out of high school, to remain unmarried, and to live in poverty. The children of teenage mothers fare worse than other children on economic, social, and cognitive dimensions (Hoffman and Maynard 2008). If teen childbearing causes large adverse consequences, then the natural response is to consider policies that can potentially reduce the likelihood of a teen pregnancy: sex education, abstinence promotion, improved contraceptive access, and related interventions. Kearney (2010) reviews the evidence on the effectiveness of various teen pregnancy prevention programs. Alternatively, however, it could be that outside factors may *cause* both a teen to give birth and to have inferior outcomes. For example, if those who lack economic opportunity are more likely to give birth as a teen, they and their children are likely to have inferior outcomes regardless of when they give birth. Providing free contraception, for instance, could (modestly) reduce the likelihood of giving birth as a teen, but it does not alter the underlying calculus that leads disadvantaged women to "drop out" of the mainstream climb toward economic and social prosperity—the path of completing school, investing in human capital, and putting marriage before motherhood.

This section reviews the most compelling evidence to date on whether teen childbearing itself causes adverse outcomes for teen mothers and their children.[11] We also review a related, important issue regarding "unintended" pregnancies. A large share of teen (and nonmarital) births are reported by the mother to be "unintended," which would suggest that providing teens with better access to sex education, contraception, and related policies could help them achieve their "intended" goal of not becoming pregnant. However, we believe that many births that are labeled as "unintended" actually reflect a degree of ambivalence on the part of the teen mother, in which case the policy prescription is less clear.

### The Effect of Teen Childbearing on Mothers and their Children

To what extent are the inferior outcomes of teen mothers driven by the event of having given birth as a teenager, as opposed to other factors, such as growing up in disadvantaged circumstances? A comparison of the outcomes of women who did and who did not give birth as teens is inherently biased by selection effects: teenage girls who "select" into becoming pregnant and subsequently giving birth (as opposed to choosing abortion) are different in terms of their background characteristics and potential future outcomes than teenage girls who delay childbearing.

We know that girls who grow up in poverty are more likely to become pregnant and to give birth as teenagers. Tabulating data from the 2003 Panel Study of Income Dynamics, we report that in a sample of women age 20 to 35, 24 percent give birth before age 20; but among the subsample of those women who were born into poverty, 49 percent give birth before age 20 (Kearney and Levine 2010).

A number of authors have tried to isolate the causal effect of teenage childbearing on subsequent outcomes, holding constant family background characteristics. To isolate the effect of teen childbearing, Geronimus and Korenman (1992) employ a "within-family" estimation approach that compares differences in subsequent socioeconomic status of sisters who experienced their first births at different ages. They analyze samples from three datasets: the National Longitudinal

[11] A separate question is the extent to which teen childbearing causes adverse outcomes for society more broadly. This is a topic that has been addressed by others (see Hoffman and Maynard 2008, for a recent review), although this literature has struggled with the issue of credibly identifying causal effects.

Survey Young Women's Sample (NLSYW), the Panel Study of Income Dynamics (PSID), and the 1979 National Longitudinal Survey of Youth (NLSY79). They find that cross-sectional comparisons that do not control for detailed family background greatly overstate the costs of teen childbearing. In fact, once background characteristics are controlled for, the differences are quite modest. Furthermore, even these modest differences likely overstate the costs of teen childbearing, since the sister who gives birth as a teen is likely to be "negatively" selected compared to her sister who does not.

In addition to differences in observed and unobserved family background characteristics, girls who are more committed to achieving higher levels of educational attainment and economic success may be more committed to preventing a pregnancy from occurring during their teenage years. Such girls may also be more inclined to choose abortion if they do get pregnant. From a research perspective, we ideally want to observe a sample of women who have the same potential outcomes and the same inclination to get pregnant and give birth, but by random chance, some do and some do not become teen mothers. A number of papers attempt to exploit quasi-experimental variation in who becomes a teen mother to isolate the causal consequences of teen childbearing.

Hotz, Mullen, and Sanders (1997) and Hotz, McElroy, and Sanders (2005) exploit the fact that some women who become pregnant as teenagers experience a miscarriage and thus do not have a birth. Their strategy essentially identifies the effect of delaying childbearing for women who become pregnant as teenagers. Using data from the 1979 National Longitudinal Survey of Youth (NLSY79) on women who were aged 13 to 17 between 1971 and 1982, the authors begin by replicating previous findings of a correlation between teen childbearing and later life outcomes.[12] But when the authors of these studies employ their miscarriages "experiment," and thereby avoid confounding selection effects, none of the differences are statistically significant, and some are even reversed in sign. Hoffman (2008) reexamines this data and finds that the estimated impacts of a teen birth are more negative for teen mothers who had births in the early 1980s relative to 1970s.[13] Ribar (1994) also employs an instrumental variables framework using the NLSY79 data. He uses age at menarche in an instrumental variables framework, noting that earlier age at menarche leads to more years at risk of becoming pregnant. The negative consequences of teen birth for high school completion rate also seem to disappear with this instrumental variable approach.

[12] An interesting statistic they tabulate in their data is that among women who become pregnant before age 18, those who choose to end their pregnancy in abortion on average have family incomes that are 40 percent higher than those who give birth. This supports the claim that among pregnant teens, there are important selection effects driving the decision to become a teen mother.

[13] Ashcraft and Lang (2010) build on the Hotz et al. study using data from the 1995 wave of the NSFG, which they argue is better suited for measuring pregnancy outcomes than the NLSY79. They additionally adjust for the fact that some girls will abort their pregnancy before a miscarriage occurs, leading to a selected sample. Their empirical estimates suggest that the effects of teen childbearing are more negative than suggested by Hotz, McElroy, and Sanders (2005), but that the adverse consequences on socioeconomic outcomes are quite small in magnitude.

David Levine and Gary Painter (2003) use a within-school propensity-score matching estimator—using quantitative methods to "match" individuals with an equal propensity to give birth as a teen and then comparing outcomes for those matched pairs who differ in teen childbearing outcomes—to identify causal effects on educational attainment. They use data from the 1988 National Educational Longitudinal Survey. The authors view their findings as suggesting (p. 898) that a "substantial portion of the relation between teen childbearing and high school completion is due to preexisting disadvantages of the young women, not due to the childbirth itself . . . Thus, half or more of the young mother's disadvantages would not have been eliminated by the young women waiting until their twenties to have children."

The evidence regarding the outcomes for the children of teen mothers similarly finds that observed differences reflect selection on the part of who becomes a teen mother, not the negative consequences of teen childbearing itself. Geronimus, Korenman, and Hillemeier (1994) employ a "within family" estimation strategy that compares outcomes for first cousins born to sisters of different ages. This analysis finds that children of teen mothers appear to score no worse on measures of development than their first cousins whose mothers had first births after their teen years. A more recent paper by Mullin (2005) employs the empirical approach of Hotz, Mullin, and Sanders (1997)—a bounded instrumental variables treatment framework relying on the occurrence of "miscarriages"—to assess the effect of teen childbearing on children. His analysis finds that delaying childbearing for nonblack teen mothers actually leads to *inferior* outcomes for the children. Building on the literature, Sepulveda (2010) extends the Ashcraft and Lang (2010) approach to the case of children. He estimates the causal effect of teen childbearing on children's outcomes to be a tightly bounded zero. Our reading of the most rigorous empirical studies to date is that the data reject the hypothesis that the children of teenage mothers would have experienced better outcomes had those same mothers delayed pregnancy until after age 19.

Taken as a whole, previous research has had considerable difficulty finding much evidence in support of the claim that teen childbearing has a causal impact on mothers and their children. Instead, at least a substantial majority of the observed correlation between teen childbearing and inferior outcomes is the result of underlying differences between those who give birth as a teen and those who do not.

### Intendedness

In 1994 and 2001, respectively, 77 percent and 82 percent of teen pregnancies were reported as "unintended." Indeed, among all women of childbearing age, half of all pregnancies are reported to be unintended (Finer and Henshaw 2006). Unintended pregnancies are related to a host of inferior outcomes for mothers and children, including higher rates of low birth weight and infant mortality, depression, and domestic violence, among others (Brown and Eisenberg 1995).

However, interpreting these statistics is difficult (Trussell, Vaughan, and Stanford 1999; Santelli et al. 2003). The common sources for pregnancy "intention" are the National Survey of Family Growth (NSFG) and the Pregnancy Risk Assessment

Monitoring System (PRAMS). A first difficulty is that the questions about intention are asked retrospectively: the NSFG asks the question after the child is born and the PRAMS asks the question when the woman is pregnant. For women in their 20s, a common belief in the research community is that too few births get reported as "unintended," because after women have become pregnant and given birth they view the event as a happy one. In fact, Joyce, Kaestner, and Korenman (2002) report that it is not uncommon for reported pregnancy intentions to change between the time of the pregnancy to a period after the child was born, with more women changing towards a more positive assessment than the reverse. When one focuses on teen fertility, we would argue that it is not generally socially acceptable to report that you "wanted" to get pregnant as a teenager, either at the time of conception, after pregnancy, or after the child is born. So survey rates of "unintendedness" would be biased upward.

Reports of pregnancy intendedness also contain inconsistencies that are hard to reconcile. Trussell, Vaughan, and Stanford (1999), for instance, report that one-third of women who report that their pregnancies began with a contraceptive failure also reported that the pregnancy was intended. Moreover, 41 percent of those who reported a contraceptive failure that led to an unintended pregnancy also report that they were happy about their pregnancy.

A plausible interpretation here is that "intendedness" is not a dichotomous variable, but instead reflects a continuum. A woman may not have planned in advance to become pregnant, but she may have been ambivalent towards whether a pregnancy might occur.

We use data from the Behavioral Risk Factor Surveillance System to tabulate why some sexually active women report not using birth control. We are interested in gaining some insight into whether unintended pregnancies are more often associated with *unwantedness* as opposed to ambivalence. In addition, we hope to gain some insight into how often limited information or limited access to contraception is to blame for a would-be unintended pregnancy. The purpose of these data is to track health outcomes and risky behaviors for adults (age 18 and over). We identify a sample of 230 women age 18 and 19 in the 2002 and 2004 surveys who are sexually active, unmarried, and not using birth control. Only 15 percent report that they were not expecting to have sex, and 11 percent report that they did not think that they could become pregnant. Remarkably, only 2 percent report that they could not afford birth control. In fact, 20 percent of them report that they either want to get pregnant or do not care if they get pregnant. The most common answer by far—45 percent—that women gave for not using birth control is "other reason." Although these findings pertain specifically to older teens, our interpretation of these data at least for this age group is that the non-use of contraception among these sexually active women reflects to a large degree a general sense of ambivalence toward becoming pregnant.

Hohmann-Marriott (2011) investigates the concept of ambivalence by comparing a woman's report of what she labels attitudes, behavior, and emotion using data from the Early Childhood Longitudinal Study—Birth Cohort (ECLS-B). "Attitudes" reflect standard wantedness and timing questions. "Behavior" addresses contraceptive use and reasons for non-use. "Emotions" address happiness towards the birth. She defines ambivalence towards a birth based on contradictory answers to the categories of questions. For instance, a woman whose pregnancy was categorized as unintended, but who did not use contraception prior to pregnancy is defined to have been ambivalent towards the pregnancy. Overall, Hohmann-Marriott finds that 39 percent of births reflect ambivalence on the part of the mother, as compared to much lower rates of 13 percent being truly unintended. However, these data only include women who gave birth, and who are more inclined towards positive intentions than a sample of all pregnancies because many unintended pregnancies end in abortion.

Our reading of the evidence is that ambivalence towards a pregnancy is reasonably common. Ambivalence toward pregnancy has very different policy implications than unwantedness.[14]

## Discussion and Policy Implications

We believe that the high rate of teen childbearing in the United States matters because it is a marker of a social problem, rather than the underlying social problem itself. If a teenager has a baby because her life chances seem so limited that her life will not be any better if she delays childbearing, then teen childbearing is unlikely to be causing much of a detrimental effect. Our review of the evidence is consistent with this position.

This conclusion has important implications for public policies. We do not believe that policies targeted directly at teen pregnancy prevention—sex education, improved access to contraception, abstinence counseling, and the like—are likely to improve outcomes much for disadvantaged young women. Instead, we believe that with improved economic opportunities, reduced poverty, and improved prospects for other adult outcomes, teen pregnancy would also decline. Levine and Zimmerman (2010), for instance, conclude that early childhood education programs and improved access to financial aid to attend college are two such types of interventions that have been proven effective in enhancing those opportunities in a cost-effective manner. A more ambitious prescription would be to attack the inequality and lack of mobility itself, although it is difficult to know just how to do that, and it might involve very large costs to do so.

We have focused here on the determinants and consequences of teen motherhood in the United States. One thing that we have not done is explain the dramatic decline in teen childbearing in the United States over the past 20 years. Although we believe that inequality and lack of opportunity explains a substantial share of the geographic variation in teen childbearing, it is not a candidate explanation for the downward trend in the United States over the past two decades, primarily because the 50/10 ratio that we rely on as a measure of inequality has not changed much during this period (although our results are insensitive to the specific measure

[14] For an alternative interpretation, see Kaye, Suellentrop, and Sloup (2009).

used).[15] In our analysis of recent trends that we described briefly earlier (Kearney and Levine 2012), we conclude that most targeted policies had no effect on teen birth rates. The two policy changes that do seem to matter some, expanded family planning services through Medicaid and reduced welfare benefits, can only combine to explain 12 percent of the decline in teen childbearing between 1991 and 2008. Demographic changes—in particular, the increasing Hispanic share of the teenage population—also cannot explain the recent decline, and in fact, would seem to have worked to increase teen childbearing rates. Additional work to understand the causes of the decline in U.S. teen childbearing should be a priority for future research.

Another issue related to teen childbearing is the consequences of nonmarital childbearing at any age. Single mothers have high poverty rates as well and the vast majority (87 percent in 2008; Martin et al. 2010) of teen births are nonmarital births. According to 2006 Census figures, 5.7 percent of people living in married couple families live below the federal poverty threshold as compared to 30.5 percent of people living in female-headed households. Broader discussions of nonmarital fertility, however, are complicated by the disparate types of women who encounter this outcome. Even among women beyond their teen years, 35.2 percent of births are to unmarried women. In fact, women in their teens represent 10 percent of all women giving birth; the number of nonmarital births to nonteens is three times larger than the number of teen nonmarital births. Clearly, these older women face different issues. It is a separate, complex issue to determine how much better off women and children would be if policies could successfully increase rates of two-parent biological families among economically disadvantaged populations.

[15] Other indicators of inequality like the 90/50 ratio (the ratio of the income at the 90th percentile to that at the median) have risen; but that would predict an upward trend in teen childbearing.

## References

**Akerlof, George A., Janet L. Yellen, and Michael L Katz.** 1996. "An Analysis of Out-of-Wedlock Childbearing in the United States." *Quarterly Journal of Economics* 111(2): 277–317.

**An, Chong-Bum, Robert Haveman, and Barbara Wolfe.** 1993. "Teen Out-of-Wedlock Births and Welfare Receipt: The Role of Childhood Events and Economic Circumstances." *Review of Economics and Statistics* 75(2): 195–208.

**Ashcraft, Adam, and Kevin Lang.** 2010. "The Consequences of Teenage Childbearing: Consistent Estimates when Abortion Makes Miscarriage Nonrandom." Unpublished manuscript.

**Becker, Gary S., and H. Gregg Lewis.** 1973. "On the Interaction between the Quantity and Quality of Children." *Journal of Political Economy* 81(2): S279–S288.

**Boonstra, Heather.** 2002. "Teen Pregnancy: Trends and Lessons Learned." *Guttmacher Report on Public Policy* 5(1): 7–10.

**Brooks-Gunn, Jeanne, and Frank F. Furstenberg, Jr.** 1989. "Adolescent Sexual Behavior." *American Psychologist* 44(2): 249–57.

**Brooks-Gunn, Jeanne, and Roberta Paikoff.** 1997. "Sexuality and Developmental Transitions during Adolescence." In *Health Risks and Developmental Transitions during Adolescence,* edited by John Schulenberg, Jennifer L. Maggs, and

Klaus Hurrelmann, 190–245. Cambridge, UK: Cambridge University Press.

**Brown, Sarah, and Leon Eisenberg, eds.** 1995. *The Best Intentions: Unintended Pregnancy and the Well-Being of Children and Families.* Washington, DC: Institute of Medicine.

**Clark, Kenneth B.** 1965. *Dark Ghetto: Dilemmas of Social Power.* New York: Harper & Row.

**Darroch, Jacqueline E., Susheela Singh, Jennifer J. Frost, and the Study Team.** 2001. "Differences in Teenage Pregnancy Rates among Five Developed Countries: The Roles of Sexual Activity and Contraceptive Use." *Family Planning Perspectives* 33(6): 244–50, 281.

**Duncan, Greg J., and Saul D. Hoffman.** 1990."Welfare Benefits, Economic Opportunities, and Out-of-Wedlock Births among Black Teenage Girls." *Demography* 27(4): 519–36.

**Duncan, Greg J, W. Jean Yeung, Jeanne Brooks-Gunn, and Judith R. Smith.** 1998. "How Much Does Childhood Poverty Affect the Life Chances of Children?" *American Sociological Review* 63(3): 406–23.

**Edin, Kathryn, and Maria Kefalas.** 2005. *Promises I Can Keep: Why Poor Women Put Motherhood before Marriage.* Berkeley, CA: University of California Press.

**Finer, Lawrence B., and Stanley K. Henshaw.** 2006. "Disparities in Rates of Unintended Pregnancy in the United States." *Perspectives on Sexual and Reproductive Health* 38(2): 90–96.

**Geronimus, Arline T., and Sanders Korenman.** 1992. "The Socioeconomic Consequences of Teen Childbearing Reconsidered." *Quarterly Journal of Economics* 107(4): 1187–1214.

**Geronimus, Arline T., and Sanders Korenman.** 1993. "The Socioeconomic Costs of Teen Childbearing: Evidence and Interpretation." *Demography* 30(2): 281–90.

**Geronimus, Arline T., Sanders Korenman, and Marianne M. Hillemeier.** 1994. "Does Young Maternal Age Affect Child Development? Evidence from Cousin Comparisons in the United States." *Population and Development Review* 20(3): 585–609.

**Godeau, Emannuelle, Saoirse Nic Gabhainn, Céline Vignes, Jim Ross, Will Boyce, and Joanna Todd.** 2008. "Contraceptive Use by 15 Year Old Students at the Last Sexual Intercourse." *Archives of Pediatric & Adolescent Medicine* 162(1): 66–73.

**Grogger, Jeffrey, and Lynn A. Karoly.** 2005. *Welfare Reform: Effects of a Decade of Change.* Cambridge, MA: Harvard University Press.

**Guttmacher Institute.** 2010. *U.S. Teenage Pregnancies, Births and Abortions: National and State Trends and Trends by Race and Ethnicity.* New York: Guttmacher Institute.

**Hamilton, Brady E., Joyce A. Martin, and Stephanie J. Ventura.** 2010. "Births: Preliminary Data for 2009." *National Vital Statistics Report,* December 21, 59(3).

**Hardy, Janet B., and Laurie Schwab Zabin.** 1991. *Adolescent Pregnancy in an Urban Environment.* Washington, DC: Urban Institute Press.

**Hoffman, Saul D.** 2008. "Updating the Teen Miscarriage Experiment: Are the Effects of a Teen Birth Becoming More Negative?" Working Papers 08-08, University of Delaware, Department of Economics.

**Hoffman, Saul D., and Rebecca A. Maynard, eds.** 2008. *Kids Having Kids: Economic Costs and Social Consequences of Teen Pregnancy.* Washington, DC: Urban Institute.

**Hohmann-Marriott, Bryndl.** 2011. "Ambivalent Intentions for Pregnancy: Measurement, Partner Effects, and Future Intentions." Presented at the 2011 meeting of the Population Association of America.

**Hotz, V. Joseph, Susan Williams McElroy, and Seth Sanders.** 2005. "Teenage Childbearing and Its Life Cycle Consequences: Exploiting a Natural Experiment." *Journal of Human Resources* 40(3): 683–715.

**Hotz, V. Joseph, Charles Mullin, and Seth Sanders.** 1997. "Bounding Causal Effect Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing." *Review of Economic Studies* 64(4): 575–603.

**Joyce, Ted, Robert Kaestner, and Sanders Korenman.** 2002. "On the Validity of Retrospective Assessments of Pregnancy Intention." *Demography* 39(1): 199–213.

**Kaye, Kelleen, Katherin Suellentrop, and Corinna Sloup.** 2009. *The Fog Zone: How Misperceptions, Magical Thinking, and Ambivalence Put Young Adults at Risk for Unplanned Pregnancy.* Washington, DC: National Campaign to Prevent Teen and Unplanned Pregnancy.

**Kearney, Melissa S.** 2004. "Is There an Effect of Incremental Welfare Benefits on Fertility Behavior? A Look at the Family Cap." *Journal of Human Resources* 39(2): 295–325.

**Kearney, Melissa S.** 2010. "Teen Pregnancy Prevention." Chap. 8 in *Targeting Investments in Children: Fighting Poverty when Resources are Limited,* edited by Phillip B. Levine and David J. Zimmerman. *University of Chicago Press.*

**Kearney, Melissa S., and Phillip Levine.** 2009. "Subsidized Contraception, Fertility, and Sexual Behavior." *Review of Economics and Statistics* 91(1): 137–51.

**Kearney, Melissa S., and Phillip B. Levine.** 2010. "Socioeconomic Disadvantage and Early Childbearing." Chap. 6 in *The Problems of Disadvantaged Youth: An Economic Perspective,* edited by Jonathan Gruber. University of Chicago Press.

**Kearney, Melissa S., and Phillip B. Levine.** 2011.

"Income Inequality and Early, Non-marital Childbearing: An Economic Exploration of the 'Culture of Despair.'" NBER Working Paper 17157.

**Kearney, Melissa S. and Phillip B. Levine.** 2012. "Explaining Recent Trends in the U.S. Teen Birth Rate." NBER Working Paper 17964.

**Levine, David I., and Gary Painter.** 2003. "The Schooling Costs of Teenage Out-of-Wedlock Childbearing: Analysis with a Within-School Propensity-Score Matching Estimator." *Review of Economics and Statistics* 85(4): 884–900.

**Levine, Phillip B.** 2002. "The Impact of Social Policy and Economic Activity throughout the Fertility Decision Tree." NBER Working Paper 9021.

**Levine, Phillip B.** 2004. *Sex and Consequences: Abortion, Public Policy, and the Economics of Fertility*. Princeton, NJ: Princeton University Press.

**Levine, Phillip B., and David J. Zimmerman, eds.** 2010. *Targeting Investments in Children: Fighting Poverty when Resources are Limited*. University of Chicago Press.

**Lewis, Oscar.** 1969. "The Culture of Poverty." In *On Understanding Poverty: Perspectives from the Social Sciences,* edited by Daniel P. Moynihan, 187–200. New York: Basic Books.

**Lundberg, Shelly, and Robert D. Plotnick.** 1995. "Adolescent Premarital Childbearing: Do Economic Incentives Matter?" *Journal of Labor Economics* 13(2): 177–200.

**Martin, Joyce A., Brady E. Hamilton, Paul D. Sutton, Stephanie J. Ventura, T. J. Mathews, and Michelle Osterman.** 2010. *Births: Final Data for 2008*. National Vital Statistics Reports vol. 59, no. 1.

**Martin, Joyce A., Brady E. Hamilton, Stephanie J. Ventura, Michelle J. K. Osterman, Sharon Kirmeyer, T. J. Mathews, and Elizabeth C. Wilson.** 2011. *Births: Final Data for 2009*. National Vital Statistics Reports, vol. 60, no. 1.

**Moffitt, Robert A.** 1998. "The Effect of Welfare on Marriage and Fertility: What Do We Know and What Do We Need to Know?" Chap. 4 in *Welfare, the Family, and Reproductive Behavior,* edited by R. Moffitt. Washington: National Research Council, National Academy of Sciences Press.

**Moffitt, Robert A.** 2003. "The Negative Income Tax and the Evolution of U.S. Welfare Policy." *Journal of Economic Perspectives* 17(3): 119–40.

**Mullin, Charles.** 2005. "Bounding Treatment Effects with Contaminated and Censored Data: Assessing the Impact of Early Childbearing on Children." *Advances in Economic Analysis & Policy* 5(1): Article 8.

**O'Donoghue, Ted, and Matthew Rabin.** 1999. "Doing It Now or Later." *American Economic Review* 89(1): 103–124.

**Ribar, David C.** 1994. "Teenage Fertility and High School Completion." *Review of Economics and Statistics* 76(3): 413–24.

**Santelli, John S., Laura Duberstein Lindberg, Lawrence B. Finer, and Susheela Singh.** 2007. "Explaining Recent Declines in Adolescent Pregnancy in the United States: The Contribution of Abstinence and Improved Contraceptive Use." *American Journal of Public Health* 97(1): 150–56.

**Santelli, John, Roger Rochat, Kendra Hatfield-Timajchy, Brenda Colley Gilbert, Kathryn Curtis, Rebecca Cabral, Jennifer S. Hirsch, and Laura Schieve.** 2003. "The Measurement and Meaning of Unintended Pregnancy." *Perspectives on Sexual and Reproductive Health* 35(2): 94–101.

**Santelli, John, Theo Sandfort, and Mark Orr.** 2008. "Transnational Comparisons of Adolescent Contraceptive Use." *Archives of Pediatric and Adolescent Medicine* 162(1): 92–94.

**Sedgh, Gilda, Stanley K. Henshaw, Susheela Singh, Akinrinola Bankole, and Joanna Drescher.** 2007. "Legal Abortion Worldwide: Incidence and Recent Trends." *International Family Planning Perspectives* 33(3): 106–116.

**Sepulveda, Carlos Eduardo.** 2010. "Teenage Childbearing: Effects on the Child." Unpublished manuscript.

**Trussell, James, Barbara Vaughan, and Joseph Stanford.** 1999. "Are All Contraceptive Failures Unintended Pregnancies? Evidence from the 1995 National Survey of Family Growth." *Family Planning Perspectives* 31(5): 246–47, 260.

**United Nations, Department of Economic and Social Affairs.** 2011. *United Nations Demographic Yearbook, 2009–2010.* New York: United Nations. http://unstats.un.org/unsd/demographic/products/dyb/dybsets/2009-2010.pdf.

**United Nations Economic Commission for Europe (UNECE).** UNECE Statistical Database. Available at http://w3.unece.org/pxweb/. Accessed June 1, 2011.

**U.S. Department of Labor, Office of Policy Planning and Research.** 1965. *The Negro Family: The Case for National Action*. Washington, DC: United States Department of Labor.

**Wilson, William Julius.** 1987. *The Truly Disadvantaged: The Inner City, the Underclass and Public Policy*. Chicago: University of Chicago Press.

# Why was the Arab World Poised for Revolution? Schooling, Economic Opportunities, and the Arab Spring[†]

Filipe R. Campante and Davin Chor

In December 2010, the self-immolation of a Tunisian fruit vendor sparked what has come to be termed the "Arab Spring." What first appeared as an isolated act of protest against local authorities quickly gained broader significance, as it was followed by a series of demonstrations that has shaken the grip of autocratic regimes across the Arab world. A year later, three longstanding dictators—Zine El Abidine Ben Ali of Tunisia, Hosni Mubarak of Egypt, and Muammar el-Qaddafi of Libya—have been ousted, after varying degrees of violence. Syria, Yemen, and Bahrain have all witnessed extensive turmoil, raising serious questions about the legitimacy and survival of their rulers. Elsewhere, the political leaders of Morocco, Algeria, and Jordan have also been pressured into enacting reforms to try to assuage public demands.

What underlying long-term conditions set the stage for the Arab Spring? Clearly, the possibilities are manifold. For example, *The Economist*'s (2011a) "shoe-thrower's index" of the vulnerability of Arab regimes lists many possible determinants of instability: years in power of the incumbent; size of the youth population cohort; democracy, corruption, press freedom, and GDP per capita (as a summary measure of economic performance). In this paper, we do not seek to rule out a contributing role for these and other structural forces that have been proposed; in places, we will acknowledge and comment on some of these other forces. However, we are convinced that education and its connection with the economic environment

■ *Filipe R. Campante is Assistant Professor of Public Policy, Harvard Kennedy School, Cambridge, Massachusetts. Davin Chor is Assistant Professor of Economics, Singapore Management University, Singapore. Their e-mail addresses are ⟨Filipe_Campante@harvard.edu⟩ and ⟨davinchor@smu.edu.sg⟩.*

[†] To access the Appendix, visit
http://dx.doi.org/10.1257/jep.26.2.167.

doi=10.1257/jep.26.2.167

deserve prominent consideration in any inquiry into the Arab Spring and other similar episodes of political change.

The argument that education positively affects the political development and democratization of nations has a long and venerable vintage (Dewey 1916; Lipset 1959, 1960). At a more micro level, a vast body of evidence confirms that individuals with a higher educational attainment consistently exhibit a greater propensity to participate in the full spectrum of political activities, from milder forms of engagement such as voting or discussing politics to more public forms of mobilization such as demonstrations.[1] This positive relationship holds true in virtually any survey dataset that asks about political engagement, even after controlling extensively for other individual traits such as age, gender, and income. Indeed, Putnam (1995a, p. 68) claims with some justification that education is "the best individual-level predictor of political participation."

But education and economic opportunity can be mismatched. Huntington (1968, p. 48) discusses how higher education in many modernizing countries often failed to provide skills that were relevant to the countries' needs, churning out graduates faster than jobs could be created, and thus leading instead to alienation and instability: "The higher the level of education of the unemployed, . . . the more extreme the destabilizing behavior which results." Along similar lines, Davies (1962) famously posited that rising expectations associated with expansions in education could, when left unmet, spill over into political violence, and several observers have highlighted the potential for this combustible mix of conditions in the Arab world (Courbage and Todd 2007; Noland and Pack 2007).

Indeed, the Arab world has experienced a remarkable expansion of education in recent decades. Table 1 shows a ranking of countries around the world by the observed increase in average years of schooling in the population aged 15 and over between 1980 and 2010. We use the latest version of the Barro and Lee (2010) dataset, updated in September 2011, which provides quantitative information on educational achievement across 146 countries spanning all inhabited continents. Eight of the top 20 countries as ranked by schooling increases in the general population are members of the Arab League, including notably Tunisia, Egypt, and Libya. A ninth country, Iran, is from the broader Middle East region and has been the site of a series of mass protests since mid-2009.[2]

[1] This positive empirical relationship between political participation and schooling has been explored extensively by political scientists and economists, including: Verba and Nie (1987), Rosenstone and Hansen (1993), Putnam (1995b), Verba, Schlozman, and Brady (1995), Bénabou (2000), Schlozman (2002), Dee (2004), Freeman (2004), Milligan, Moretti, and Oreopoulos (2004), Hillygus (2005), Glaeser, Ponzetto, and Shleifer (2007), Sondheimer and Green (2010), and Campante and Chor (forthcoming). A smaller body of work finds contrary evidence in some instances on the relationship between schooling and voting specifically, but we argue later that this detracts little from our main thesis.

[2] It may appear odd that Germany ranks so highly on this top-20 list; this happens largely because the Barro–Lee data suggest relatively low levels of schooling attainment in Germany in 1980. On this, Barro and Lee (2001, p. 559–60) have noted the discrepancy between their estimates, based on data from UNESCO, and other estimates by the OECD which tend to ascribe higher education levels to Germany in past decades. They attribute these differences mostly to the fact that the UNESCO data classified

*Table 1*
**Increases in Schooling Attainment across the World**
*(top 20 countries, 1980–2010)*

| *Country* | *Years of schooling, 1980* | *Years of schooling, 2010* | *Increase in years of schooling* |
|---|---|---|---|
| 1. Botswana | 3.12 | 9.56 | 6.44 |
| 2. Germany | 5.61 | 11.82 | 6.21 |
| **3. Iran** | **3.34** | **8.59** | **5.25** |
| **4. Algeria** | **3.06** | **8.30** | **5.24** |
| **5. United Arab Emirates** | **3.88** | **9.12** | **5.23** |
| 6. Gabon | 3.33 | 8.35 | 5.02 |
| 7. Brazil | 2.77 | 7.54 | 4.77 |
| **8. Bahrain** | **4.92** | **9.59** | **4.67** |
| **9. Jordan** | **4.58** | **9.23** | **4.65** |
| **10. Libya** | **3.26** | **7.85** | **4.59** |
| 11. France | 5.96 | 10.53 | 4.58 |
| 12. Malaysia | 5.69 | 10.14 | 4.46 |
| 13. Bolivia | 5.47 | 9.91 | 4.44 |
| **14. Egypt** | **2.65** | **7.08** | **4.43** |
| 15. El Salvador | 3.58 | 7.97 | 4.39 |
| 16. Mexico | 4.89 | 9.11 | 4.22 |
| 17. Spain | 6.17 | 10.38 | 4.22 |
| **18. Saudi Arabia** | **4.38** | **8.48** | **4.10** |
| **19. Tunisia** | **3.25** | **7.32** | **4.07** |
| 20. Latvia | 6.69 | 10.60 | 3.91 |

*Source:* Calculated from the Barro–Lee dataset (2010, version 1.2), based on the average total years of schooling in the population aged 15 and above.
*Note:* Arab League countries and Iran are in bold.

We develop our argument by first presenting evidence that this expansion of education in the Arab world was indeed matched with poor labor market prospects, and particularly so in the countries that have been at the heart of the protest wave. We show that this set of conditions is in turn associated with an increased propensity at the individual level towards engaging in political activities of a protest nature, especially among those who have received more education. We then argue that these conditions are useful predictors of political instability and incumbent change at the country level, although we caution that more work is needed to establish the causal nature of this relationship. Finally, we conclude by discussing several potential research directions, as well as the implications of our framework for some ongoing developments and debates. For convenience, we will refer to the Arab world and Middle East interchangeably. Our discussion will include Iran, which is not an Arab League member but largely follows the patterns discussed here. It will however exclude Israel and Turkey, which are geographically in the Middle East but face different economic and political challenges.

vocational and nonacademic educational training, which are very important in Germany, as part of lower levels of schooling.

## Schooling and Economic Opportunities in the Arab World

In recent decades, the Arab region has been characterized by an expansion in schooling coupled with weak labor market conditions. This pattern is especially pronounced in those countries that saw significant upheaval during the first year of the Arab Spring uprisings.

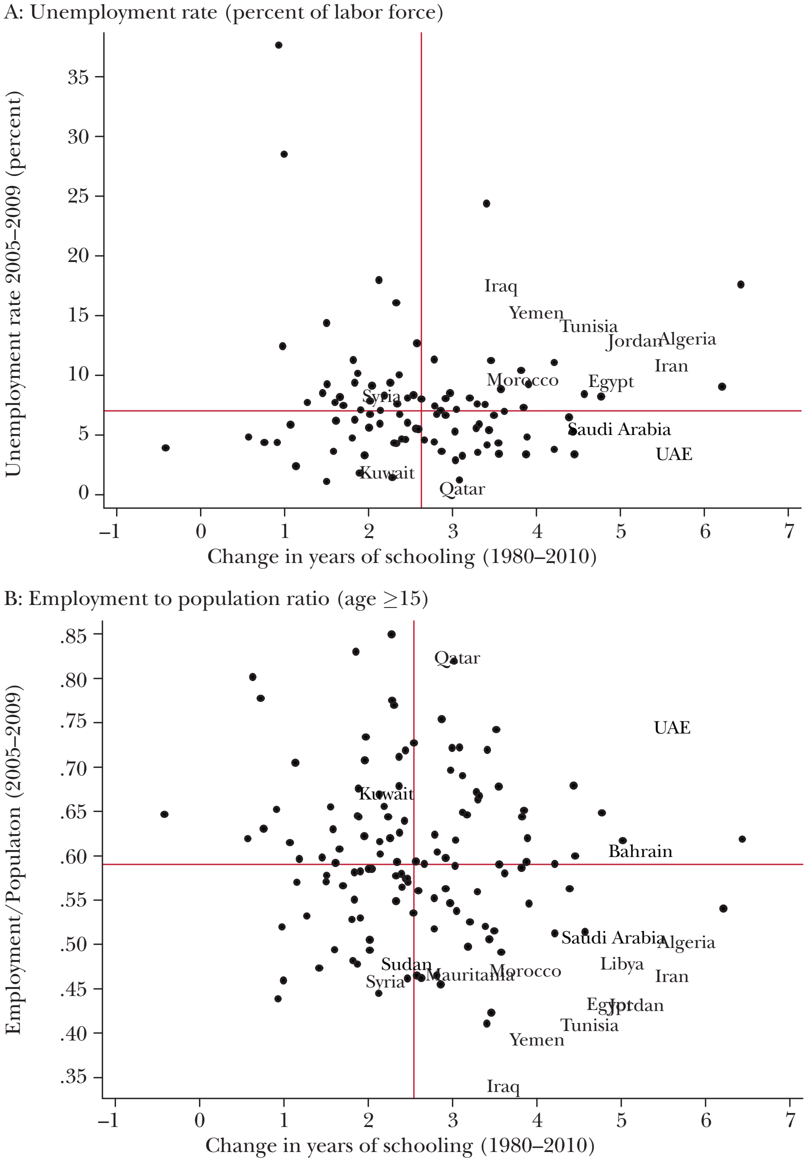
Figure 1 compares the Arab world vis-à-vis the rest of the world along the two respective dimensions of schooling gains and labor market conditions. For ease of visualization, only the Arab League and Iran are labeled by their country names in the scatterplot. In both panels, the horizontal axis plots the change in average years of schooling between 1980 and 2010 in the general population—this is the same variable that was reported in Table 1, calculated from the Barro and Lee (2010) dataset. The Arab countries are generally above the sample median (indicated by the vertical line) in terms of the gains in schooling attainment that they achieved. While many developing countries elsewhere also saw large improvements in schooling during this period, the Arab world had in many cases the advantage of oil revenues that supported the expansion of the education system (Daun and Arjmand 2002). These data admittedly only provide a rough gauge of educational achievement. For instance, they say nothing about the content of education, and a comparatively high proportion of students' time in Arab countries is devoted to religious studies that may be linked to lower labor market rewards (Daun and Arjmand 2002). Notwithstanding this caveat, the available data do suggest large gains in human capital in the Arab world.

No correlation between the two variables is obviously apparent in Figure 1—nor would we have expected one—but one message that clearly emerges is that, in the Middle East, the gains in schooling were often not accompanied by abundant labor market opportunities. Figure 1A plots the unemployment rate in the general population averaged over 2005–2009 on the vertical axis, as made available by the World Bank's World Development Indicators. These data are drawn from the International Labor Organization (ILO), which compiles labor market statistics from labor force or household surveys conducted by country authorities around the world. These are convenient as a summary indicator of labor market prospects for a broad sample of countries. However, it should be noted that while the ILO defines the unemployment rate as the share of the labor force that is without work, but is available for and seeking employment, the definitions of what constitutes the labor force and unemployed status can differ from country to country.

Figure 1A confirms that many Middle East countries have experienced unemployment rates above the sample median (horizontal line). Indeed, many of the countries that have seen large-scale protests in the first year of the Arab Spring are in the upper-right quadrant of this figure. Egypt, Tunisia, Yemen, and Iran stand out in this regard, although Morocco and Jordan have also seen significant though less violent protests. On the other hand, those Middle East countries with the lowest unemployment rates, namely Qatar, Kuwait, and the United Arab Emirates, have seen some rise in political expression (for press coverage, see *The Economist* 2011c), but have not seen mass protests.

*Figure 1*

**Where the Arab Countries Stand: Labor Market Conditions against Schooling Investments**


A: Unemployment rate (percent of labor force)

B: Employment to population ratio (age ≥15)

*Source:* Data on schooling from Barro and Lee (2010); other data from World Bank World Development Indicators.

*Notes:* Arab League countries and Iran are labeled by their country names, while all other countries are indicated by circular markers. The *x*-axis plots the change in years of schooling in the general population aged 15 and over, between 1980–2010. For the *y*-axis, Panel A plots the unemployment rate in the total labor force, while Panel B plots the employment-to-population ratio for the general population aged 15 and over. All labor market variables are averaged over 2005–2009. Lines indicating the median values of the *x*- and *y*-axis variables are included.

A very similar picture emerges when we consider an alternative measure of the state of the labor market in Figure 1B, namely the employment-to-population ratio for the population aged 15 and older, once again averaged over 2005–2009. It measures the share of the population employed in the market production of goods and services. While these data are also taken from the World Development Indicators and ultimately draw from the same ILO source as the unemployment data, they are potentially less subject to some idiosyncrasies that could distort the accuracy of unemployment rates as a description of labor market conditions. For instance, a country might have a relatively low unemployment rate because individuals drop out of the labor force in response to bleak job prospects. Most of the Middle East countries that have been at the center of violent protests exhibited large schooling gains and poor employment-to-population ratios, appearing in the lower-right quadrant of Figure 1B. In particular, this quadrant now includes Libya, for which the unemployment data were unavailable for Figure 1A. On the other hand, countries such as Qatar and the United Arab Emirates, which remained relatively peaceful amid the ongoing events in nearby countries, once again show up in the quadrant combining large gains in schooling and a relatively healthy labor market.[3]

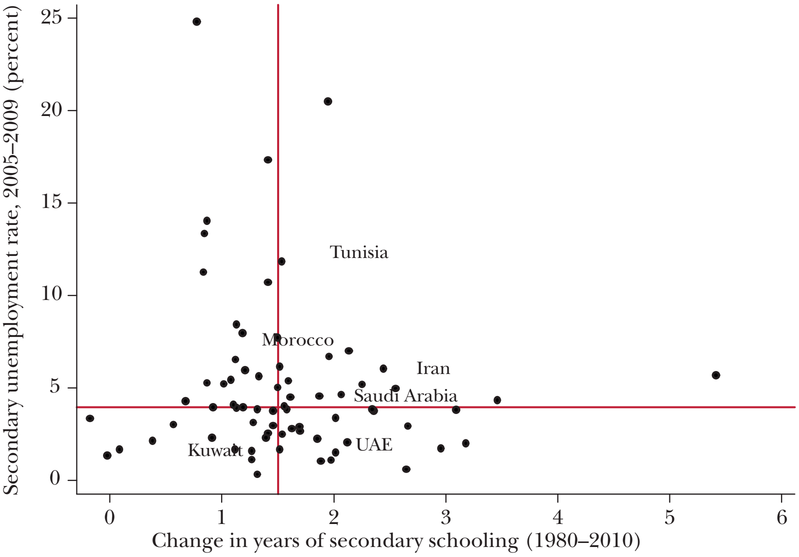
The labor market prospects faced by skilled or educated workers in the Arab world appear to be especially weak. Figure 2 presents one such illustration based on the secondary unemployment rate over 2005–2009, namely the percentage of workers with some secondary education who are unemployed.[4] This variable is analogously plotted against the change between 1980 and 2010 in the average years of secondary schooling among those aged 15 and over. A key caveat here is that unemployment data by education status is reported by fewer countries, so the scatterplot is much sparser; for example, Egypt and Algeria both drop out of the sample. The available data nevertheless provide a picture that is consistent with Figure 1: Protest-prone Tunisia, Iran, and Morocco all saw relatively large increases in secondary schooling yet had high secondary unemployment rates. Kuwait and the United Arab Emirates on the other hand appear to have been the most successful in the region in providing job opportunities for the secondary-educated while averting major unrest at the same time.

An alternative approach to assessing the labor market rewards for educated individuals would be to examine the prevailing skill premium in Arab countries. On

[3] A very similar figure emerges when using the employment-to-population ratio for the male workforce only (also from the World Development Indicators), indicating that our conclusions are not overly driven by the relatively low levels of female participation in labor markets in the Arab world. We also obtained a similar set of scatterplots when using real GDP per capita from the Penn World Tables as an alternative proxy for economic conditions on the vertical axis. Both of these figures are available in the online Appendix.

[4] The available variable from the ILO actually reports the percentage of the unemployed who have a secondary schooling attainment. To convert this to the percentage of those with some secondary education who are unemployed, we multiply this by the overall unemployment rate and divide by the percentage of those aged 15 and above with some secondary education (from the Barro–Lee dataset). Data availability issues prevent us from building the same figure for tertiary unemployment rates for a large enough number of countries.

*Figure 2*
**Where the Arab Countries Stand: Unemployment Rates against Secondary Schooling Investments**



*Notes:* Arab League countries and Iran are labeled by their country names, while all other countries are indicated by circular markers. The *x*-axis plots the change in years of secondary schooling in the general population aged 15 and over, between 1980–2010. The *y*-axis plots the unemployment rate among workers with some secondary education, averaged over 2005–2009 (calculation described in the main text). Lines indicating the median values of the *x*- and *y*-axis variables are included.

this count, the data from Clemens, Montenegro, and Pritchett (2009, table 2) on immigrant wages in the United States uncover some indirect evidence. For instance, they show that very skilled Egyptians with at least some college-level education who migrated to the United States earned more than 13 times as much as comparable individuals who remained in their home country. While less-skilled Egyptians would also earn more in the United States than their counterparts back home, the premium in their case turns out to be smaller, namely a factor of ten for those with no more than a secondary education. A similar pattern also applies to Yemen, where the corresponding wage factors are eleven and nine for college-educated and secondary-educated immigrants to the U.S. economy, respectively. This pattern is in fact unusual for the sample of countries in Clemens, Montenegro, and Pritchett (2009), as the relative premium for immigrants with some tertiary education is often on par or even lower than the premium for immigrants with only some secondary education (relative to counterparts in their countries of origin). This is true in particular for the other two Arab countries in their sample, Morocco and Jordan, which incidentally did not see the Arab Spring boil over to the same degree: College-educated immigrants from Jordan earned a premium that was a factor of

4.15, lower than the corresponding factor of 4.82 for those with some secondary education; in the case of Morocco, the immigrant-premium was essentially the same for both education categories (a factor of two). Egypt and Yemen thus seem to have had particularly unrewarding labor markets for the relatively skilled.

Still on this note, it is especially telling that the Tunisian street vendor, Mohamed Bouazizi, whose act of protest marked the start of the Arab Spring, was himself rumored to be a university graduate. Although this detail about his schooling was apocryphal (Fahim 2011), the fact that the rumor gained such traction is revealing of the strong current of job-related discontent amongst university graduates in Tunisia. Many observers have since drawn attention to the high unemployment rates that prevail among the increasing ranks of the educated in countries such as Tunisia and Egypt, pinning blame on the overbearing presence of an inefficient, heavily regulated state that crimps the development of independent enterprises and business activity (Ammous and Phelps 2011; Cassidy 2011; *The Economist* 2011b). For example, in the recent World Bank report on "Doing Business in the Arab World" (Doing Business 2011), the region as a whole ranked only ahead of South Asia and Sub-Saharan Africa on a composite index designed to capture how easy it is for small private firms to do business.[5]

Taken together, these different pieces of information build a narrative that suggests that the combination of rising levels of education and poor job prospects—particularly for the relatively skilled—was present in the Arab world, and particularly so in those countries that have witnessed the Arab Spring in its fullest bloom.

## The Links to Political Protest

Why would a large expansion in schooling and weak prospects for the workforce open the door to political instability? A very robust and widespread body of empirical evidence has shown that individuals with higher educational attainment are more likely to engage in all types of political acts—whether because education increases awareness of political issues, fosters the socialization needed for effective political activity, or generally increases so-called civic skills (for example, Brady, Verba, and Schlozman 1995; Glaeser, Ponzetto, and Shleifer 2007). Indeed, evidence from randomized and quasi-experimental settings (such as Sondheimer and Green 2010), as well as from instrumental variables approaches (such as Milligan, Moretti, and Oreopoulos 2004), suggest that the relationship from more education to greater political involvement is likely to be causal.

We believe that the attractiveness of the labor market returns for the skills acquired through education will influence the extent to which education would

[5] This average score for the Arab world masks a lot of diversity within the region: Saudi Arabia, Bahrain, and the United Arab Emirates were actually among the top 40 countries in the world on this index, with Qatar, Tunisia, and Oman not too far behind. The report also documents a number of positive reforms that have been undertaken by Arab countries in recent years.

also raise one's participation in political activities. In Campante and Chor (forthcoming), we flesh out a formal framework for understanding why this would be the case. Suppose that the skills and human capital acquired through education are useful both in production and political activities. In this setting, an economic environment in which human capital is more valuable in production will also be one where individuals are less likely to direct their human capital towards political participation, given the higher opportunity cost of the production income from labor markets that would be foregone. This tradeoff should be especially acute for those forms of political engagement that are effort-intensive, and so the economic environment should have a larger effect on the propensity to devote one's human capital to labor-intensive activities like public protests as compared to less labor-intensive acts like voting. In addition, weak economic conditions that affect all workers' incomes proportionately should lead to a relatively larger response in political participation from more-educated individuals, given the higher value of the foregone income that such individuals should in principle command. In fact, the importance of the opportunity cost of political activity has been used in a number of ways in the literature. For example, it has been emphasized in the study of regime transitions (Acemoglu and Robinson 2001, 2005; Brückner and Ciccone 2011), civil wars (Grossman 1991; Collier and Hoeffler 2004), and political violence in general (Besley and Persson 2011).

In the years leading up to the events of the Arab Spring, the expansion of schooling in the Arab world increased the pool of individuals who had completed primary and attained some secondary schooling (and beyond) but who had not seen that education rewarded in the labor market. The low opportunity cost of political participation would thus make such individuals more likely instead to channel their efforts towards political action, and political protest in particular. We can explore this prediction using response data from the World Values Survey, a comprehensive study on sociocultural and political attitudes conducted periodically around the world. We draw on the most recent complete wave of this survey, which was conducted from 2005–2007. Several questions in the survey pertain to one's propensity to engage in a number of different political activities, including "attending lawful demonstrations" (question E026). The response options to this question are "Would never do," "Might do," or "Have done," which we coded respectively as 0, 1, and 2 to provide a score of an individual's self-reported propensity towards this effort-intensive form of political participation. We took a simple average of this individual score for those respondents in each country who reported having at least some secondary education (usually numbering several hundred per country).[6] We focus specifically on these respondents, because our argument implies that country differences in the propensity to demonstrate should be even starker for individuals with higher levels of education. It is unlikely, after all, that primary education focused on basic literacy and numeracy would have the same effect on political involvement as

[6] This corresponds to individuals who reported at least a 3 on the World Values Survey 8-point scale of education status.

the teaching of critical thinking skills in secondary and tertiary schooling. We were able to compute this country score for only 40 countries, admittedly a relatively small number.

Figure 3 compares the average country scores within each of the quadrants of Figure 1A—that is, grouping countries according to whether they are above or below the world median in terms of schooling gains and the unemployment rate. Figure 3A shows that there is a higher average propensity to attend lawful demonstrations in countries that were above the median in terms of the increase in years of schooling. Most notably, this difference is entirely driven by the subgroup of countries that also saw relatively high unemployment rates. In short, greater gains in education at the country level appear to be associated with a stronger individual propensity towards protest activities, but much more so when they are combined with poor labor market conditions.

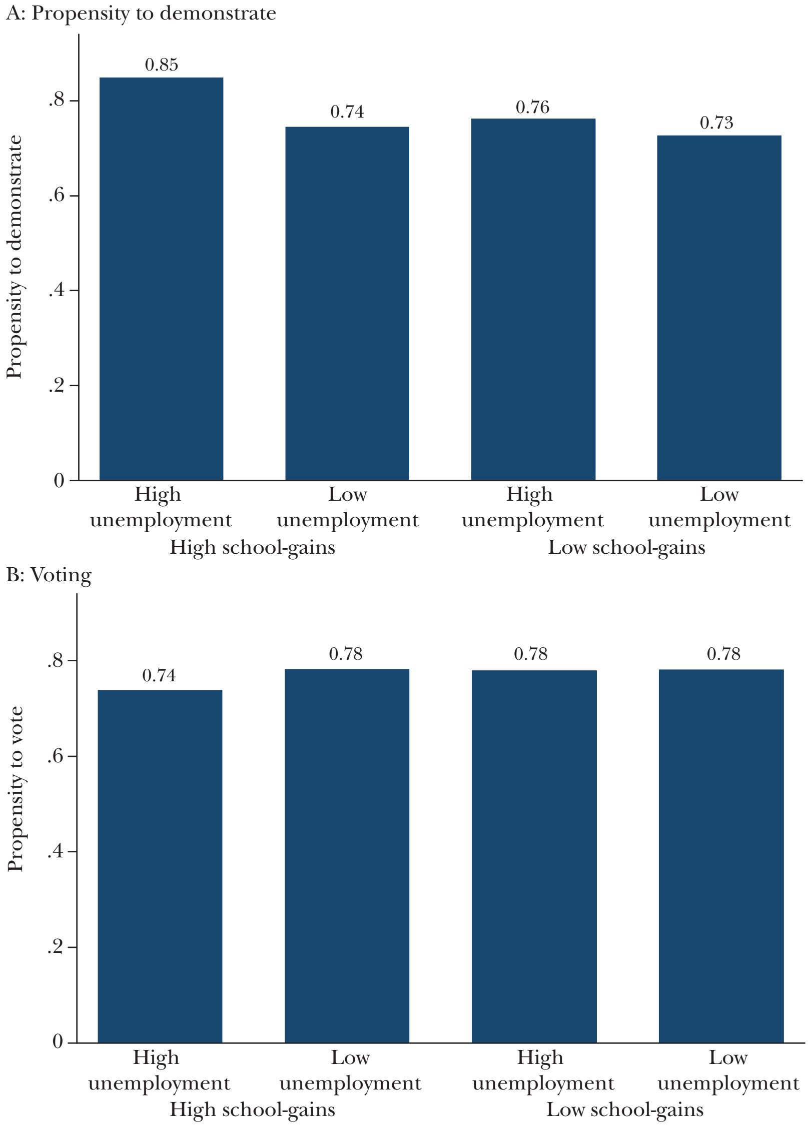
On the other hand, a less distinct pattern emerges when we consider a form of political activity that is less effort intensive, namely voting. Voting has been described as "the only political act requiring relatively little initiative" (Verba and Nie 1987, p.77) as well as being the least demanding in terms of civic skills (Brady, Verba, and Schlozman 1995). Figure 3B reproduces the above exercise using instead the response to a binary voting variable from the World Values Survey (question E257)—whether the individual voted in the most recent parliamentary election—to compute the country political participation score. The interaction between education and economic circumstances does not appear to operate as clearly in promoting this form of political activity, which is consistent with the lower opportunity cost of time and effort associated with voting.

Of course, the illustrative presentation of the data patterns here is only a starting point, but we have explored these patterns of political participation more formally using data from the World Values Survey and the Comparative Study of Electoral Systems in Campante and Chor (forthcoming). For example, in countries well endowed with resources that are likely to decrease the relative returns to human capital in production—for example, resources such as land that are associated with less skill-intensive activities—the positive correlation between education and participation at the individual level was stronger precisely in effort-intensive political activities such as attending demonstrations or occupying buildings. At a more micro level, we also found that individuals who worked in unskilled or manual occupations showed a greater propensity than those in skilled occupations to devote some of any incremental human capital towards political activities.[7] In Campante and Chor (2011), we have undertaken a more rigorous empirical exploration of these ideas using the full individual-level data in the World Values Survey. We found that while more-educated individuals are more likely to engage

[7] In a similar vein, Charles and Stephens (2011) provide evidence that positive labor market shocks tend to reduce voter turnout in local U.S. elections. They attribute this finding to the increased opportunity cost of using one's time to pay attention to and gather information on political developments related to municipal issues, which tend to receive less media coverage than national issues.

*Figure 3*

**Individual Engagement in Political Activities across Different Sets of Countries (Restricted to Individuals with at Least Some Secondary Education)**



*Source:* Authors based on data from the World Values Survey.

*Notes:* The propensity to demonstrate scores in Panel A are calculated from Wave 5 of the World Values Survey, averaged over individuals in each country to obtain a country score. The range of values of this score is from 0 to 2. The propensity to vote in Panel B is calculated from a binary variable on participation in voting from Wave 5 of the World Values Survey, averaged over individuals in each country to obtain a country score. The range of values of this variable is from 0 to 1. All country scores are calculated using only individuals with some secondary education (at least 3 on the Survey's 8-point scale of education status). The columns in each panel report the mean country scores as grouped by the four quadrants in Figure 1A, namely: above/below the median unemployment rate and above/below the median change in schooling years.

in political acts such as demonstrations, boycotts, and strikes, this link between education and political protest is stronger among those individuals who underperform in the labor market. To be specific, we constructed a measure of income underperformance as the extent to which an individual's reported income status falls below that predicted by a regression model based on a comprehensive set of observable characteristics for the individual, including importantly their level of schooling. We further verified that many of the Arab countries present in the World Values Survey also tended to display very high average levels of income underperformance, relative to other countries, for individuals who have some secondary education. In fact, Morocco, Iraq, Jordan, Egypt, and Algeria were among the worst-ranked countries on this count when compared with the rest of the world.

At this juncture, we should distinguish our argument from the long line of thought that has held that "grievances" can provide the fuel for demonstrations. For instance, Opp (1988) links grievances to participation in social movements, and Verba, Nie, and Kim (1987, p.161) argue that a "group [that is] particularly motivated by a sense of grievance vis-à-vis other groups . . . may become much more active than its socioeconomic level would predict." Such "grievance" effects driven by economic frustration were certainly present during the Arab Spring. Take for example the results of a poll conducted by the International Republican Institute in Egypt in April 2011, shortly after the resignation of President Mubarak. Out of the 28 percent of respondents who claimed that they had taken part in the recent protests, 64 percent cited "low living standards/lack of jobs" as their primary motivation, far above the 19 percent who mentioned "lack of democracy and political reform" (International Republican Institute 2011). Indeed, 41 percent of the respondents indicated that they "have trouble feeding [themselves] and [their] family and buying even the most essential things for survival."

We certainly do not discount the importance of the pure "grievance" effect. Nevertheless, we would argue that it becomes difficult to explain the patterns in the data fully if we do not complement a grievance-based explanation with the opportunity cost effect that we emphasize. After all, one could envision that the public could be angry at the economic situation but apathetic, particularly if the time and effort cost of engaging in political action is too high. In our interpretation, individuals would be more prone to devote their energies to political protest not only because they are aggrieved, but also because it is less costly to do so when labor market conditions are weak. And the weak labor market conditions lower the opportunity cost of political activity more for the highly educated than the less educated (whose time is less valued by the labor market regardless of broader economic circumstances) and more for effort-intensive activities than for less intensive ones. Thus, one can now explain why the combination of more schooling and poor economic conditions will manifest itself in effort-intensive forms of political participation without having to assume that the grievance motive would somehow both gain traction with more education and express itself especially through those forms of participation.

From a broader perspective, the opportunity cost argument also helps us to make sense of some evidence that leans against the mainstream finding of a positive impact of schooling on political participation at the individual level. Several papers have reported finding a small or even insignificant causal effect of education (for example, Tenn 2007; Kam and Palmer 2008; Berinsky and Lenz 2010). There is also some evidence that this correlation is considerably weaker in low-income countries (Pande 2011). However, all of these exceptions have to do with voting, a relatively less effort-intensive activity for which we would expect the opportunity cost forces to be weaker. Such findings underscore the need to distinguish between different forms of political engagement, which are quite possibly affected by the economic and institutional environment in different ways.

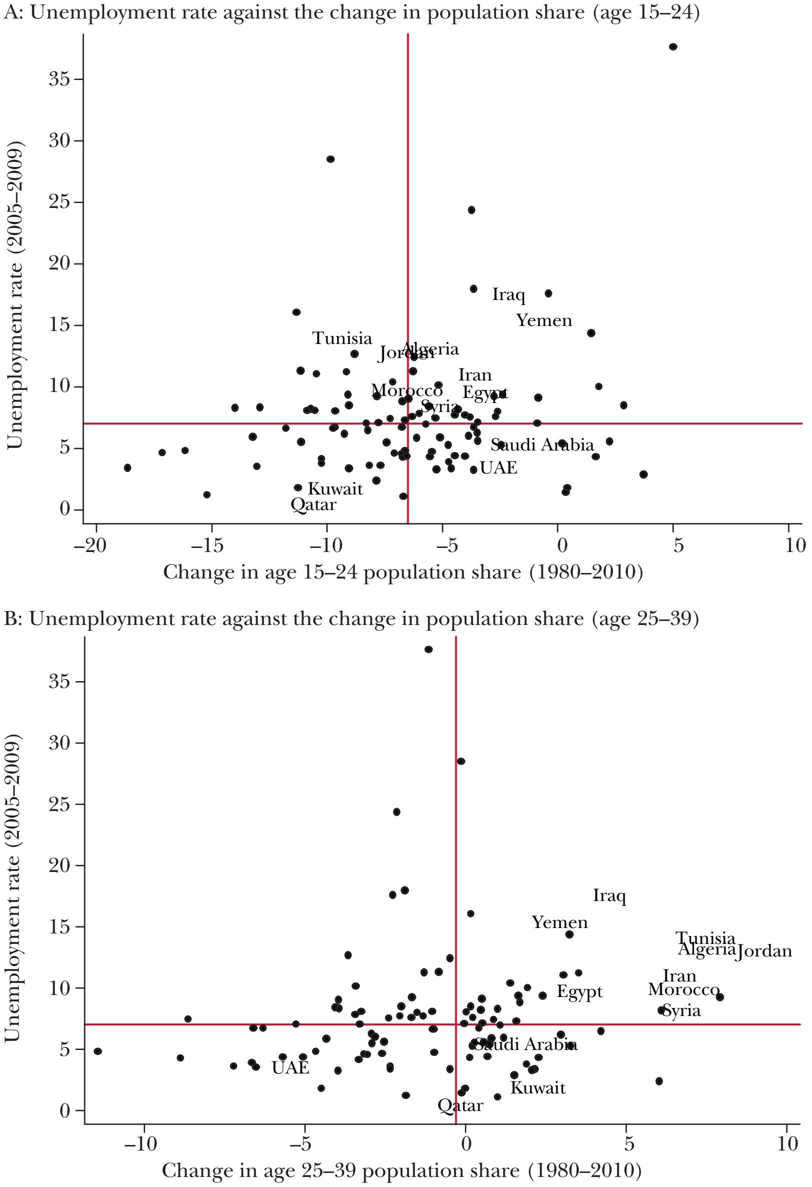
All in all, we have offered suggestive evidence that the combination of education and unrewarding economic circumstances is associated with an increased propensity towards political protest. Since we have also argued that the Arab world indeed witnessed both substantial investment in education and poor labor market conditions, it is plausible to think that this combination was important as a root cause of the Arab Spring.

Of course, a number of other structural forces have been mentioned in connection with the Middle East turmoil, and it is useful to acknowledge these and how they might complement our framework. In particular, the youthful demographic profile of the affected countries has featured prominently in these discussions, motivated by the evidence, from casual observation, and from arguments by political scientists like Urdal (2006), that youth are more prone to acts of political protest or violence. We briefly explore this "youth hypothesis" in Figure 4. The upper panel plots the change in the share of the population aged 15–24 (out of the population aged 15 and over) from the Barro and Lee (2010) dataset on the horizontal axis. A quick look at this figure reveals that Syria, Egypt, Yemen, and Iran did see a large increase in this youth cohort share from 1980 to 2010. On the other hand, countries like Tunisia, Jordan, Algeria, and Morocco actually experienced fairly large declines in the relative size of this youth cohort during this period.

The picture changes considerably however if we focus on the share of the population aged 25–39 instead. Figure 4B illustrates that all the countries that were affected by significant uprisings during the Arab Spring were also places where the share of the population aged 25–39 had increased considerably from 1980 to 2010 amid a labor market climate featuring high unemployment rates. It thus seems that the demographic shifts in this "not-so-young" age cohort might actually be more relevant for the recent Arab experience. It is perhaps no coincidence that the age 25–39 cohort consists of young workers entering the prime of their working age years, whose political behavior would be more liable to respond to prevailing conditions in the labor market. In other words, Figure 4B suggests that the "youth hypothesis" refers most fully to younger members of the workforce with poor economic prospects, who would be particularly inclined to revolt. We view this as entirely consistent with the opportunity cost mechanism we emphasize.

*Figure 4*

**Was it the Youth Revolution? Unemployment Rates against the Population Shares of Young Cohorts**



A: Unemployment rate against the change in population share (age 15–24)

B: Unemployment rate against the change in population share (age 25–39)

*Notes:* Arab League countries and Iran are labeled by their country names, while all other countries are indicated by circular markers. The *y*-axis plots the unemployment rate in the total labor force, averaged over 2005–2009. For the *x*-axis, Panel A plots the change in the population share aged 15–24 (as a share of that aged 15 and over), while Panel B plots the change in the population share aged 25–39 (as a share of that aged 15 and over). Both population share changes are calculated from 1980 to 2010. Lines indicating the median values of the *x*- and *y*-axis variables are included.

## Links to Political Change

Even if education and poor economic rewards tend to be correlated with a greater propensity towards protest activities at the individual level, it does not follow immediately that, at the aggregate level, these variables would be associated with political change. Of course, it is not enough to point to the Arab Spring as vindication: Why then did the Middle East experience this sudden wave of uprisings, but not other countries with similar structural conditions? In other words, why have the other countries in the upper-right quadrant of Figure 1A not gone through their own "springtime"? Perhaps they will in the future, but more generally it seems plausible that revolutionary episodes are not deterministic processes, so their occurrence might not even be guaranteed even in relatively favorable conditions.

Can we nevertheless say something more systematic about how the interaction of increased education and economic circumstances affects the probability of episodes of political change? To take a stab at these questions, we use a measure of political change that we have built based on turnover data on country leaders compiled from WorldStatesmen.org, an encyclopedia that provides detailed chronologies of heads of state and heads of government around the world.[8] We construct a binary indicator for whether there is any change in the chief executive running the country during a given five-year window, namely 1990–1994, 1995–1999, and so on. Note that, for this variable, we do not distinguish between changes that are brought about by peaceful democratic processes or violent revolutions, although we will return to a discussion on this issue later below. We estimate probit regressions on this indicator of executive change, where the key explanatory variables are the same measures of the unemployment rate and the change in the average years of schooling that we have already been using. Due to how sparse the unemployment data become across countries as we go further back in time, we focus our analysis on the period between 1990 and 2009, taking averages of the unemployment rate over each five-year window. (For the change in schooling variable, we calculate this for the 1990–1994 observation as the difference in average years of schooling between 1985 and 1990, and so on.) Given the panel nature of our approach, we include country and time dummies in our regressions, while also clustering the standard errors by country. The country fixed effects in particular are useful for controlling for omitted variables such as social or even geographical features. One might suspect for instance that the concentration of Sunnis, or whether a country is a Gulf versus non-Gulf state, might otherwise be correlated with a country's investment in education or its unemployment rate.

Table 2 contains the results of this exercise. The first column shows that the incidence of executive change is positively correlated with the unemployment rate; on the other hand, its correlation with changes in years of schooling

[8] To guard against concerns that might be raised over the open-source nature of the website, we have compared the records in WorldStatesmen.org against other data sources as a cross-check for the years in which political transitions occurred. See Campante, Chor, and Do (2009) for more details.

*Table 2*
**Labor Market Conditions and Schooling Gains: Implications for Incumbent Stability**

*Dependent Variable: Change in Executive (1 = Yes, 0 = No)*

| | *Probit regressions* | | | |
|---|---|---|---|---|
| *Independent variable* | (1) | (2) | (3) | (4) |
| Unemployment Rate | 0.086**<br>(0.037) | 0.044<br>(0.041) | 0.041<br>(0.041) | 0.082*<br>(0.048) |
| Change in Years of Schooling | 0.023<br>(0.411) | –1.448**<br>(0.586) | –1.440**<br>(0.582) | –1.298**<br>(0.581) |
| Unemployment Rate<br>× Change in Years of Schooling | | 0.130***<br>(0.043) | 0.129***<br>(0.043) | 0.105**<br>(0.047) |
| Change in Population Share 15–24 | | | 0.055<br>(0.124) | –0.113<br>(0.166) |
| Unemployment Rate<br>× Change in Population Share 15–24 | | | –0.005<br>(0.010) | 0.018<br>(0.014) |
| Change in Population Share 25–39 | | | | –0.207<br>(0.148) |
| Unemployment Rate<br>× Change in Population Share 25–39 | | | | 0.029**<br>(0.013) |
| Country dummies? | Yes | Yes | Yes | Yes |
| Year dummies? | Yes | Yes | Yes | Yes |
| Observations | 274 | 274 | 274 | 274 |
| Number of countries | 70 | 70 | 70 | 70 |
| Pseudo $R^2$ | 0.132 | 0.155 | 0.156 | 0.167 |

*Notes:* Standard errors in parentheses are clustered by country. All columns are probit regressions with country and year fixed effects. The data used run from 1990 to 2009. Observations are constructed for five-year windows, for example 1990–1994. The dependent variable is a binary variable indicating the occurrence of a change in the chief executive governing the country during a given five-year window. The unemployment rate is the average over the five years. The change in years of schooling and change in population shares are calculated as the value observed at the start of the window minus that observed at the start of the preceding five-year window (for 1990–1994, this is the 1990 value minus the 1985 value).
***, **, and * denote significance at the 1, 5, and 10 percent levels respectively.

is statistically insignificant. The second column uncovers how a combination of schooling increases and weak labor markets is associated with a greater likelihood of incumbent change: the coefficient of the interaction term between the change in average years of schooling and the unemployment rate is positive and significant at the 1 percent level. In column 3, we incorporate variables that would help to account for the possible role of a youthful demography in driving political change, specifically by controlling for the lagged five-year change in the share of the population aged 15–24, as well as its interaction with the prevailing unemployment rate. We find evidence here that the role of schooling increases is indeed statistically robust to and distinct from the "youth hypothesis," at least when the latter is represented by the age 15–24 cohort. Our central findings do seem to be complemented, though, by the change in the population cohort share that is aged 25–39, as shown in column 4. Countries that saw large increases in the size of this "not-so-young" cohort amid a labor market with a high unemployment rate are in turn more liable to experience a change in political leadership. This in fact augments the role of the underlying expansion of schooling, as the significance of the interaction of this previous term with the unemployment rate remains significant. Overall, we view this as consistent with the idea that the opportunity cost in labor markets helps to explain the patterns in the political turnover data, as this slightly older cohort should be most active in labor markets and not obviously more (or less) likely to hold grievances.

While we would refrain from placing too much emphasis on the specific numbers from these illustrative regressions, the lesson we draw is that our story of interest can be quite important and useful for predicting the likelihood of turnover.[9] These implications for regime stability are explored further and more formally in a cross-country regression framework in Campante and Chor (2011).

As previously mentioned, our cross-country analysis does not distinguish between the means by which regime change was triggered, in part to avoid having to make arbitrary choices regarding what constitutes an instance of turnover driven by a revolution. We would moreover argue that our mechanism applies just as well to a situation where an incumbent might be brought down within an existing democratic framework as a result of political pressure from mass demonstrations or strikes. Having said this, the Middle East countries do differ significantly from the other countries in the upper-right quadrant of Figure 1A on one key dimension: democracy. To illustrate this, we use the Polity IV database, which gives countries a "Democracy" score on a 0 to 10 scale, based on the assessment of a list of institutional features meant to capture the characteristics of a well-functioning democracy (Marshall and Jaggers 2011). The data are available on a yearly basis, and for each country in the upper-right quadrant of Figure 1A, we compute an average 2005–2009 score. As it turns out, the average score for the non–Middle East countries in that quadrant is around 8.3, in contrast with a Middle East average of 1.1. In other words, one dimension that sets the Middle East countries apart is that they are highly nondemocratic when compared with the other countries in that quadrant. Thus, one interpretation is that in the absence of democratic mechanisms for regime change, the societal pressures that had been building up in the Middle East against incumbents were pent up and eventually found expression in popular outbursts of protest. Alternatively, those pressures exerted by the combination of

[9] In terms of the quantitative implications of this probit model, the column 4 estimates suggest that a one standard deviation increase in the unemployment rate (roughly six percentage points for the countries in the sample in 2005–2009) would be associated with an increase in the probability of turnover of about 36 percent. About seven-tenths of this increase can be attributed to the main effect of the unemployment rate alone, with the remaining three-tenths due to the interaction of the unemployment rate with increases in schooling. (We evaluate all other country variables at their mean values in 2005–2009.)

schooling increases and poor economic circumstances might have found a more peaceful resolution within the democratic institutional framework in other countries, through, for example, the removal of incumbents via the electoral process.

This interpretation does receive some support from the data on incumbent changes. Of the 19 non–Middle East countries in the upper-right quadrant of Figure 1A, 13 had experienced at least one change in the identity of the chief executive between 2005 and 2009. In contrast, among the eight Middle East countries in that quadrant, only two saw a change, namely Iran and Iraq—and of course, Iraq is a peculiar case associated with a transition away from a foreign intervening power. Moreover, there is only one non–Middle East country, Kazakhstan, in this high unemployment, high schooling-gains quadrant, that shares a similarly low democracy score as the Middle East. While we would refrain from making overconfident predictions, we would not be shocked to see some political instability in Kazakhstan in the not-too-distant future.

## What Can We Hope to Learn?

We have argued that the lack of adequate economic opportunities for an increasingly educated populace can help us understand episodes of regime instability such as the Arab Spring. Our work in this area can be viewed in the context of the long-running debate over the "modernization hypothesis." In the classic view put forth by Lipset (1960), economic and institutional development tend to go hand in hand, and so "modernization"—including the expansion of education—naturally begets democracy. In an alternative view memorably put forward by Huntington (1968), modernization can instead be destabilizing in the absence of the necessary institutional infrastructure to support the process of change. Przeworski and Limongi (1997) survey and assess this extensive literature.

The major difficulty in distinguishing between these two camps has been empirical in nature. Cross-country studies generally confirm the strong positive raw correlation linking education and democracy (for example, Barro 1999; Przeworski, Alvarez, Cheibub, and Limongi 2000). However, the empirical work has struggled to show a clear causal effect from within-country increases in schooling to improvements in democracy. For example, Glaeser, La Porta, Lopez-de-Silanes, and Shleifer (2004) find Lipsetian results, while Acemoglu, Johnson, Robinson, and Yared (2009) argue that these findings are spurious, in that they are driven by the joint increase over the years of both education and the spread of democracy across rather than within countries.[10]

[10] Bobba and Coviello (2007) offer an econometric perspective on this debate. A similar challenge has emerged when it comes to the closely-related problem of establishing a causal effect flowing from income growth to democratization (Acemoglu, Johnson, Robinson, and Yared 2008, 2009). Moral-Benito and Bartolucci (2011) argue that there is evidence for a nonlinear effect, namely that rising income fosters democracy, but only up to a certain level of income.

Our approach seeks to break down the broader theories about modernization into more specific underlying mechanisms: in our case, the interaction between the schooling background and economic circumstances faced by individuals. We have focused on how economic circumstances affect the opportunity cost of political participation, but other mechanisms are certainly possible. In one recent example, Friedman, Kremer, Miguel, and Thornton (2011) have explored how schooling affects political views and attitudes using a randomized field experiment in Kenya that increased individual schooling levels via the assignment of merit scholarships. Their results point at increased knowledge within the treatment group, but also a greater disenchantment with politics and (perhaps surprisingly) a greater acceptance of political violence. Another approach looks at what is taught and how. Algan, Cahuc, and Shleifer (2011) find that different teaching methods—for example, group discussion versus lecturing—seem to influence students' views and beliefs.

Our broader thesis may also apply beyond the Middle East. For instance, several observers have pointed to China as an example of a developing country that has recently seen an increased scarcity of job opportunities for university graduates against the backdrop of a rapid expansion of the tertiary education sector in the past decade (Jacobs 2010; Eichengreen 2011). Our interpretation of recent events in the Arab world reinforces the view that China may well face a rise in political instability if the Chinese economy does not sustain a pace of growth that generates sufficient jobs to keep up with the education profile of its population. All in all, the connections between education, the economic environment, individual political engagement, and institutional change will most certainly continue to play a large role in driving political developments and dynamics in the years to come.

■ *We thank the editors and Chris Avery for their supportive and helpful comments.*

## References

**Acemoglu, Daron, and James Robinson**. 2001. "A Theory of Political Transitions." *American Economic Review* 91(4): 938–63.

**Acemoglu, Daron, and James Robinson.** 2005. *Economic Origins of Dictatorship and Democracy.* Cambridge, UK: Cambridge University Press.

**Acemoglu, Daron, Simon Johnson, James Robinson, and Pierre Yared.** 2008. "Income and Democracy." *American Economic Review* 98(3): 808–842.

**Acemoglu, Daron, Simon Johnson, James Robinson, and Pierre Yared.** 2009. "Reevaluating the Modernization Hypothesis." *Journal of Monetary Economics* 56(8): 1043–1058.

**Algan, Yann, Pierre Cahuc, and Andrei Shleifer.** 2011. "Teaching Practices and Social Capital." http://www.economics.harvard.edu/faculty/shleifer/files/teaching_practices_oct2011.pdf.

**Ammous, Saifedean, and Edmund Phelps.** 2011. "Tunisians Set Off on the Road from Serfdom." *Financial Times,* January 24.

**Barro, Robert.** 1999. "Determinants of

Democracy." *Journal of Political Economy* 107(S6): 158–83.

**Barro, Robert, and Jong-Wha Lee.** 2001. "International Data on Educational Attainment: Updates and Implications." *Oxford Economic Papers* 53(3): 541–63.

**Barro, Robert, and Jong-Wha Lee.** 2010. "A New Data Set of Educational Attainment in the World, 1950–2010." NBER Working Paper 15902.

**Bénabou, Roland**. 2000. "Unequal Societies: Income Distribution and the Social Contract." *American Economic Review* 90(1): 96–129.

**Berinsky, Adam J., and Gabriel S. Lenz.** 2010. "Education and Political Participation: Uncovering the Causal Link." *Political Behavior,* forthcoming.

**Besley, Timothy, and Torsten Persson.** 2011. "The Logic of Political Violence." *Quarterly Journal of Economics* 126(3): 1411–45.

**Bobba, Matteo, and Decio Coviello.** 2007. "Weak Instruments and Weak Identification in Estimating the Effects of Education on Democracy." *Economic Letters* 96(3): 301–307.

**Brady, Henry E., Sidney Verba, and Kay Lehman Schlozman.** 1995. "Beyond SES: A Resource Model of Political Participation." *American Political Science Review* 89(2): 271–94.

**Brückner, Markus, and Antonio Ciccone.** 2011. "Rain and the Democratic Window of Opportunity." *Econometrica* 79(3): 923–47.

**Campante, Filipe R., and Davin Chor.** 2011. "'The People Want the Fall of the Regime': Schooling, Political Protest and the Economy." HKS Faculty Research Working Paper RWP11-018.

**Campante, Filipe R., and Davin Chor.** Forthcoming. "Schooling, Political Participation, and the Economy." *Review of Economics and Statistics.*

**Campante, Filipe R., Davin Chor, and Quoc-Anh Do.** 2009. "Instability and the Incentives for Corruption." *Economics and Politics* 21(1): 42–92.

**Cassidy, John.** 2011. "Prophet Motive." *The New Yorker,* February 28.

**Charles, Kerwin Kofi, and Melvin Stephens Jr.** 2011. "Employment, Wages, and Voter Turnout." NBER Working Paper 17270.

**Clemens, Michael A., Claudio E. Montenegro, and Lant Pritchett.** 2009. "The Place Premium: Wage Differences for Identical Workers across the US Border." HKS Faculty Research Working Paper RWP09-004.

**Collier, Paul, and Anke Hoeffler.** 2004. "Greed and Grievance in Civil War." *Oxford Economic Papers* 56(4): 563–95.

**Courbage, Youssef, and Emmanuel Todd.** 2007. *Le Rendez-vous des Civilisations.* Paris: Seuil.

**Daun, Holger, and Reza Arjmand.** 2002. "Arab Countries: Oil Boom, Religious Revival and Non-Reform." In *Educational Restructuring in the Context of Globalization and National Policy,* edited by Holger Daun, 205–225. London: Routledge.

**Davies, James.** 1962. "Toward a Theory of Revolution." *American Sociological Review* 27(1): 5–19.

**Dee, Thomas.** 2004. "Are there Civic Returns to Education?" *Journal of Public Economics* 88(9–10): 1697–1720.

**Dewey, John.** 1916. *Democracy and Education.* New York: The Macmillan Company.

**Doing Business.** 2011. *Doing Business in the Arab World 2011.* Washington, DC: World Bank.

***Economist, The.*** 2011a. "The Shoe-Thrower's Index: Where Is the Next Upheaval?" February 10.

***Economist, The.*** 2011b. "The Economics of the Arab Spring: Open for Business?" June 23.

***Economist, The.*** 2011c. "The Arab Awakening: Revolution Spinning in the Wind." July 14.

**Eichengreen, Barry.** 2011. "Why Egypt Should Worry China." *Project Syndicate,* February 8.

**Fahim, Kareem.** 2011. "Slap to Man's Pride Set Off Tumult in Tunisia." *New York Times,* 21 January.

**Freeman, Richard B.** 2004. "What, Me Vote?" In *Social Inequality,* edited by Kathryn M. Neckerman, 703–728. New York: Russell Sage Foundation.

**Friedman, Willa, Michael Kremer, Edward Miguel, and Rebecca Thornton.** 2011. "Education as Liberation?" NBER Working Paper 16939.

**Glaeser, Edward, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer.** 2004. "Do Institutions Cause Growth?" *Journal of Economic Growth* 9(3): 271–303.

**Glaeser, Edward, GiacomoPonzetto, and Andrei Shleifer.** 2007. "Why Does Democracy Need Education?" *Journal of Economic Growth* 12(2): 77–99.

**Grossman, Herschel.** 1991. "A General Equilibrium Model of Insurrection." *American Economic Review* 81(4): 912–21.

**Hillygus, D. Sunshine.** 2005 "The Missing Link: Exploring the Relationship between Higher Education and Political Engagement." *Political Behavior* 27(1): 25–47.

**Huntington, Samuel P.** 1968. *Political Order in Changing Societies.* New Haven, CT: Yale University Press.

**International Republican Institute.** 2011. "Egyptian Public Opinion Survey, April 14–April 27, 2011" http://www.iri.org/sites/default/files/2011%20June%205%20Survey%20of%20Egyptian%20Public%20Opinion,%20April%2014-27,%202011_0.pdf.

**Jacobs, Andrew.** 2010. "China's Army of Graduates Struggles for Jobs." *New York Times,* December 11.

**Kam, Cindy D., and Carl L. Palmer.** 2008. "Reconsidering the Effects of Education on Political Participation." *Journal of Politics* 70(3): 612–31.

**Lipset, Seymour.** 1959. “Some Social Requisites of Democracy: Economic Development and Political Legitimacy.” *American Political Science Review* 53(1): 69–105.

**Lipset, Seymour.** 1960. *Political Man: The Social Basis of Modern Politics,* New York: Doubleday.

**Marshall, Monty G., and Keith Jaggers.** 2011. “Political Regime Characteristics and Transitions, 1800–2010.” Polity IV Project, University of Maryland. http://www.systemicpeace.org/polity/polity4.htm.

**Milligan, Kevin, Enrico Moretti, and Philip Oreopoulos.** 2004. “Does Education Improve Citizenship? Evidence from the United States and the United Kingdom.” *Journal of Public Economics* 88(9–10): 1667–95.

**Moral-Benito, Enrique, and Cristian Bartolucci.** 2011. “Income and Democracy: Revisiting the Evidence.” Banco de España Working Paper 1115.

**Noland, Marcus, and Howard Pack.** 2007. *The Arab Economies in a Changing World.* Washington, DC: Peterson Institute.

**Opp, Karl-Dieter.** 1988. “Grievances and Participation in Social Movements.” *American Sociological Review* 53(6): 853–64.

**Pande, Rohini.** 2011. “Can Informed Voters Enforce Better Governance? Experiments in Low-Income Democracies.” *Annual Review of Economics* 3: 215–37.

**Przeworski, Adam, Michael E. Alvarez, Jose A. Cheibub, and Fernando Limongi.** 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990,* Cambridge: Cambridge University Press.

**Przeworski, Adam, and Fernando Limongi.** 1997. “Modernization: Theories and Facts.” *World Politics* 49(2): 155–83.

**Putnam, Robert D.** 1995a. “Bowling Alone: America’s Declining Social Capital.” *Journal of Democracy* 6(1): 65–78.

**Putnam, Robert D.** 1995b. “Tuning In, Tuning Out: The Strange Disappearance of Social Capital in America.” *PS: Political Science & Politics* 28(4): 664–83.

**Rosenstone, Steven J., and John M. Hansen.** 1993. *Mobilization, Participation and Democracy in America.* New York: MacMillan.

**Schlozman, Kay Lehman.** 2002. “Citizen Participation in America: What Do We Know? Why Do We Care?” In *The State of the Discipline,* edited by Ira Katznelson and Helen Milner. New York: W. W. Norton.

**Sondheimer, Rachel Milstein, and Donald P. Green.** 2010. “Using Experiments to Estimate the Effects of Education on Voter Turnout.” *American Journal of Political Science* 54(1): 174–89.

**Tenn, Steven.** 2007. “The Effect of Education on Voter Turnout.” *Political Analysis* 15(4): 446–64.

**Urdal, Henrik.** 2006. “A Clash of Generations? Youth Bulges and Political Violence.” *International Studies Quarterly* 50(3): 607–629.

**Verba, Sidney, and Norman H. Nie.** 1987. *Participation in America: Political Democracy and Social Equality.* Chicago, IL: University of Chicago Press.

**Verba, Sidney, Norman H. Nie, and Jae-On Kim.** 1987. *Participation and Political Equality: A Seven-Nation Comparison,* Chicago, IL: University of Chicago Press.

**Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady.** 1995. *Voice and Equality: Civic Voluntarism in American Politics.* Cambridge, MA: Harvard University Press.

*Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 189–206*

# Using Internet Data for Economic Research

Benjamin Edelman

The data used by economists can be broadly divided into two categories. First, structured datasets arise when a government agency, trade association, or company can justify the expense of assembling records. The Internet has transformed how economists interact with these datasets by lowering the cost of storing, updating, distributing, finding, and retrieving this information. Second, some economic researchers affirmatively collect data of interest. Historically, assembling a dataset might involve delving through annual reports or archives that had not previously been organized into a format ready for research. In some cases researchers would survey stores, factories, consumers, or workers; or they could carry out an experiment. For researcher-collected data, the Internet opens exceptional possibilities both by increasing the amount of information available for researchers to gather and by lowering researchers' costs of collecting information. In this paper, I will explore the Internet's new datasets, present methods for harnessing their wealth, and survey a sampling of the research questions these data help to answer.

The Internet had 346 million sites as of June 2011, growing at a rate of 30 percent per year (Netcraft 2011). Many websites that are household names offer data of interest to economists. For example, Monster.com organizes available jobs; Amazon.com offers data on prices and sales of all sorts of items; eBay posts records of the bidding process for every listing; and Facebook and Twitter organize information about social connections, consumption choices, and preferences for privacy. The first section of this paper discusses "scraping" the Internet for data—that is, collecting data on prices, quantities, and key characteristics that are already available on websites but not yet organized in a form useful for economic research.

■ *Benjamin Edelman is Assistant Professor of Business Administration, Harvard Business School, Boston, Massachusetts. His website is ⟨www.benedelman.org⟩.*

http://dx.doi.org/10.1257/jep.26.2.189. doi=10.1257/jep.26.2.189

The second main section of the paper then considers online experiments. This category includes experiments that the economic researcher observes but does not control—for example, when Amazon or eBay alters site design or bidding rules. It also includes experiments in which a researcher participates in design, including researcher-designed experiments; experiments conducted in partnership with a company or website; and online versions of laboratory experiments. Finally, I discuss certain limits to this type of data collection, including both "terms of use" restrictions on websites and concerns about privacy and confidentiality.

A wealth of data exists online mainly because the general public wants access to it. With so much data readily accessible, researchers often need not obtain any special permission to obtain the data they seek. Furthermore, easy data availability helps avoid selection bias: data providers tend to decline researcher requests for data that (they believe) reflect unfavorably on them, but comprehensive online postings sometimes reveal surprisingly detailed information. The panorama of available online data may especially benefit researchers early in their careers. Online data is often available without delay and thus allows the preparation of original empirical research under tight time constraints—particularly helpful for students writing papers within the constraint of an academic term. Unrestricted access also assists anyone whose credentials might fail to satisfy the gatekeepers who evaluate requests for internal data from companies and organizations. Meanwhile, the costs and difficulty of collecting such data are modest. Most undergraduate computer science students can design a basic system to collect data from a website, using tools as basic as macros or scripts, so lack of advanced programming skills need not stand in the way.

Online data can speak to almost every field of economics, including subjects well beyond software, networks, and information technology. The next section identifies representative examples, focusing on novel data sources while flagging useful tools and techniques as well as recurring challenges.

## Scraping the Internet to Collect Data

Consumers and competitors push websites to post remarkable amounts of information online. For example, most retail booksellers would hesitate to share information about which items they sold. Yet eBay posts the full bid history for every item offered for sale, and Amazon updates its rankings of top-selling items every hour. Surveying job seekers can be time consuming, but sites like Monster.com organize available jobs, making it easier and quicker to track job search among narrower groups. Researchers in many fields of economics have discovered the benefits of online data collection. Table 1 presents papers, each using online data, drawn from every top-level category in the *Journal of Economic Literature* classification system. Below, I turn to specific methods of online data collection, with details on their respective applications as well as representative research using each method.

*Table 1*
**Diverse Papers Grounded in Online Data**

History of Economic Thought

**Azar, Ofer H.** 2007. "The Slowdown in First-Response Times of Economics Journals: Can it Be Beneficial?" *Economic Inquiry* 45(1): 179–87.

Examines trends in the timing of journals' responses to submitted manuscript, collecting response time data from journals websites.

Microeconomics

**Bajari, Patrick, and Ali Hortacsu.** 2003. "The Winner's Curse, Reserve Prices, and Endogenous Entry: Empirical Insights from eBay Auctions." *RAND Journal of Economics* 34(2): 329–55.

Bid data from coin sales on eBay reveal bidder behavior in auctions, including the magnitude of the winner's curse.

Macroeconomics and Monetary Economics

**Cavallo, Alberto.** 2011. "Scraped Data and Sticky Prices." January 27. http://www.mit.edu/~afc/papers/Cavallo-Scraped.pdf.

Daily price data from online supermarkets reveal price adjustment and price stickiness.

International Economics

**Philipp Maier.** 2005. "A 'Global Village' without Borders? International Price Differentials at eBay." DNB Working Paper No. 044, De Nederlandsche Bank.

Purchases at eBay reveal differences in real prices between countries, including differences between countries without currency friction.

Financial Economics

**Antweiler, Werner, and Murray Z. Frank.** 2004. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *Journal of Finance* 59(3): 1259–94.

Finds that online discussions help predict market volatility; effects on stock returns are statistically significant but economically small.

Public Economics

**Ellison, Glen, and Sara Fisher Ellison.** 2009. "Tax Sensitivity and Home State Preferences in Internet Purchasing." *American Economic Journal: Economic Policy* 1(2): 53–71.

At a comparison shopping service, click patterns reveal users' efforts to avoid sales taxes.

Health, Education, and Welfare

**Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant.** 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature,* February 19, Volume 457(7232): 1012–14.

Trends in users' web searches identify influenza epidemics.

Labor and Demographic Economics

**Edelman, Benjamin.** Forthcoming. "Earnings and Ratings at Google Answers." *Economic Inquiry*.

Measures labor market outcomes in an online research service, including higher earnings for experience, flexibility, and disfavored work schedules.

Law and Economics

**Bhattacharjee, Sudip, Ram D. Gopal, Kaveepan Lertwachara, and James R. Marsden.** 2006. "Impact of Legal Threats on Online Music Sharing Activity: An Analysis of Music Industry Legal Actions." *Journal of Law and Economics* 49(1): 91–114.

Observes the prevalence of users sharing music files via peer-to-peer networks, and analyzes users' response to an increased likelihood of litigation.

Industrial Organization

**Chevalier Judith, and Austan Goolsbee.** 2003. "Measuring Prices and Price Competition Online: Amazon.com vs. BarnesandNoble.com." *Quantitative Marketing and Economics* 1(2): 203–222.

Uses publicly available price and rank data to estimate demand elasticities at two leading sellers of online books, finding greater price sensitivity at Barnes & Noble than at Amazon.

*Table 1—continued*

Business Administration and Business Economics; Marketing; Accounting

**Brynjolfsson, Erik, Yu (Jeffrey) Hu, and Duncan Simester.** 2011. "Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales." *Management Science* 57(8): 1373–86.

> Data from an online clothing store show that the Internet reduces product search times and enables successful niche products.

Economic History

**Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden.** 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science,* January 14, 331(6014): 176–82.

> Analyzing approximately 4 percent of all books ever printed, as scanned by Google Books, the authors profile changes in word choice and grammar, the duration of celebrity, and the problems of censorship and suppression.

Economic Development, Technological Change, and Growth

**Seamans, Robert, and Feng Zhu.** 2010. "Technology Shocks in Multi-Sided Markets: The Impact of Craigslist on Local Newspapers." NET Institute Working Paper 10-11.

> Explores the response of local newspapers to entry by Craigslist, including increase in subscription price, decreasing advertising price, and decreasing classified price.

Economic Systems

**Hsieh, Chang-Tai, Edward Miguel, Daniel Ortega, and Francisco Rodríguez.** 2011. "The Price of Political Opposition: Evidence from Venezuela's Maisanta." *American Economic Journal: Applied Economics* 3(2): 196–214.

> Drawing on a "punishment list" published by Hugo Chavez, containing the names of people who signed a petition against him, this paper estimates the "price of political opposition"—the lost earnings of disfavored persons.

Agricultural and Natural Resource Economics

**Camacho, Adriana, and Emily Conover.** 2011. "The Impact of Receiving Price and Climate Information in the Agricultural Sector." IDB Working Paper IDB-WP-220.

> Finds that farmers who received SMS (text message) information about price and weather had a narrower dispersion in expected price of their crops and a significant reduction in crop loss.

Urban, Rural, and Regional Economics

**Kroft, Kory, and Devin G. Pope.** 2008. "Does Online Search Crowd Out Traditional Search and Improve Matching Efficiency? Evidence from Craigslist." http://faculty.chicagobooth.edu/devin.pope/research/pdf/JPE_Final_with_figures.pdf.

> Measures the number of Craigslist posts in metropolitan areas to observe Craigslist's effect on classified ads, apartment and housing rental vacancy, and unemployment.

## The Basics of Online Data Collection

For a researcher seeking to collect data from the Internet, the simplest approach is typically a one-time collection of structured information from a single site. For example, Roth and Ockenfels (2002) retrieve data from about 480 eBay and Amazon auctions, finding that Amazon's automatic extension of auction closing time obviated the incentive for late bidding (which was observed much more often at eBay). Carlton and Chevalier (2001) retrieve data about prices of fragrances, DVD players, and refrigerators from an array of retail sites to explore the relationship between retail promotion, online availability, and price dispersion. They find that manufacturer websites tend to charge higher prices than retail websites, and that manufacturers who limit distribution of their product in the physical world also tend to avoid having their product appear on websites that offer especially deep discounts. Freedman and Jin (2011) retrieve data about lenders and borrowers at the microfinance person-to-person lending site Prosper.com, assessing the magnitude of asymmetric information, efforts to mitigate information asymmetry, and pricing of risk. They find that early lenders apparently did not understand the risks involved, and as the lenders learned to evaluate these risks, high-risk borrowers found themselves unable to receive loans through this mechanism.

Online data collection typically calls for a three-step procedure: A first component retrieves web pages from a web server, often checking an input file for a list of pages to retrieve, terms to search for, or similar instructions. Next, a parser receives web server replies and extracts the data elements of interest to the researcher, storing this data in an output file. Finally, a researcher can use standard tools to import and analyze the output file. Of course numerous refinements on this approach are possible, and subsequent sections present some of the variations.

Researchers can implement data collection systems in a number of programming languages. When demonstrating data collection in a classroom or assignment, I often offer Visual Basic for Applications (VBA) code that stores data directly in Excel—letting Excel handle storage of both inputs and outputs. Figure 1 presents sample code from this approach. For larger projects, researchers often choose Perl, PHP, or Python.

**Long-Term Data Retrieval**

It is sometimes helpful to collect data over an extended period, which yields a larger sample as well as insight into changes over time. When data collection is repeated periodically—a step that can be programmed to occur at preset time intervals—a one-time script can be converted into a system for continued data retrieval.

For example, Ellison and Ellison (2009) collect data on consumers making online purchases of computer memory modules. They examine hourly data over a year, using the Pricewatch.com search engine, which displays lists of products being sold by participating online retailers at their own website. To measure users' incentive to buy from out-of-state online sellers (often thereby avoiding state sales tax), Ellison and Ellison need variation in users' choice sets to draw conclusions about user preferences. Because choices change as online sellers adjust their prices, more-frequent data improves the power of the analysis. They find that states with higher sales taxes for off-line purchases tend to have more online purchases—with buyers thereby often avoiding the state sales tax.

Long-term data collection can be particularly important for analysis of price levels and inflation. Cavallo (2011) collected daily prices of 80,000 supermarket products taken from the public web pages of online retailers in four countries over a period of three years. This price data is then used to explore patterns in price stickiness. Cavallo finds a bimodal pattern of price changes—that is, price changes are significantly positive or negative but few changes are close to zero, confirming

*Figure 1*
**Sample Worksheet and Script for Basic Online Data Collection**

|   | A | B |
|---|---|---|
| 1 | ISBN | Rank |
| 2 | 0300151241 |  |
| 3 | 0691143285 |  |
| 4 | 0195340671 |  |

```
Sub GetAmazonDataDemo()
    ‘retrieves sales ranks from Amazon
    ‘INPUT:         sheet1 column A – gives a list of ISBN-10 numbers
    ‘OUTPUT:        sheet 1 column B – reports Amazon sales ranks

     Dim curcell As Range, htmlresponse As String     ‘required variables

     ‘iterate through rows of the first column
     For Each curcell In ActiveWorkbook.Sheets(“Sheet1”).Range(“A:A”)
         ‘leave the loop when encounter a blank value in column A
         If curcell.Value = “” Then Exit For

         ‘row 1 is the header row – don’t do anything there
         If curcell.Row > 1 Then
             ‘get Amazon product detail page with this ISBN
             htmlresponse = GetURL(“http://www.amazon.com/o/ASIN/” & Trim(curcell.Value))

             parse1 = GetBetween(htmlresponse, “Sales Rank:”, “ in”)
             rankval = GetAfter(parse1, “#”)

             ‘store rank in column B
             curcell.Offset(0, 1).Value = rankval
         End If
    Next curcell      ‘proceed to next row
End Sub

Function GetURL(url)
    ‘INPUT:    a URL to be retrieved
    ‘OUTPUT:the HTTP body at the specified URL
    ‘REQUIREMENTS:           a HTTP object: Winhttp, Xmlhttp, Winhttprequest, or Serverxmlhttp
    ‘ERROR HANDLING:         none – should be added for production use!

     On Error Resume Next
     Set objHTTP = CreateObject(“WinHttp.WinHttpRequest.5.1”)
     If objHTTP Is Nothing Then Set o = CreateObject(“Microsoft.xmlhttp”)
     If objHTTP Is Nothing Then Set o = CreateObject(“winhttp.winhttprequest”)
     If objHTTP Is Nothing Then Set o = CreateObject(“MSXML2.ServerXMLHTTP”)
     objHTTP.Open “GET”, url, False
     objHTTP.send
     GetURL = objHTTP.responsetext
End Function

Function GetBetween(s, s1, s2)
    ‘INPUT:    s - string to search;
              s1 - string that marks start of retrieval
              s2 - string that marks end of retrieval
    ‘OUTPUT: the portion of s that comes strictly between s1 and s2
    ‘ERROR HANDLING: if s1 or s2 is not found, returns a blank string

    p1 = InStr(s, s1)
    If p1 = 0 Then Exit Function              ‘s1 was not found
    p2 = InStr(p1 + Len(s1), s, s2)
    If p2 = 0 Then Exit Function              ‘s2 was not found
    GetBetween = Mid(s, p1 + Len(s1), p2 - p1 - Len(s1))
End Function

Function GetAfter(ByVal s As String, ByVal s1 As String)
    ‘INPUT: s - string to search; s1 - string that marks start of retrieval
    ‘OUTPUT: the portion of s that comes strictly after s1; if s1 is not found, all of s
    p1 = InStr(s, s1)
    If p1 = 0 Then GetAfter = s: Exit Function
    GetAfter = Mid(s, p1 + Len(s1), Len(s) - p1 - Len(s1) + 1)
End Function
```

the predictions of menu cost models. He also finds synchronization of prices for goods that are close competitors within product categories.

### Multistep Data Retrieval

More complex data retrieval systems can monitor multiple sources and adjust their configurations based on what occurs. For example, in Edelman (2002), I use a two-step process to measure the effects of recommendations by Amazon's editorial staff. A first system identifies which books Amazon's editorial staff recommend, running repeatedly to uncover new books soon after the recommendations begin. Then a second system tracks the sales rank of each such book—allowing measurement of the sales increase attributable to Amazon's recommendation. This approach allows examination of interaction between multiple economic actors, and if one set of events is at least locally exogenous, this approach can identify causal effects.

To measure users' sharing on peer-to-peer networks, Bhattacharjee, Gopal, Lertwachara, and Marsden (2006) also rely on a multistep collection process. A first system searches for randomly selected genres to retrieve a list of users sharing music. For selected users, a second system then activates a "Find More From Same User" function to retrieve information about the total number of songs that user is sharing. The authors collect weekly data for a year, yielding a measurement of users' response to threatened litigation by the recording industry. Notably, Bhattacharjee et al. collect data not from sites presented to users in browsers such as Internet Explorer and Firefox, but rather from Kazaa and WinMx, two file-sharing programs users can install on their computers. Data collection in this context calls for scripting to operate programs' menus and buttons and to capture results from program windows.

### Collecting Data from Secondary Sources

For some purposes, researchers may find it preferable to collect data from aggregators that assemble data from multiple underlying sources. For example, Baye, Morgan, and Scholten (2004) examine retailers' prices at a comparison shopping service (a website that lists prices of selected items at multiple retailers) to measure the breadth of price dispersion. Baye, Morgan, and Scholten find substantial price dispersion and little evidence of "the law of one price."

For any researcher needing information about the history of a website—whether on a one-off basis or for large-sample analysis—the Internet Archive is the natural choice. With copies of 150 billion pages dating back to 1996 (Internet Archive 2011), the Internet Archive provides no-charge access to prior versions of most online materials, facilitating all manner of historic analysis. Seamans and Zhu (2010) use the Internet Archive to gather historic data on Craigslist postings to explore relationships between the entry of Craigslist into a market and newspaper circulation and pricing.

While Internet Archive preserves historic materials, certain secondary sources analyze and tabulate *current* user behavior. For example, Google Trends reports the frequency of particular searches at Google. Using Google Trends data, Choi and Varian (2009) predict future filings for unemployment benefits, finding an improvement over official government forecasts. Wu and Brynjolfsson (2009) use Google searches to predict housing prices and sales, while Ginsberg, Mohebbi, Patel, Brammer, Smolinski, and Brilliant (2009) use data from Google searches to detect influenza epidemics.

### Web 2.0 Data Sources

The term "Web 2.0" denotes the Internet's change from simple screens of unchanging information to interactive communications in which users contribute ever more content—often making sharply more information available to the general public and to researchers. Facebook and Twitter are prominent examples in this vein.

Researchers seeking to study activity at Facebook benefit from default privacy settings that let the general public view each user's name, friends, networks, wall posts, photos, likes, and more. The resulting data can facilitate research on myriad topics. For example, Baker, Mayer, and Puller (2011) use Facebook to assess the diversity effects of randomized dormitory assignment. They find that students randomly exposed to persons of a different race have more friends of that race within the dormitory, but no greater diversity in social networks outside that environment.

Iyengar, Han, and Gupta (2009) look at data from Cyworld, a social networking site in Korea, in which users often decorate their mini-homepages with items like wallpaper or music purchased from Cyworld. Using 10 weeks of purchase and nonpurchase data from 208 users, they identify a low-status group that is not affected by the purchases of others; a medium-status group that has a positive correlation with the purchases of others; and a high-status group that has a negative correlation with the purchases of others. Acquisti and Gross (2006) examine demographic and behavioral differences in users' views of privacy. They find that privacy concerns expressed in survey results did not seem to limit which people joined Facebook nor how much information they revealed—in part, because those who joined did not fully understand what data was public. Default privacy settings at the short-message service Twitter also facilitate research: with few exceptions, Twitter messages are public, and Twitter also publishes the list of all authors each Twitter user is "following." While few researchers have embraced Twitter data, Vincent and Armstrong (2010) assess high-frequency trading strategies grounded in messages on Twitter, finding a profit opportunity in fast-breaking Twitter discussions.

The Internet's newest services also facilitate research on users' views of companies and organizations. Every company and organization page on Facebook includes a "like" button, and Facebook, Google, and others now let sites present "like," "+1," and similar buttons to garner user endorsements. The number and/or identity of users clicking these buttons is often available to researchers and the interested public—facilitating research about trends and trendsetters, reaction to news, and more.

### Monitoring Network Activity

The computer science literature features papers that analyze users' network traffic to draw conclusions about online activity. For example, Saroiu, Gribble, and Levy (2004) monitor the online activity of students at a university to identify computers infected with "spyware"—software like Gator, Cydoor, SaveNow, and eZula that gathers information about computer use and relays it to a third party, often without the consent (or informed consent) of the computer user. The authors monitor infections to identify types of users and behaviors particularly likely to suffer from spyware. Karagiannis, Broido, Brownless, claffy, and Faloutsos (2004) observe network activity at two Internet service providers to measure the prevalence of peer-to-peer file-sharing software, finding that contrary to a common belief at that time, peer-to-peer file sharing was not declining in response to concerns over its legality, but was in fact continuing to increase. In principle, economists could use similar methods. As Saroiu, Gribble, and Levy (2004) and Karagiannis et al. (2004) demonstrate, network monitoring systems are notable for the breadth of data they can observe—in principle, every online activity of users on the corresponding networks. They are also notable for their ability to collect data no individual site can, or cares to, assemble.

On a larger scale, commercial services analyze network traffic to identify trends in user behavior and site popularity. Best known is comScore (2011), which offers a two-million member panel of users recruited to install tracking software that monitors their browsing, purchasing, and other online activities. Compete offers a similar service, also via a panel of participating users, while Hitwise collects data from both users and Internet service providers.

### What Types of Data are Abundant or Scarce?

Online data collection projects tend to reveal certain data far more readily than others. For example, researchers seeking prices are easily satisfied; myriad sites post product price information as they offer items for purchase. But a researcher needing sales quantities faces greater difficulty: while a few sites, most notably eBay, post substantial information about each sale, most sites have no business need to distribute sales information systematically and publicly.

Work-arounds can yield quantity data. Some sites report the quantity of in-stock inventory available for each listed product. Slow decreases in this quantity are typically interpreted as indicating items sold. Large increases are understood to reflect arrival of additional inventory. By checking sufficiently frequently, a researcher can infer purchases.

Other sites report sales ranks, which offer insight into sales quantities. Best known in this area is Amazon, where many an author diligently tracks sales rank. Chevalier and Goolsbee (2003) pioneered procedures for converting Amazon sales ranks to sales quantities using multiple methods including cross-checks with publishers (who report that a given sales rank in a given week matches a given quantity), controlled experiments (purchasing and/or returning a given quantity of books and monitoring changes in rank), and fitting a portion of the distribution using publicly known sales quantities (for example, for bestsellers whose sales are occasionally revealed to the public). Depending on available data, these approaches can be used to extract quantities from other ranking systems.

In certain contexts, researchers may be able to infer useful information about sales from other materials that are posted at sites, such as "top sellers," "recommended items," or "people who bought this also bought . . ." suggestions. However, such methods are not yet well developed.

### Data Requiring Company or Site Cooperation

While the preceding sections identify data that researchers can collect directly from websites, some research questions require data that sites decline to publish to the public. In this context, it has proven fruitful to request additional data from the companies and organizations that run such sites. For example, Hitsch, Hortaçsu, and Ariely (2010) obtained records from an online dating site revealing all aspects of users' activities, including browsing, viewing photos, and sending and receiving messages. This data is necessary for Hitsch, Hortaçsu, and Ariely's evaluation of the quality of users' matches; they find that at the dating site, "the actual matches are approximately efficient." Data providers are often concerned about distribution of their internal records. Below, I discuss methods to address such concerns and protect privacy.

## Online Experiments

Scraping the web for existing data can raise concerns about whether observed changes can be treated as exogenous or endogenous. After all, when users and sites make decisions based on changing external circumstances to advance their respective objectives, it can be difficult to draw inferences about what factor caused what outcome. An experimental method can yield better insight into causation. Online experiments can also observe long-run behavior changes, whereas most short-run experiments risk overemphasizing short-run substitution effects.

In the taxonomy of List (2011) in this journal, online experiments can take four forms: "Natural experiments" arise from exogenous changes created by third parties (or nature) that mimic the conditions of an experiment. "Laboratory experiments" place the agents in an artificial game-like setting to see how they react. "Field experiments" present agents with randomized variation in conditions—changes agents experience while carrying out their usual online activities in their homes or at work, though they are aware that an experiment is being conducted. Finally, "natural field experiments" present agents with randomized variation in natural settings without informing agents that they are involved in an experiment.

### Natural Experiments

Sometimes, a site or service changes design parameters arbitrarily or in a time or manner unlikely to be correlated with other outcomes. Such circumstances can create a natural experiment yielding insight into lines of causation.

For example, Roth and Ockenfels (2002) point out that system designers set the ending rules for eBay and Amazon auctions without consideration of implications for bidder behavior. Thus, these sites provide a reasonable context to assess the effect of such rules. At the time of their study, eBay auctions closed at a fixed time, while Amazon auctions continued until ten minutes had gone by without a bid. Bidders reacted to such rules, submitting a higher fraction of late bids on eBay.

Other exogenous variation comes from unexpected changes. For example, Miller (2011) relies on a large increase in the amount of information available about borrowers on Prosper.com, a change which made lenders more selective among high-risk borrowers. Chiou and Tucker (2011) note a 2009–2010 dispute between Google and Associated Press that led to the temporary removal of AP stories from Google News (which aggregates news content from many sources). During that period, Google News referred fewer users to *all* traditional news sites, compared to other news aggregators that continued to host AP articles.

### Researcher-Designed Experiments

Some researchers participate in online markets in order to build what are, in essence, online field experiments. For example, Hossain and Morgan (2006) list items on eBay with varying listing prices and shipping prices, showing that users undervalue shipping cost relative to item price. By varying reserve policies in listings at eBay, Katkar and Reiley (2006) find that secret reserve prices deter bidder entry and reduce the likelihood of a listing resulting in a sale. Resnick, Zeckhauser, Swanson, and Lockwood (2006) auction matched pairs of items on eBay, some using a seller's well-established identity and others using new identities, thereby identifying the willingness of buyers to pay for seller reputation.

Many companies have recognized the benefit of experiments in improving their own operations. For example, online marketers test dozens of alternative advertisements. Tools like Optimizely and Visual Website Optimizer let a designer evaluate user behavior in multiple variants of a site—testing alternative layout, color, text, and more. Varian (2010) discusses the benefits Google has achieved through comprehensive experiments to evaluate possible changes to its services.

In lieu of a researcher running experiments, Einav, Kuchler, Levin, and Sundaresan (2011) flag the possibility of a researcher identifying experiments others are already running. If an eBay seller is testing variations in item listing (perhaps which format, description, or pricing achieves the highest price), a researcher can find these variations, retrieve data about both the experimenter's changes and the public's response, and thereby draw conclusions about the effect of the changes at issue. Einav et al. find that of the 100 million listings on eBay each day, more than half will reappear on the site, often with differing parameters for the sale. Assembling a dataset with hundreds of thousands of such matching pairs during a single year, these authors examine questions about price dispersion, bidding under different sets of rules, and customers' reaction to shipping fees. Practitioners' experiments can offer a vastly larger sample than researcher-implemented experiments—in turn, yielding more precise estimates. Practitioners' experiments also often occur across product categories, whereas practical concerns often limit researcher-implemented experiments to narrow categories, impeding inferences about other areas.

### Experiments in Partnership with a Company or Site

Some kinds of online field experiments tend to require cooperation from a company or site operator. For example, Chen, Harper, Konstan, and Li (2010) look at MovieLens, an online site that offers recommendations for movies. Submitting movie recommendations is a public good—benefiting all other users of the site, at some cost to the specific user who makes time to contribute. Many movies had too few recommendations for the software to match them with potential users, but Chen et al. find that when MovieLens subscribers are informed of their standing in terms of how many recommendations they make relative to other users, they tend to contribute more recommendations. Chen et al. partnered with Movielens in order to provide such notifications to a random set of users. Online advertising has proven particularly well suited to experiments with company cooperation. Reiley, Li, and Lewis (2010) change the number of advertisements presented at the top of the page at an Internet search engine, finding that when more such advertisements are presented, users click more often on the top-most advertisement. Ostrovsky and Schwarz (2011) adjust reserve prices in Yahoo! auctions for online advertisements, finding large revenue increases when reserve prices are set optimally.

With additional technical complexity, researchers may be able to conduct experiments entailing modification of a website even without participation by or cooperation from that site. In Edelman and Gilchrist (2010), my coauthor and I build a proxy that presents some users with modified search result pages showing hypothetical alternative advertisement labels: in place of the usual "sponsored link" or "ad" labels, some users instead saw labels reading "paid advertisement." Users with low education or little online experience benefit most from the "paid advertisement" label, which the Federal Trade Commission has sought in other media. Similarly, Schechter, Dhamija, Ozment, and Fischer (2007) present varying security warnings as users attempt to access online banking applications, finding that few users recognize the warnings intended to flag possible attacks. They also flag the importance of realistic experimental conditions: users who participated in the experiments as role players were far less concerned with security than those who used their own actual passwords.

### Online Lab Experiments

Online experiments can address many of the questions historically explored in real-world economics laboratories. For example, Horton, Rand, and Zeckhauser (2011) replicate three classic lab experiments in an online lab: the extent of cooperation in a one-shot prisoners' dilemma game; playing the prisoners' dilemma game after being "primed" by reading various religious or nonreligious texts (which tends to reduce rates of defection); and testing the "framing" result of Kahneman and Tversky(1979) that choices will differ depending on how questions are framed, because people are risk averse in the domain of gains but risk seeking as to losses.

Horton, Rand, and Zeckhauser also conducted a natural field experiment in which subjects were offered the opportunity to be paid to transcribe a simple passage of text in return for a compensation that had been randomly determined—demonstrating an upward-sloping supply of labor. Mason and Suri (2011) argue that online lab experiments using Mechanical Turk (which allows hiring people anywhere in the world to carry out tasks that can be performed online) can offer important benefits over traditional laboratory experiments, including easier access to a large and diverse subject pool, low cost, and faster deployment of new experiments.

Online lab experiments also present potential downfalls. As in physical economics laboratories, participants who sign up to participate are unlikely to be representative of the population as a whole, and their differences might be correlated with some treatments. Online subjects can exit a study more readily than subjects in a lab, which may be a concern if certain treatments disproportionately prompt early exit. Communication between online subjects may be possible, both during an experiment and between experiments, depending on a researcher's method of recruiting subjects. To deter communication among subjects, Horton, Rand, and Zeckhauser (2011) suggest running online experiments quickly and avoiding notoriety.

If researchers so choose, online experiments can blur the boundaries between the lab and the field. For example, Centola (2010) builds an online social network where participants can see the health behaviors of selected other participants assigned to be their "health buddies." From one perspective, this appears to be a natural field experiment: users participate from home or work, for an extended period, not knowing that they are subject to randomized variation in an academic research project. Yet participants are interacting in an environment constructed from scratch specifically for research purposes. Just as designers of a lab experiment design the rules of their system, Centola controlled most aspects of what participants could see, say, and do. With the right design, online experiments may be able to combine positive aspects of lab and field experiments.

## Limits to Internet-Based Data Collection

### Terms of Use and Similar Restrictions

Most websites present a "Terms of Use" or similar document that purports to restrict the methods and purposes of data access. Such statements are widespread; for example, Amazon, eBay, and Google all include provisions stating that users must not copy data from their sites. A series of court cases hold that such agreements are enforceable against competitors seeking to copy data for reasons courts view as improper. For example, in *eBay vs. Bidders' Edge* (100 F.Supp.2d 1058 [N.D. Cal. 2000]), Bidders' Edge sought to copy eBay data to build an auction aggregation service—a service which, if successful, would have undercut eBay's competitive advantage. eBay could therefore offer a cogent notion of harm—not just a few extra requests for its web server to answer, but a genuine business loss.

In contrast, researchers are far less disruptive to the site whose data is being copied. For example, researchers' activities are usually limited to analyzing data but not republishing or redistributing—and certainly not reselling—the information they collect. Furthermore, most researchers access online data that sites produce and distribute incidental to other activity, and researchers' activities do not interfere with sites' core business models. In this context, researchers typically perceive that they have strong defenses to any claims that data providers might bring. Finally, for lack of an urgent business harm, a target site is less likely to press the point.

In practice, sites most often respond to researchers' activities not by filing lawsuits, but by blocking access from computers that send too many requests. Seeing many requests from a single IP address (roughly, a single computer), it is usually straightforward for a site to configure its web server to deny further requests from that computer. Such a blockage often suffices to prompt a researcher to scale back data collection. That said, many researchers nonetheless continue requesting data even after a blockage—for example, using a different computer or a different IP address. Excessive requests could slow access by others and invite further bans, but most researchers' data requirements can be adequately addressed using a data collection system that operates at a rate similar to an ordinary user, sending perhaps one request every few seconds.

To date, to the best of my knowledge, no data provider has filed suit against a researcher collecting data that is available, in smaller quantities, to the general public without charge. Of course a researcher facing notable problems—perhaps accessing data that is otherwise made available only under a paid license—might do well to seek guidance from a qualified attorney.

### Privacy and Confidentiality

A researcher collecting online data must also consider privacy concerns. If collecting data about individuals, a researcher should consult the appropriate human subjects committee. That said, many human subjects committees will readily give the researcher permission for an online data collection project to proceed quickly and without conditions. In particular, human subject committees often give blanket permission or even waivers for studies that involve observation of public behavior, analysis of existing data, and gathering information in a way that that presents minimal risk to subjects.

When collecting data from a secure or semisecure site or when analyzing company data or other internal data, researchers may want to shield themselves from user-specific data. Sometimes, researchers need not even receive sensitive information. For example, when Hitsch, Hortaçsu, and Ariely (2010) analyzed user behavior at a dating website, they analyzed user data that contained no names, contact information, or images. But other research requires at least limited analysis of sensitive information or data derived from sensitive information. For example, Ian Larkin and I analyzed working paper downloads at the SSRN (Social Science Research Network) website to investigate whether the numbers were being "gamed" by authors downloading their own papers repeatedly to increase reported download counts. In doing this analysis, we did not want to see the names of the authors whose papers were most downloaded in circumstances suggesting gaming, but we did seek to analyze relationships between gaming and authors' professional standing, coauthors, and peers. To limit data in this way, we kept author names in a restricted table with limited access rights. When we needed analysis of authors' resumes and biographical information, we provided our research assistants with access to authors' names, but our assistants could not view information about paper downloads, and they did not know the purpose of our study. These procedures prevented anyone, including us, from connecting particular download data to a particular author. We found limited evidence of gaming due to demographic factors and career concerns, but strong evidence of gaming driven by social comparisons with various peer groups (Edelman and Larkin 2009).

When requesting data from companies, additional protections can help address concerns about confidentiality and about the possibility of readers uncovering company identity. It is routine to describe a corporate data source in general terms (like sector and approximate size) but to decline to name the specific company. But creative researchers can do more to protect data details. For example, when analyzing effectiveness of an online advertising campaign performed by Yahoo! and a major retailer, Lewis and Reiley (2011) created a database of over one million customers matched in the databases of the two companies. However, they then hired an outside vendor to render the data anonymous by merging together all of the personally-identifying information about users' online and offline activities. In addition, the vendor multiplied actual sales amounts by an undisclosed number between 0.1 and 10—preventing readers, or even the researchers, from learning the true amount of the company's advertising costs, incremental revenue, or other dollar figures.

Even data not intended to identify individuals may prove easily linked to specific persons. For example, in 2006 AOL's research division posted search data from 650,000 users—a dataset AOL intended to offer for academic research by anyone interested. AOL believed users' privacy was adequately protected because AOL published only users' search requests, not their names, usernames, or e-mail addresses. But some users could be identified from their unusual searches—including searches for their own names, value of their homes, and the like (Barbaro and Zeller 2006). Indeed, even a narrow range of possibilities for users' Social Security numbers can be inferred based on birthplace and date of birth, which are often publicly available (Acquisti and Gross 2009). With re-identification of individuals possible in unexpected circumstances, protecting privacy requires careful planning and ongoing vigilance.

## Opportunities

Opportunities for research using the Internet expand every year. New sites and services collect and retain ever more data, while mobile devices collect data even more widely. Sites and users have been remarkably willing to share much of their data with anyone interested, and the opportunities for economic research are limited primarily by researcher time and creativity. Furthermore, with certain kinds of data increasingly widely available, future researchers may be able to replicate results with similar methods and different datasets rather than using different methods on the same sections.

Meanwhile, in the realm of experiments, online data offers advances on questions of exogeneity and identification. A website design change is often plausibly exogenous, whereas real-world events like government policies are typically correlated with other events. Online systems also make it particularly easy—and, in some contexts, increasingly routine—for different users to receive different treatments on a widespread and ongoing basis. These circumstances combine the identification offered by experiments with the realism of naturally occurring data—potentially giving researchers the best of both methodologies.

■ *I thank Paul Kominers and Xiaoxiao Wu for excellent research assistance.*

## References

**Acquisti, Alessandro, and Ralph Gross.** 2006. "Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook." Presented at the Sixth Workshop on Privacy Enhancing Technologies, Cambridge, United Kingdom, June 28 –30.

**Acquisti, Alessandro, and Ralph Gross.** 2009. "Predicting Social Security Numbers from Public Data." PNAS 106(27): 10975–80.

**Baker, Sara, Adalbert Mayer, and Steven L. Puller.** 2011. "Do More Diverse Environments Increase the Diversity of Subsequent Interaction? Evidence from Random Dorm Assignment." *Economics Letters* 110(2): 110–112.

**Barbaro, Michael, and Tom Zeller, Jr.** 2006. "A Face is Exposed for AOL Searcher No. 4417749." *New York Times,* August 9.

**Baye, Michael R., John Morgan, and Patrick Scholten.** 2004. "Price Dispersion in the Small and in the Large: Evidence from an Internet Price Comparison Site." *Journal of Industrial Economics* 52(4): 463–96.

**Bhattacharjee, Sudip, Ram D. Gopal, Kaveepan Lertwachara, and James R. Marsden.** 2006. "Impact of Legal Threats on Online Music Sharing Activity: An Analysis of Music Industry Legal Actions." *Journal of Law and Economics* 49(1): 91–114.

**Carlton, Dennis W., and Judith A. Chevalier.** 2001. "Free Riding and Sales Strategies for the Internet." *Journal of Industrial Economics* 49(4): 441–61.

**Cavallo, Alberto.** 2011. "Scraped Data and Sticky Prices." January 27. http://www.mit.edu/~afc/papers/Cavallo-Scraped.pdf.

**Centola, Damon.** 2010. "The Spread of Behavior in an Online Social Network Experiment." *Science,* September 3, 329(5996): 1194–97.

**Chen, Yan, F. Maxwell Harper, Joseph Konstan, and Sherry Xin Li.** 2010. "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens." *American Economic Review* 100(4):1358–98.

**Chevalier Judith, and Austan Goolsbee.** 2003. "Measuring Prices and Price Competition Online: Amazon.com vs. BarnesandNoble.com." *Quantitative Marketing and Economics* 1(2): 203–222.

**Chiou, Lesley, and Catherine Tucker.** 2011. "Copyright, Digitization, and Aggregation." NET Institute Working Paper No. 11-18.

**Choi, Hyunyoung, and Hal Varian.** 2009. "Predicting Initial Claims for Unemployment Benefits." July 5. http://research.google.com/archive/papers/initialclaimsUS.pdf.

**comScore.** 2011. "Methodology." Information on a webpage. http://www.comscore.com/About_comScore/Methodology.

**Daniel Kahneman, and Amos Tversky.** 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47(2): 263–92.

**Edelman, Benjamin.** 2002. "The Effect of Editorial Discretion Book Promotion on Sales at Amazon.com." http://www.benedelman.org/publications/thesis-intro.pdf.

**Edelman, Benjamin, and Duncan S. Gilchrist.** 2010. "'Sponsored Links' or 'Advertisements'?: Measuring Labeling Alternatives in Internet Search Engines." Harvard Business School Working Paper 11-048.

**Edelman, Benjamin, and Ian Larkin.** 2009. "Demographics, Career Concerns or Social Comparison: Who Games SSRN Download Counts?" Harvard Business School Working Paper 09-096.

**Einav, Liran, Theresa Kuchler, Jonathan D. Levin, and Neel Sundaresan.** 2011. "Learning from Seller Experiments in Online Markets." NBER Working Paper 17385.

**Ellison, Glen, and Sara Fisher Ellison.** 2009. "Tax Sensitivity and Home State Preferences in Internet Purchasing." *American Economic Journal: Economic Policy* 1(2): 53–71.

**Freedman, Seth, and Ginger Zhe Jin.** 2011. "Learning by Doing with Asymmetric Information: Evidence from Prosper.com." NBER Working Paper 16855.

**Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant.** 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature,* February 19, Volume 457(7232): 1012–14.

**Hitsch, Gunther J., Ali Hortaçsu, and Dan Ariely.** 2010. "Matching and Sorting in Online Dating." *American Economic Review* 100(1): 130–63.

**Horton, John J., David G. Rand, and Richard J. Zeckhauser.** 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14(3): 399–425.

**Hossain, Tanjim, and John Morgan.** 2006. ". . . Plus Shipping and Handling: Revenue (Non) Equivalence in Field Experiments on eBay." *Advances in Economic Analysis & Policy* 6(2): Article 3.

**Internet Archive.** 2011. "About the Wayback Machine." Information on a webpage. http://www.archive.org/web/web.php.

**Iyengar, Raghuram, Sangman Han, and Sunil Gupta.** 2009. "Do Friends Influence Purchases in a Social Network?" HBS Marketing Unit Working Paper 09-123.

**Karagiannis, Thomas, Andre Broido, Nevil Brownlee, kc claffy, and Michalis Faloutsos.** 2004. "Is P2P Dying or Just Hiding?" Presented at the 47th Global Telecommunications Conference, Globcom 2004, Dallas TX, Nov. 29–Dec. 3.

**Katkar, Rama, and David H. Reiley.** 2006. "Public versus Secret Reserve Prices in eBay Auctions: Results from a Pokémon Field Experiment." *Advances in Economic Analysis and Policy* 6(2): Article 7.

**Lewis, Randall A., and David H. Reiley.** 2011. "Does Retail Advertising Work? Measuring the Effects of Advertising on Sales via Controlled Experiment on Yahoo!" http://www.davidreiley.com/papers/DoesRetailAdvertisingWork.pdf.

**List, John A.** 2011. "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off." *Journal of Economic Perspectives* 25(3): 3–16.

**Mason, Winter, and Siddharth Suri.** 2011. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods.* Online publication, June 30. doi: 10.3758/s13428-011-0124-6.

**Miller, Sarah.** 2011. "Information and Default in Consumer Credit Markets: Evidence from a Natural Experiment." https://netfiles.uiuc.edu/smille36/www/InformationDefault062011.pdf.

**Netcraft.** 2011. "June 2011 Web Server Survey." http://news.netcraft.com/archives/2011/06/07/june-2011-web-server-survey.html.

**Ostrovsky, Michael, and Michael Schwarz.** 2011. "Reserve Prices in Internet Advertising Auctions: A Field Experiment." Presented at the 12th ACM Conference on Electronic Commerce, San Jose, California, June.

**Reiley, David H., Sai-Ming Li, and Randall A. Lewis.** 2010. "Northern Exposure: A Field Experiment Measuring Externalities between Search Advertisements." *Proceedings of the 11th ACM Conference on Electronic Commerce* (EC-2010), pp. 297–304. ACM Digital Library.

**Resnick, Paul, Richard Zeckhauser, John Swanson, and Kate Lockwood.** 2006. "The Value of Reputation on eBay: A Controlled Experiment." *Experimental Economics* 9 (2): 79–101.

**Roth, Alvin E., and Axel Ockenfels.** 2002. "Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet." *American Economic Review* 92(4): 1093–1103.

**Saroiu, Stefan, Steven Gribble, and Henry Levy.** 2004. "Measurement and Analysis of Spyware in a University Environment." Presented at the First Symposium on Networked Systems Design and Implementation (NSDI '04), San Francisco, CA, March 29–31.

**Schechter, Stuart, Rachna Dhamija, Andy Ozment, and Ian Fischer.** 2007. "The Emperor's New Security Indicators." Presented at the IEEE Symposium on Security and Privacy, Oakland, CA, May.

**Seamans, Robert, and Feng Zhu.** 2010. "Technology Shocks in Multi-Sided Markets: The Impact of Craigslist on Local Newspapers." NET Institute Working Paper 10-11.

**Varian, Hal R.** 2010. "Computer Mediated Transactions." *American Economic Review* 100(2): 1–10.

**Vincent, Arnaud, and Margaret Armstrong.** 2010. "Predicting Break-Points in Trading Strategies with Twitter." SSRN Working Paper 1685150.

**Wu, Lynn, and Erik Brynjolfsson.** 2009. "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales." Presented at the NBER meeting on Technological Progress & Productivity Measurement, December 4.

Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 207–222

# Jonathan Levin: 2011 John Bates Clark Medalist

Liran Einav and Steve Tadelis

Jonathan Levin, the 2011 recipient of the American Economic Association's John Bates Clark Medal, has established himself as a leader in the fields of industrial organization and microeconomic theory. Jon has made important contributions in many areas: the economics of contracts and organizations; market design; markets with asymmetric information; and estimation methods for dynamic games. Jon's combination of breadth and depth is remarkable, ranging from important papers in very distinct areas such as economic theory and econometric methods to applied work that seamlessly integrates theory with data. In what follows, we will attempt to do justice not only to Jon's academic work, but also try to sketch a broader portrait of Jon's other contributions to economics as a gifted teacher, dedicated advisor, and selfless provider of public goods. Numerical references to Jon's papers cited in this essay are listed in Table 1.

## Biography

Unlike most economists, Jon's exposure to economics started not in the classroom, but at the dinner table. He was born in New Haven to Richard and Jane Levin, who both went on to prominent careers at Yale. Richard, a Professor of Economics,

■ *Liran Einav is Associate Professor of Economics, Stanford University, Stanford, California, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Steve Tadelis is Associate Professor of Economics and Public Policy, Haas School of Business, University of California—Berkeley, Berkeley, California, and Distinguished Economist, eBay Research Labs, San Jose, California. Their e-mail addresses are 〈leinav@stanford.edu〉 and 〈stadelis@haas.berkeley.edu〉.*

http://dx.doi.org/10.1257/jep.26.2.207. doi=10.1257/jep.26.2.207

**Jonathan Levin**

is now serving as President of Yale University, while Jane, whose Ph.D. is in English literature, serves as the Director of Yale's selective interdisciplinary program for freshmen in Western Civilization. Jon and his three younger siblings were debating economic policy around the dinner table before most children learned how to spell "economics." We can only imagine how engaging these family conversations were by the fact that many years later, Jon was recruited by his father to help him in his role as the co-chair of the National Academy committee on patent reform. The paper that resulted from this collaboration [7] was important in shaping the America Invents Act, which was signed into law this past September.

We were told that, during a recruiting dinner at his home, Jon's puzzle-solving skills as a high school student helped his father recruit Barry Nalebuff to Yale. This experience even inspired Barry to employ Jon as a very young research assistant soon afterwards. Barry's engaging puzzles in math and economics (a number of which were published in the early issues of this journal from 1987 to 1990) piqued Jon's interests and led to a productive research collaboration. In fact, their joint paper [1] on vote-counting schemes, published in this journal, was Jon's first academic publication.

Jon pursued undergraduate studies at Stanford University, where he chose a double-major that reflected his two diverse passions, math and English literature. This combination of majors, which is not common for economists, provided superb training for a successful career in economics: be rigorous and write well! At Stanford,

*Table 1*
**Selected Papers by Jonathan Levin**

1. “An Introduction to Vote-Counting Schemes,” (with Barry Nalebuff). 1995. *Journal of Economic Perspectives* 9(1): 3–26.
2. “An Optimal Auction for Complements.” 1997. *Games and Economic Behavior* 18(2): 176–92.
3. “Information and Competition in U.S. Forest Service Timber Auctions,” (with Susan Athey). 2001. *Journal of Political Economy* 109(2): 375–417.
4. “Information and the Market for Lemons.” 2001. *RAND Journal of Economics* 32(4): 657–66.
5. “Multilateral Contracting and the Employment Relationship.” 2002. *Quarterly Journal of Economics* 117(3): 1075–1103.
6. “Relational Incentive Contracts.” 2003. *American Economic Review* 93(3): 835–47.
7. “Patent Oppositions,” (with Richard Levin). 2003. Chap. 13 in *Economics for an Imperfect World: Essays in Honor of Joseph Stiglitz*, edited by R. Arnott, B. Greenwald, R. Kanbur, and B. Nalebuff. Cambridge: MIT Press.
8. “Profit Sharing and the Role of Professional Partnerships,” (with Steve Tadelis). 2005. *Quarterly Journal of Economics* 120(1): 131–71.
9. “Matching and Price Competition,” (with Jeremy Bulow). 2006. *American Economic Review* 96(3): 652–68.
10. “Estimating Dynamic Models of Imperfect Competition,” (with Patrick Bajari and Lanier Benkard). 2007. *Econometrica* 75(5): 1331–70.
11. “Liquidity Constraints and Imperfect Information in Subprime Lending,” (with William Adams and Liran Einav). 2009. *American Economic Review* 99(1): 49–84.
12. “Empirical Industrial Organization: A Progress Report,” (with Liran Einav). 2010. *Journal of Economic Perspectives* 24(2): 145–62.
13. “Online Advertising: Heterogeneity and Conflation in Market Design,” (with Paul Milgrom). 2010. *American Economic Review* 100(2): 603–07.
14. “Beyond Testing: Empirical Models of Insurance Markets,” (with Liran Einav and Amy Finkelstein). 2010. *Annual Review of Economics* 2(1): 311–36.
15. “Contracting for Government Services: Theory and Evidence from U.S. Cities,” (with Steve Tadelis). 2010. *Journal of Industrial Economics* 58(3): 507–41.
16. “Early Admissions at Selective Colleges,” (with Christopher Avery). 2010. *American Economic Review* 100(5): 2125–56.
17. “Comparing Open and Sealed Bid Auctions: Evidence from Timber Auctions,” (with Susan Athey and Enrique Seira). 2011. *Quarterly Journal of Economics* 126(1): 207–57.
18. “The Value of Information in Monotone Decision Problems,” (with Susan Athey). 2001. http://kuznets.fas.harvard.edu/~athey/VOI.pdf.
19. “Winning Play in Spectrum Auctions,” (with Jeremy Bulow and Paul Milgrom). 2009. NBER Working Paper 14765.
20. “The Impact of Information Technology on Consumer Lending,” (with Liran Einav and Mark Jenkins). 2012. http://www.stanford.edu/~leinav/Credit_Scoring.pdf.
21. “Contract Pricing in Consumer Credit Markets,” (with Liran Einav and Mark Jenkins). Forthcoming. *Econometrica*.
22. “Pricing and Welfare in Health Plan Choice,” (with Kate Bundorf and Neale Mahoney). Forthcoming. *American Economic Review*.
23. “Set-Asides and Subsidies in Auctions.” (with Susan Athey and Dominic Coey). 2011.
24. “The Economics of Internet Markets.” Forthcoming. In *Advances in Economics and Econometrics*, vol. 4, edited by D. Acemoglu, M. Arellano, and E. Dekel. Cambridge University Press.
25. “Designing Advanced Market Commitments for New Vaccines,” (with Michael Kremer and Christopher Snyder). 2009. Unpublished paper.
26. “Learning from Seller Experiments in Online Markets,” (with Liran Einav, Theresa Kuchler, and Neel Sundaresan). 2011. NBER Working Paper 17385.
27. “Sales Mechanisms in Online Markets: What Happened to Online Auctions?” (with Liran Einav, Chiara Farronato, and Neel Sundaresan). In progress.
28. “Sales Taxes and Internet Commerce,” (with Liran Einav, Dan Knoepfle, and Neel Sundaresan). In progress.

Jon studied intermediate microeconomics with Donald Brown, whom Jon had known as a child in New Haven. We were told that this engaging and demanding course was a turning point for Jon, persuading him that economics was his true calling.

After graduating from Stanford in 1994, Jon went on to pursue a Master's degree at Oxford University on a Fulbright scholarship. At Oxford, Jon took his first steps into auction theory under the supervision of Paul Klemperer. That early effort produced the second of Jon's publications [2]. From Oxford, Jon went on to obtain his Ph.D. from MIT, with Bengt Holmstrom, Glenn Ellison, and Jerry Hausman as his thesis advisors. Jon successfully completed his Ph.D. in 1999 after only three years, though "successfully" is somewhat of an understatement: four papers he wrote during his time at MIT were later published in the *American Economic Review* [6], *Quarterly Journal of Economics* [5], *Journal of Political Economy* [3], and *RAND Journal of Economics* [4]. Not a bad start.

After a year with a Cowles Foundation Post-Doctoral Fellowship, Jon accepted a position at Stanford in 2000, where he is now Professor of Economics and department chair. He was elected as a fellow of the Econometric Society in 2008, and is now on the editorial board of five leading journals.

## Contracts and Organizations

Long-term contractual relationships govern many economic transactions such as employer and employee, buyer and supplier, lender and borrower, and regulator and industry firms. What is striking, as Macaulay (1963) observed, is that many firms that engage in long-term contracts do not fully specify the contractual terms that govern their relationships. Instead, they rely on their ongoing relationships to fill gaps in the contract and to maintain gains from trade. In other words, the parties realize that the long-term value of their relationship can help direct behavior to be mutually beneficial.

The "relational contacts" literature has developed to study the conditions under which parties can use the threat of foregoing the value of their ongoing relationship to mitigate the hazards of one party pursuing short-term gains at the expense of the other party. In this setting, mutually beneficial trade is supported by self-enforcing contracts that do not require courts to intervene and impose sanctions. Early seminal developments by Telser (1980), Klein and Leffler (1981), and MacLeod and Malcomson (1989) laid the grounds for this prolific research agenda. However, these and most other papers study specialized contacts that are not necessarily optimal. Moreover, variations in the information structure available to the parties will influence their ability to form relational contracts, and the literature had not offered a consistent framework to address these important issues.

In [6], Jon unifies, generalizes, and extends the literature to study optimal contracts with a variety of information structures. To appreciate Jon's contribution, consider an ongoing (infinite horizon) relationship between an employer and employee. The employee's per-period effort determines per-period output, which is owned by the employer. A relational contract specifies, for each period, (enforceable) per-period unconditional wages, a discretionary bonus payment that the employer promises to pay the employee, and the effort that the employee promises to exert. The earlier literature focused on the case where effort is observed and output is determined by effort. Jon's analysis generalizes the investigation to include two central informational variations: 1) the employee's effort is unobservable and output is stochastically determined by the employee's effort; and 2) the employee observes some random shock to production costs that is unobserved by the employer and the employee chooses how much to produce. The first involves the application of relational contract analysis to a dynamic version of Holmstrom's (1979) static "hidden action" problem, while the second involves the application of relational contract analysis to a dynamic version of Mirrlees' (1971) static "hidden information" problem.

Jon's analysis provides three key insights. First, when searching for the best (optimal) relational contract that the parties can achieve, it suffices to restrict attention to "stationary contracts" where in every period the contract is essentially the same. This striking result arises because unlike standard repeated games where players rely on "continuation payoffs" (promises of future payments) to provide incentives, in this setting the parties have *two* instruments to provide incentives: bonus payments made today and promised continuation payoffs from future surplus. As Jon shows, these instruments are perfect substitutes with risk-neutral parties. Hence, any contract in which the employer provides incentives using variation in continuation payoffs can be replaced using variation in bonus payments that provide the same incentives. Second, Jon shows that there is a "dynamic enforcement" constraint that characterizes the optimal stationary contract: namely, discretionary bonus payments can be neither too small (otherwise the employee will leave the relationship) nor too large (otherwise the employer will renege and terminate the relationship). This limits what the parties can enforce in lieu of court-enforced contracts that have no such limits. Last but not least, Jon also establishes important connections between the relational contracts framework and the well-established one-shot contracts frameworks of Mirrlees (1971) and Holmstrom (1979). This paper is Jon's most cited piece, and it has become the standard reference and workhorse model for this growing literature.

In [5], Jon applies his framework of relational contracts to address an important practical concern of firms: the design of workforce compensation schemes and incentives contracts. He opens the paper with some classic questions: "Why do firms hesitate to cut pay or lay off workers in economic downturns? Why do some firms promote the idea of lifelong employee commitment, while others explicitly back away from such policies or hire temporary workers with low tenure expectations?" Jon refers to studies and cases that suggest that many firms have encountered severe problems, or believe they would encounter problems, if they would change employees' benefit packages, or if they deviate from expected wage increases or perhaps even attempt to lower wages. Jon approaches the problem as one in which firms form contractual relationships with their workforce as a whole, rather than have a collection of individual contractual relationships. Jon compares a firm's commitments to its workforce as a whole, which he refers to as "multilateral relational contracting," with commitments to individuals or smaller groups of employees, called "bilateral relational contracts." Using this approach, Jon reveals an important tradeoff. Multilateral contracting improves the firm's ability to offer implicit self-enforcing commitments, which in turn improves incentives. However, when these implicit commitments cover the whole workforce, it becomes difficult to adjust to changes in the business environment. In contrast, bilateral contracts facilitate workforce changes, at the cost of restricting the set of self-enforcing contracts and hence reducing incentives. Jon's analysis sheds light on the use of relative performance evaluation and the adoption of multitiered workforces, in which different well-defined groups are employed using distinct relational contracts. This paper showcases one of Jon's trademarks: begin with an interesting and important real-world question, offer a rigorous and elegant model to shed light on the question, and then use the model's insights to improve our understanding of the economic reasons for actual practices.

In addition to the choice of explicit and implicit contracts, a firm's choice of organizational form will also influence the incentives and productivity of its employees. The corporate form of organization dominates some industries, such as manufacturing, technology, and many service industries, while partnerships have been prominent in human capital–intensive professional services such as law, accounting, investment banking, management consulting, advertising, and medicine. Why do these patterns persist, and what can explain them? In [8], Jon and Steve Tadelis take the defining feature of a partnership to be the redistribution of profits among its partners, whereas with corporations, employees earn wages and owners are the residual claimants of profits. They show that profit-sharing leads individuals to be particularly selective as to whom they take on as partners, resulting in a higher quality of service than that of a corporation. The intuition is straightforward: if a group commits to split its output, then it will not take on an additional partner if the marginal output of that individual is less than the current average output. In contrast, a corporation will hire an employee as long as the marginal output of that employee exceeds the market wage. As a result, if clients are disadvantaged in assessing service quality, then partnerships tend to be a preferable mode of organization relative to a profit-maximizing corporation because they create an internal incentive to select high-quality agents. Thus, partnerships will emerge when two features are prominent: 1) human capital is responsible for product quality; and 2) clients are at a disadvantage relative to firms in assessing the ability of the firm's workforce. These two conditions are typical of professional services, but not of manufacturing or technology industries where partnerships are quite unusual, thus explaining the patterns that motivated the research question.

In another joint paper [15], Jon and Steve consider the fundamental question in the economics of organization as posed by Coase (1937): When should a firm choose to do part of the production process within the firm, and when should it buy inputs from the market? Their paper offers a model that combines standard agency theory with transaction cost economics in the spirit of Williamson (1985), in which some processes are more complex and harder to contract for than others. The model highlights the make-or-buy trade-off along two dimensions. The choice to buy ensures productive efficiency, which is obtained by contracting with high-powered incentives, yet the organization must bear the costs of contract administration. The choice to make rather than buy unloads the burdens of administering contracts, but involves low-powered incentives that reduce productive efficiency. The application of their approach is less conventional in the organizational economics literature because it considers local governments as the "firms" with the choice of whether to provide services with their own employees or by contracting with private or public sector providers. The empirical analysis suggests that economic efficiency concerns seem to play an important role in the decisions of local governments to contract for services, but not surprisingly, politics plays a role too.

## Market Design

Market design scholars study how market rules and institutions affect outcomes, such as the efficient allocation of resources or the profits of the market designer. There are two central applications of market design: auction markets, and the design and engineering of matching markets. Matching markets describe situations in which one side of the market must be matched with another side, but for institutional reasons prices are not allowed to play a defining role. Jon has been active in both areas, integrating economic theory, novel empirical methods, and data to obtain interesting new insights.

As mentioned earlier, Jon's interest in auctions dates back to his Master's thesis at Oxford University where he studied under the supervision of Paul Klemperer. In [2], Jon considers the optimal selling mechanism for complementary items, and he outlines conditions for which the bundling of items in a standard auction is optimal, yet shows that in general bundling the items does not maximize revenue. Thus, a tension can arise between efficient allocation of items across buyers and the seller's profit-maximization objective.

Jon continued his work on market design as a student at MIT, where he and Susan Athey (who taught there at the time) started a prolific partnership. They studied competitive bidding for federally owned timber, with the goal of understanding how different auction rules used by the government have affected competition. These papers combine rigorous theory and sophisticated estimation techniques to shed light on important government policy decisions. It was also one of the first empirical studies to focus on auctions with "scoring rules," which are common in government procurement decisions. In [3], Jon and Susan investigate how firms can strategically exploit the scoring rule often used in Forest Service auctions. In these auctions, government agencies offer for sale tracts of forest where more than one kind of timber grows. The government agency first publishes its own estimate of how many trees of each type are present in the tract that is up for bidding. Bidding firms can then form their own estimates, after which each firm lists a price per type of tree, which constitutes the firm's bid. The highest bidder is the firm for which the simple product of bid-prices and government-estimated quantities is the highest. The paper highlights the fact that firms can strategically skew their bids without affecting their total price, by raising bids on types of trees for which the government overestimated quantities while lowering prices on those that were underestimated.[1] By developing an equilibrium model of scoring auctions, Jon and Susan use a combination of bidding behavior and performance data after the auction to document strategically skewed bidding by participating firms. The paper highlights the consequences of skewed bidding for both the allocation of tracts across firms and the resulting revenues to the government.

In [17], Jon, Susan, and their student Enrique Seira study the relative performance of oral and sealed bid auctions, again using Forest Service auctions. The most famous result in auction theory, the "revenue equivalence theorem," states that under certain specific conditions, there should be no difference between these auction designs in terms of revenue, allocation, or entry. When these specific conditions fail, however, then different auction formats will result in different outcomes, affecting competition, revenues, and whether resources are allocated efficiently. The paper uses data on timber auctions in the Montana–Idaho border area, in which both oral and sealed bid auctions have been used. An advantage of the setting is that the choice of auction format was driven by exogenous factors and occasionally by explicit randomization. The paper shows that, consistent with some recent development in auction theory, sealed-bid auctions favor "weak" participants (those with a lower expected value from the auctioned object) and that prices in these auctions are higher than those in oral auctions. The analysis suggests that a lack of aggressive competition among the few large bidders could explain the departures from more competitive bidding in oral auctions. In addition to its applied contribution to the design of timber auctions, the paper brings together an impressive collection of methods that are needed to capture the deviations of timber auctions from the standard textbook auction setting, such as bidder asymmetries, unobserved heterogeneity, and a joint decision of bidding and participation.

In a follow-up paper [23], Jon, Susan, and their student Dominic Coey use an empirical approach similar to the one used in [17] to study alternative methods

[1] For example, imagine that the government estimates 500 Douglas fir trees ($D$) and 400 Western Hemlock trees ($W$). Imagine that Firm 1 bids \$90 for each $D$ tree and \$80 for each $W$ tree, while Firm 2 bids \$80 and \$100. The government uses its estimates, together with the firms' bids, to create a total bid per firm. The total bid of Firm 1 is then (\$90 × 500) + (\$80 × 400) = \$77,000 while the total bid of Firm 2 is (\$80 × 500) + (\$100 × 400) = \$80,000, and Firm 2 is the winning bidder. Now, imagine that Firm 2 sends a surveyor to count the trees and as a result estimates that there are in fact 600 $D$ trees and only 300 $W$ trees. Instead of bidding \$80 and \$100 for the trees, it can change its bids to \$40 and \$150, respectively. Its total bid is still \$80,000, so it will win the auction, but if its estimates were correct then Firm 2 would only end up paying (\$40 × 600) + (\$150 × 300) = \$69,000, compared to paying (\$80 × 600) + (\$100 × 300) = \$78,000 if it kept the original bids of \$80 and \$100 (which resulted in the same total bid). In equilibrium, both firms have an incentive to produce accurate estimates and to use these strategically.

for a government to achieve distributional goals in an auction. In particular, they compare the use of targeted subsidies, where some bidders receive a subsidy that makes them more competitive, to setting aside certain contracts for those disadvantaged bidders, as is standard practice in federal procurement and natural resource sales. Economic theory does not provide clear guidance on which approach should be preferred, but their empirical results suggest that relative to setting aside some contracts, subsidies increase both revenues and efficiency.

Many other government agencies use auctions to allocate contracts and resources. The auctions used by the United States and other countries to sell spectrum rights to telecommunications firms in the mid-1990s are perhaps the most famous among economists. Indeed, these auctions sparked a revival of interest in auction theory and practice. The involvement of prominent academics in the design of these auctions, and their general success, has helped demonstrate the effectiveness of modern economic theory. Nevertheless, there is little published research on how these large and complex auctions actually work. In [19], Jon, Jeremy Bulow, and Paul Milgrom point out that the "activity rules" in these auctions—rules that govern the eligibility of bidders to continue bidding in subsequent auction rounds as a function of their earlier bids—create the potential for substantial pricing anomalies. More importantly, they show that aggregate prices appear to be driven by bidder budgets rather than by the bidder valuations emphasized in standard auction theory papers. This insight stresses the shortcoming of applying standard models to some instances of bidding behavior and offers motivation for the study of bidders with hard budget constraints. The paper then provides insights into the ways in which sophisticated bidders can take advantage of pricing anomalies and discusses one particular high-stakes case of a spectrum auction where this happened.

Turning to the design of matching markets, two of the best-known examples are the market for medical residents and the market for college admissions. Again, these are settings where the price mechanism is ruled out, leading the market designer to solicit the preferences of the market participants in order to make efficient matches.

In [9], Jon and Jeremy Bulow consider the National Residency Matching Program, which uses a market design by Al Roth (Roth and Peranson 1999) that adapts the famous Gale–Shapley (1962) algorithm to assign medical school graduates to hospital residency positions. Motivated by an antitrust case claiming that the match depressed resident wages by preventing individualized salary negotiations, Jon and Jeremy analyze matching markets in which firms are restricted from varying their salary offers across individuals, so that matching takes place with each firm offering a fixed salary. They show that in such a market the equilibrium assignment is nearly efficient but wages are low and significantly compressed, while firm profits exceed those in any competitive equilibrium. This result suggests that the ability to make discriminatory offers can have a significant impact on the distribution of surplus in matching markets.

In [16], Jon and Chris Avery analyze the market for college admissions. They focus on "early admissions," which has become a common feature of this market in the last two decades. Many top colleges offer applicants the opportunity to choose one college to which they apply early, several months before the regular application deadline. The paper combines theory and supporting evidence to explain the importance of early admissions in matching between applicants who are enthusiastic about certain colleges, and the colleges that wish to admit enthusiastic applicants. The paper argues that the early admissions policy is a market institution that allows applicants to communicate their enthusiasm about their preferred college in a credible manner. In particular, the exclusionary nature of the policy generates a real opportunity cost, so an applicant who does not apply for early decision to a college signals less enthusiasm about being admitted there.

Finally, Jon's work with Michael Kremer and Chris Snyder [25] offers one more illustration of Jon's ability to take insights from economic theory and apply them in practice. They analyze the design of Advanced Market Commitments, which are targeted subsidy programs to encourage the development and introduction of vaccines and drugs for low-income countries. The basic concept has gained substantial currency in the development community. Unfortunately, the design that was envisioned involves committing to a subsidized price for spot market purchases until funds run out, and thus suffers from a serious incentive problem. Unless supply conditions are competitive, which they typically are not, firms have little incentive to invest in capacity to serve the full market because selling larger annual quantities does not increase the overall level of funds in the advanced market commitment program. The paper shows that this problem can be overcome by tying subsidies to long-term supply commitments, with contracts allocated through a competitive bidding mechanism.

## Markets with Asymmetric Information: Subprime Lending and Health Insurance

In recent years, especially after the economic downturn that started in late 2007, there has been much interest in exploring subprime lending and its consequences. In a series of papers with Liran Einav, and former students Will Adams and Mark Jenkins, Jon examines the organization of subprime automobile-loan markets. Some good fortune was in play here, perhaps confirming Louis Pasteur's "chance favors the prepared mind." This research originated in 2005, a few years before the term "subprime" became a household name—albeit in reference to the market for home mortgages rather than auto loans.

The papers from this project provide a revealing window into low-income, high-risk credit markets. The data, obtained from a large auto sales and financing company, shows that one-third of loan applicants have no bank account, the modal (annual) interest rate is almost 30 percent, and more than 60 percent of the loans end in default. The papers seek to understand the high demand for these loans, the constraints faced by borrowers, and the informational imperfections inherent in high-risk consumer lending. The work also develops a range of empirical approaches for studying contracting markets with moral hazard and adverse selection.

In [11], Jon, Liran, and Will describe several striking facts about subprime lending. First, the individuals in the subprime population appear highly liquidity constrained. Loan applicants are extremely sensitive to the required down payment, far more sensitive than they are to changes in car prices. Moreover, purchasing activity spikes dramatically in February and March, when consumers become eligible for early tax rebates. Indeed, the stark seasonality in demand can be linked to individuals eligible for large rebates under the federal Earned Income Tax Credit program. Second, lenders in this market for high-risk consumer credit face serious problems of adverse selection and moral hazard. Given a choice, high-risk borrowers tend to self-select into the largest loans, and default rates for all borrowers are very sensitive to the level of monthly payments. Distinguishing moral hazard from adverse selection is challenging because, as is typical in empirical studies of asymmetric information, these two distinct problems give rise to similar empirical patterns, which in this context is the positive correlation between default rates and loan amounts. Central features of [11] are the use of plausible exogenous variation in the loan amount to identify moral hazard separately, and the finding that it plays a central role in causing defaults.

In [21], Jon, Liran, and Mark develop a more complete model of demand that allows for a joint analysis of the demand for cars, the down payment decision, and the repayment outcomes. The paper then analyzes optimal contract design and quantifies the value of credit-scoring information and risk-based pricing in such a context. This paper is one of the first to incorporate supply-side restrictions in markets with moral hazard and adverse selection, thus illustrating how standard techniques from industrial organization can be extended to estimate demand systems for credit or insurance contracts. Although some extrapolation is needed to leap from automobile loans to home mortgages, many features documented in this paper seem to have played a central role in the recent subprime mortgage market meltdown, in which lax down payment requirements allowed borrowers to become highly leveraged and therefore vulnerable in the face of declining house prices and underlying income or liquidity risk.

In [20], the three authors complement the analysis of the value of risk-based financing by using data from before and after the auto loan company switched from a traditional lending model, with significant discretion at the dealership level, to a modern and highly automated system that used computerized risk models. In addition to the increase in profits through an obvious channel—better information allows more efficient screening of bad risks—the paper documents a surprisingly large increase in profits driven by the ability to customize larger and more expensive cars to better risks, and describes interesting patterns across dealerships within the organization.

Many of the informational and behavioral problems in credit markets are also relevant in insurance markets. In [22], Jon, Kate Bundorf, and former student Neale Mahoney study the pricing structure in employer-sponsored health insurance markets. The nature and regulation of employer-provided health insurance leads to consumers often facing the same prices even though different plans may have a cost advantage in serving different populations. The paper develops and estimates an equilibrium model of demand and supply for differentiated products to assess the sorting inefficiency caused by this lack of variation in prices. The results suggest that plans do indeed have cost structures that favor particular sorting patterns—in particular, integrated delivery systems, such as Kaiser, seem to have a sizable cost advantage in serving those with poor health status. Yet the amount of sorting inefficiency caused by uniform pricing is limited, because demand for health insurance is relatively inelastic.

Despite the fact that the papers described in this section have been written during the more recent part of Jon's career, it is obvious that markets with asymmetric information have been on his mind for many years. In [4], which was part of Jon's Ph.D. thesis, he studies the relationship between information and trade in adverse selection markets. The paper uses a series of examples to show how changes in the degree of asymmetric information can have counterintuitive effects on the equilibrium quantity of trade. The paper is also related to another early paper in information economics [18], in which Jon and Susan Athey develop a more general approach to comparing information structures for Bayesian decisions by using monotone comparative statics methods to derive notions of informativeness for different kinds of economic decision problems.

## Estimation of Dynamic Games

Many problems in industrial organization revolve around industry dynamics: When does market leadership persist? What is the relationship between innovation and market power? How large are sunk costs of entry? What are the trade-offs between short-run pricing and long-run investment? These questions are often difficult to address using static models of competition, yet the use of dynamic models introduces a host of econometric and computational challenges. One of Jon's most cited papers [10], which is joint with Patrick Bajari and Lanier Benkard, develops a computationally feasible method for estimating dynamic models of imperfect competition. It has become the standard method that economists (mostly in industrial organization) use for empirical estimation of dynamic games.

This approach, which extends earlier influential work by Hotz and Miller (1993) on the estimation of dynamic decision problems, is one of several "two-steps approaches" that were proposed recently in order to ease the computational limitations associated with the nested fixed point approach attributed to Rust (1987). The basic idea is to first (flexibly) estimate the Markov decision rule, or "policy function." The policy function can be used to recover the agent's (expected) value function, which is done through forward simulations. Then, in the second step, one can recover the primitives of each agent's payoff function by choosing the parameters that best rationalize the observed behavior. Jon, Pat, and Lanier suggest doing this by comparing the computed value using the policy function estimated in the first stage and the value implied by alternative suboptimal policies. The parameters are chosen by minimizing the cases where the alternative is predicted to be better than the actual policy.

An important advantage of this approach is that it is never necessary to calculate the equilibrium of the dynamic game, a property that significantly reduces the computational burden. It is both straightforward to implement and also allows users of this approach to apply it for continuous controls (rather than only discrete controls, as in other approaches). In the three years since its publication, the method has become the leading approach to estimate dynamic models of imperfect competition. Its computational ease has made such questions more feasible to study, and the number of applications that use it is rapidly growing. As examples of recent papers that have attracted significant attention, Ryan (forthcoming) uses this approach to study the dynamic implications of environmental regulation of cement plants, Holmes (2011) uses it to study the rollout of Wal-Mart stores, and Sweeting (2011) uses it to analyze the product positioning of radio stations.

## Final Remarks

We have tried to summarize Jon's contributions and display the remarkable depth and breadth of his academic work. We have neglected to mention some exciting work in the pipeline, which is establishing him as a leader of a rapidy growing new area of study: the economics of Internet markets. In [24], Jon surveys what we know and what we don't about the economics of the Internet. A glimpse of Jon's future publications in this area is evident in a series of new papers. In [13], Jon and Paul Milgrom study Internet advertising. In three working papers, [26], [27], and [28], Jon, with Liran Einav, Neel Sundaresan, and current students Chiara Farronato, Dan Knoepfle, and Theresa Kuchler, analyze new data from eBay to investigate seller experimentation on the Internet, the effect of online sales taxes, and the tradeoff between auctions and fixed prices in selling items online.

This new direction reflects one of the most admirable attributes of Jon's approach to research. He works on what interests him and consequently is having fun. Perhaps best captured in Jon's own words: "I think one of the nice things about being an academic is, as long as you keep an open mind, interesting questions come up and you can start to pursue them and see where they go. That's always been the way that I've done research in the past, and I expect that's how I'll continue to do it in the future" (Nguyen 2011). Some of Jon's open-minded views about research can be found in an article about current approaches in the field of industrial organization from two years ago in this journal [12].

Although the Clark Medal is about academic and professional contributions, we would be remiss if we did not add a more personal perspective on Jon. We both have had the pleasure of being Jon's colleagues and coauthors, and we share the honor of being his friends. We stress this friendship because beyond publishing brilliant and influential work, Jon stands out for his personality, which can be summarized simply as being a real mensch.[2] It is for these reasons that we are extremely pleased to celebrate Jon's work and achievements. We willingly accept the risk of being perceived as overly adulating, knowing well that if anything, we have failed to truly convey Jon's qualities. Jon's demeanor is always that of respect and encouragement, and his genuine interest in others is apparent from the long list of students that Jon has mentored since starting his career at Stanford. Jon always gives more than he takes, and his colleagues benefit frequently from his generous professional advice.

Even as a professor, Jon does much more than just work. Sports like tennis, baseball, and hockey provided balance for Jon in his youth, and he has always been an active outdoorsman, pursuing hiking, mountain climbing, and kayaking from a young age. On one trip in 2002 with a close friend, Jon hiked up Mount Whitney, the tallest peak in the lower 48 states, with a kayak strapped to his back, before paddling down through the class V headwaters of the Kern River. This would make any proficient kayaker more than proud, but as the former President of the Stanford Kayaking Club, it was just another day in the water.

Jon and his wife Amy, who as a physician has her own demanding schedule, are together raising a lovely family of three bright and energetic children: Madeline, Benjamin, and Noah. The fact that Jon has achieved so much while being a devoted husband and father is testimony to his productivity and extraordinary work habits. Jon's parental responsibilities mean that he must be more cautious, and as such, has taken a break from the extreme outdoor activities he used to pursue. You are therefore unlikely to see him with a kayak on his back climbing a dangerous trail; his water adventures these days are taking place next to the toddlers' pool.

■ *We thank Richard Levin and Jon Levin for very helpful comments and background information.*

[2] The word "mensch," borrowed widely from Yiddish, informally means a person of integrity and honor; one that has admirable characteristics, such as fortitude and firmness of purpose. As Steve Levitt, a former Clark medalist himself, noted on his *Freakonomics* blog post (2011): "When it comes to Jon Levin I cannot remember anyone saying anything negative about him. (Of course, now that he has the Clark Medal that will likely change in a hurry!)".

## References

**Coase, Ronald H.** 1937. "The Nature of the Firm." *Economica* 4(16): 386–405.

**Gale, David, and Lloyd S. Shapley.** 1962. "College Admissions and the Stability of Marriage." *American Mathematical Monthly* 69(1): 9–14.

**Holmes, Thomas J.** 2011. "The Diffusion of Wal-Mart and Economies of Density." *Econometrica* 79(1): 253–302.

**Holmstrom, Bengt.** 1979. "Moral Hazard and Observability." *Bell Journal of Economics* 10(1): 74–91.

**Hotz, V. Joseph, and Robert A. Miller.** 1993. "Conditional Choice Probabilities and the Estimation of Dynamic Models." *Review of Economic Studies* 60(3): 497–529.

**Klein, Benjamin, and Keith Leffler.** 1981. "The Role of Market Forces in Assuring Contractual Performance." *Journal of Political Economy* 89(4): 615–41.

**Levitt, Steven D.** 2011. "Jonathan Levin: The Most Recent John Bates Clark Medal Winner." *Freakonomics* blog. June 1. http://www.freakonomics.com/2011/06/01/jonathan-levin-the-most-recent-john-bates-clark-medal-winner/.

**Macaulay, Stewart.** 1963. "Non-contractual Relations in Business: A Preliminary Study." *American Sociological Review* 28(1): 55–67.

**MacLeod, W. Bentley, and James M. Malcomson.** 1989. "Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment." *Econometrica* 57(2): 447–80.

**Mirrlees, James A.** 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies* 38(114): 175–208.

**Nguyen, Ivy.** 2011. "Professor Levin Wins Clark Medal for Econ Work." *Stanford Daily*, April, 19. http://archive.stanforddaily.com/?p=1047664.

**Roth, Alvin E., and Elliott Peranson.** 1999. "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design." *American Economic Review* 89(4): 748–80.

**Rust, John.** 1987. "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher." *Econometrica* 55(5): 999–1033.

**Ryan, Stephen P.** Forthcoming. "The Costs of Environmental Regulation in a Concentrated Industry." *Econometrica*.

**Sweeting, Andrew.** 2011. "Dynamic Product Positioning in Differentiated Product Industries: The Effect of Fees for Musical Performance Rights on the Commercial Radio Industry." http://public.econ.duke.edu/~atsweet/SWEETING_formatsjan11.pdf.

**Telser, Lester G.** 1980. "A Theory of Self-Enforcing Agreements." *Journal of Business* 53(1): 27–44.

**Williamson, Oliver E.** 1985. *The Economic Institutions of Capitalism.* Free Press: New York, NY.

Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 223–236

# Retrospectives
# The Introduction of the Cobb–Douglas Regression

Jeff Biddle

This feature addresses the history of economic terms and ideas. The hope is to deepen the workaday dialogue of economists, while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please write to Joseph Persky of the University of Illinois at Chicago at ⟨jpersky@uic.edu⟩.

## Introduction

At the 1927 meetings of the American Economic Association, Paul Douglas presented a paper entitled "A Theory of Production," which he had coauthored with Charles Cobb. The paper proposed the now familiar Cobb–Douglas function as a mathematical representation of the relationship between capital, labor, and output. The paper's innovation, however, was not the function itself, which had originally been proposed by Knut Wicksell, but the use of the function as the basis of a statistical procedure for estimating the relationship between inputs and output.[1] The paper's least squares regression of the log of the output-to-capital ratio in manufacturing on the log of the labor-to-capital ratio—the first Cobb–Douglas

[1] Wicksell proposed the function as a hypothetical representation of the relationship between inputs and output in 1896, and continued to use it in subsequent works to help illustrate theoretical propositions (Samuelson 1979). It was largely through Douglas's empirical studies, however, that the function became familiar to economists, and thus it came to be known as the Cobb–Douglas function. While the present essay deals with the early history of the function as a basis for statistical estimation of production relationships, the function also plays a role in the interesting and related history of mathematical modeling in economic theory.

■ *Jeff Biddle is Professor of Economics, Michigan State University East Lansing, Michigan. His e-mail address is ⟨biddle@msu.edu⟩.*

http://dx.doi.org/10.1257/jep.26.2.223 doi=10.1257/jep.26.2.223

regression—was a realization of Douglas's innovative vision that a stable relationship between empirical measures of inputs and outputs could be discovered through statistical analysis, and that this stable relationship could cast light on important questions of economic theory and policy.

This essay provides an account of the introduction of the Cobb–Douglas regression: its roots in Douglas's own work and in trends in economics in the 1920s, its initial application to time series data in the 1927 paper and Douglas's 1934 book *The Theory of Wages*, and the early reactions of economists to this new empirical tool.

## Paul Douglas and the Origin of the Cobb–Douglas Regression

Paul H. Douglas received his Ph.D. in economics from Columbia University in 1920. He had begun his graduate education at Columbia in 1913 and had taken his first college teaching post in 1915. In 1920, he accepted a position at the University of Chicago, where he would remain on the faculty until 1948. Douglas was a prolific researcher and began in the late teens to publish a steady stream of articles and books, usually on topics related to labor legislation and working-class living standards. In 1921, he entered an ongoing debate on the trend in real wages in the U.S. since 1890 (Douglas and Lamberson 1921), and in 1924 started work on *Real Wages in the United States, 1890–1926*, a statistical exploration of recent trends in wages, prices, employment, and unemployment rates (Douglas 1930). As Douglas assembled this evidence, he was also developing a theoretical framework through which to interpret it. In 1926 he submitted a "treatise on the theory of wages" to a competition sponsored by the clothing manufacturer Hart, Schaffner, and Marx, and was awarded the $5,000 first prize. The prize-winning manuscript, which included "a more or less original explanation of general wages drawn in terms of relative elasticities of supply" and "the theory of occupational and geographical differences in wage rates," was too long to be published, and Douglas agreed to distill it into a book. Seven years passed before this book appeared under the title *The Theory of Wages* (Douglas 1934, p. xi). It was substantially altered from its 1926 form, and at its core was the Cobb–Douglas regression.

Douglas (1971, pp. 46–47) recounted the "origin story" of the Cobb–Douglas regression in several places, including this version in his autobiography:

> One spring day in 1927, while lecturing at Amherst, I charted on a logarithmic scale three variables I had laboriously compiled for American manufacturing for the years 1899 to 1922: an index of total fixed capital corrected for the change in the cost of capital goods (C), an index of the total number of wage earners employed in manufacturing (L), and an index of physical production (P). I noticed that the index of production lay between those for capital and labor and that it was from one third to one quarter of the relative distance between the lower index of labor and the higher index of capital. After consulting with my friend Charles W. Cobb, the mathematician, we

> decided to try to find on the basis of these observations the relative contributions which each of the two factors of production, labor and capital, had upon production itself. We chose the Euler formula of a simple homogeneous function of the first degree, which that remarkable Englishman Philip Wicksteed had developed some years before ($P = bL^k C^{1-k}$). We found the values of $k$ and $1 - k$ by the method of least squares to be .75 and .25, and that $b$ was merely 1.01.[2]

Thus, Cobb and Douglas estimated $k$ in their hypothesized relationship $P = bL^k C^{1-k}$ by using the indexes Douglas had constructed to fit the linear regression $\mathrm{Log}(P/C) = b + k\,\mathrm{Log}(L/C)$.[3] They then plugged annual values of $C$ and $L$, along with the estimated values of $b$ and $k$, into their nonlinear "Euler formula" and calculated a series of predicted or "theoretical" values for $P$, denoted $P'$. Douglas was encouraged by the high correlation between actual $P$ and predicted $P'$, as well as the fact that the estimate for the share of manufacturing value added represented by wages and salaries over the period 1909–1918 from the National Bureau of Economic Research was almost identical to the estimate of $k$. The first public presentation of this research was the above-mentioned paper, "A Theory of Production," which appeared a few months later in the *American Economic Review* (Cobb and Douglas 1928). When Douglas published *The Theory of Wages* six years later, it included as a central feature a description of the statistical data, methods, and results from "A Theory of Production," accompanied by accounts of results of estimating the Cobb–Douglas regression with time series data from Massachusetts and New South Wales.

In the meantime, Cobb (1930) had published a paper on his own using the Massachusetts data and an estimation method designed to ameliorate problems created by measurement error in the data. [4] Cobb obtained bizarre results with this method, and was unable to make any sense of them. He never again published work involving the Cobb–Douglas regression. Through the years, however, Douglas would consistently give Cobb half of the credit for establishing the research program.

## The Cobb–Douglas Regression and Theories of Production and Distribution

In the 1920s, a "theory of production" was understood to be an explanation of the determinants of the level of output. The amount produced, it was generally agreed, depended upon the level of technological knowledge and the quantities

[2] Similar accounts can be found in Douglas (1948, p. 6) and Douglas (1976, p. 904). Samuelson (1979, p. 926) discusses Douglas's apparent confusion between the work of Wicksteed and Wicksell.

[3] This specification embodied the restriction that the sum of the coefficients of capital and labor equaled one. Douglas would relax this restriction in his later work with the function.

[4] Cobb's estimation method was "diagonal mean regression," a method proposed by Ragnar Frisch for estimating linear relationships between variables when all variables are measured with error. See Cobb (1939) for details on the technique.

of the factors of production employed. A theory of production offered an account of the forces making for changes in these determinants of output, and/or details about the quantitative relationships between inputs and output—for example, hypotheses about the circumstances under which the "law of diminishing return" was applicable. A theory of distribution, on the other hand, explained the determination of the division of output between various members of society. Classical economists had theorized about distribution in terms of the shares of the product received by three social classes—land owners, laborers, and capitalists—who controlled the three productive factors of land, labor, and capital, with the principles explaining the share going to land owners being distinct from those governing the shares received by laborers or capitalists. Marx's version of distribution theory ran in terms of two classes—workers and capitalists—describing the ways in which the social and economic institutions of capitalism allowed capitalists to expropriate much of the output attributable to the efforts of the working class. During the late 1800s, a number of economists introduced theories of distribution based on the now-standard principle that in a market system, the payment received by the owner of any factor of production was determined by the marginal productivity of that factor. Some of these marginal productivity theorists presented their ideas in terms of the classical trio of factors, while others rejected the relevance of those categories, preferring to emphasize the commonality, from the point of view of distribution theory, of any tool, substance, or service that could contribute to the production of final goods. By the 1920s, the economics profession displayed no consensus regarding the appropriate approach to theorizing about distribution, with approaches based on the marginal productivity principle competing with those rooted in the classical tradition and those emphasizing the ways in which social and economic institutions influenced the bargaining power of various groups. Among the important points of disagreement between advocates of the various versions of distribution theory were the extent to which distribution was a function of malleable human institutions versus relatively unchangeable aspects of human nature and the physical world, and the extent to which policies intended to alter distributive shares would influence the overall level of output.

As is clear from the title of the 1927 paper, the Cobb–Douglas regression was first presented as a contribution to production theory. Douglas opened the paper with a list of questions that could be addressed if an empirical relationship between capital, labor, and output could be discovered, including whether the increase in output apparent in the data was "purely fortuitous, whether it was primarily caused by technique, and the degree if any, to which it responded to changes in the quantity of labor and capital," and whether his proposed statistical procedure might provide "an historical approach to the theories of decreasing imputed productivity (diminishing increment to the total product)" that would open the way towards "further attempts to secure quantitative approximations to these tendencies, if indeed there should turn out to be historical validity to them" (Cobb and Douglas 1928, p. 139). The paper made no explicit reference to theories of distribution. There was a mention of the question of whether the "processes of distribution are modeled at all closely upon those of the production of value" but no explicit discussion of the link between the two provided by marginal productivity theory.

By contrast, in *The Theory of Wages* the general discussion of the empirical estimation of production relationships was embedded in a detailed explication of the marginal productivity theory and a defense of that theory as a framework for inductive study of production and distribution. The estimated elasticities of curves of marginal productivity, which as of 1934 Douglas seemed to regard as the most important quantities revealed by his innovative statistical analysis, could, in light of the marginal productivity theory, also be regarded as elasticities of aggregate demand curves for capital and labor. In the 1934 volume, there was no question as to the theoretical framework motivating Douglas's numerous comparisons between estimates of the value of labor's marginal product, derived from his regression, and measures of real wages or labor's share of the value of output.

## The Cobb–Douglas Regression and Empirical Economics in the 1920s

One can see in Douglas's innovation of 1927 a blending of several characteristics of the empirical economics literature of the 1920s. First, it reflected the period's emphasis on the importance of creating reliable statistical measures of economic activity. Because government programs for collecting economic statistics were still in their infancy, one of the more important skills for empirically oriented economists was the ability to construct, from the fragmentary statistical evidence available on a phenomenon, a credible quantitative account of that phenomenon. Thus, the researcher had to locate the relevant data sources, to extrapolate from time periods or sectors for which data were relatively complete to time periods or sectors in which they were scant, and to defend or assess the likely accuracy of the results using logic, implicit theorizing, and various consistency checks across data from different sources. Researchers also needed to persuade readers not only that the steps taken to produce the estimates were the most reasonable ones under the circumstances, but that the resulting statistical picture, with all its shortcomings, was accurate enough to be useful. This type of work was a crucial prerequisite to the estimation of the Cobb–Douglas regression in 1927, with ten of the 24 pages of "A Theory of Production" devoted to explaining and defending Douglas's methods of constructing time series for fixed capital and labor. Judged by the standards of the time, Douglas's construction of these series by itself would have been considered an important contribution to empirical economics.

Second, Douglas's use of least squares regression, as well as correlation coefficients and indexes, placed him at the cutting edge of statistical practice in economics as of 1927. Prior to World War I, with the exception of the work of a few pioneers such as Wesley Mitchell, Warren Persons, and Irving Fisher, most empirical economic research simply presented raw numbers or percentage shares, and included no derived statistical measures such as means, standard deviations, or index numbers. This began to change during the 1920s, and by the end of that decade, the well-trained empirical economist understood basic statistical theory and applied it in constructing index numbers, tabulating frequency distributions, and calculating summary statistics.[5] Douglas's own development as an economic statistician paralleled these changes in what represented good statistical practice for economists, as he moved from reporting numbers and percentages in tables and text (Douglas 1918, 1919), to calculating means and measures of average deviation to illustrate relevant points (Douglas 1930), to the use of correlation coefficients and linear regression in "A Theory of Production."

Finally, in attempting to statistically estimate the key functional relationships underlying marginal productivity theory, Douglas was expanding the boundaries of a newly emerging research program in empirical economics. During the 1920s and early 1930s, a growing literature sought to apply regression techniques to estimate the real world counterparts of theoretical supply and demand curves. As Morgan (1990) has shown, this work played an important role in shaping the approach to combining statistical methods and economic theory that would become the standard econometric practice in the later decades of the twentieth century, but in the 1920s it was still quite esoteric. Douglas (1934, p. xii), however, saw it as the wave of the future, and he explicitly linked his own research to it:

> It has long seemed to me that the inductive, statistical, and quasi-mathematical method must be used if we are ever to make economics a truly fruitful and progressive science. The neoclassical school has constructed a valuable theoretical scaffolding according to which the value of commodities and the rates of return to land, labor, and capital are fixed at the intersection of the various supply and the demand curves. This is a beginning but only a beginning. For in order to make the analysis precise, to forecast, and to detect interactions in economic society it is plainly necessary to determine the slopes of the demand and supply curves. . . . An excellent beginning has been made in this direction by such scholars as Henry L. Moore, Schultz, Ezekiel, Bean, Working, and Marschak . . .
>
> There is a need for a similar approach to the problems of distribution. We need to know whether the assumed curves of diminishing incremental productivity are merely imaginative myths or whether they are real, and if the latter, what their slopes are. We need to know more about the supply functions of the factors of production and whether the actual processes of distribution furnish any degree of corroboration to the inductive tendencies discovered. This book is an attempt to do just that.

Thus, the Cobb–Douglas regression represented a bold attempt to join up-to-date statistical methods with the still controversial marginal productivity theoretical framework.

[5] In Biddle (1999), I document this transition, while Ayres (1927) provides a contemporary account.

*Figure 1*
**One of Five "Charts" from Cobb and Douglas's "A Theory of Production"**

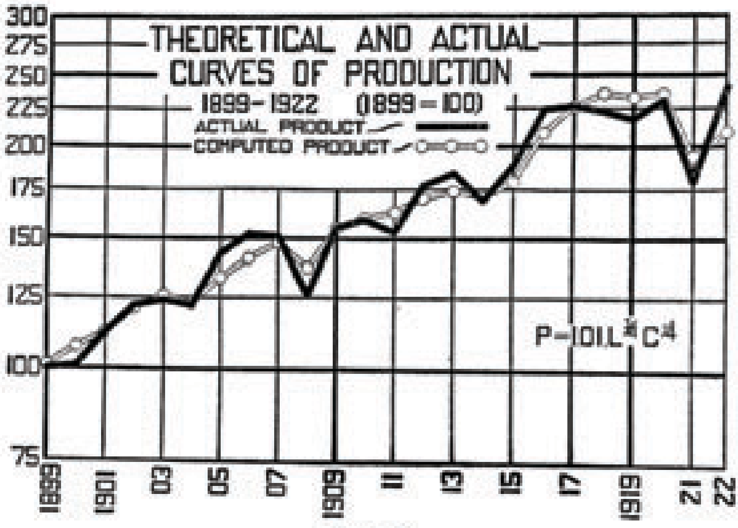


Chart II

*Source:* Chart II from "A Theory of Production" (Cobb and Douglas 1928, p. 153).

In both "A Theory of Production" and *The Theory of Wages*, Douglas needed to convince readers that his regression procedure provided reliable measures of real and interesting economic quantities, a task made more difficult by the novelty of his approach. He resorted to an array of arguments to do so. His primary argument was based on the fit of the regressions. The time series Cobb–Douglas regressions tended to generate a close correspondence between the actual values of the output variable ($P$) and values predicted by the regression in the manner described above ($P'$). This correspondence, Douglas argued, strongly suggested that the regressions captured a true relationship between inputs and output. He demonstrated this goodness of fit using correlation coefficients, tables, and graphs, such as Figure 1, reproduced from "A Theory of Production" (1928), showing the relationship between actual and "theoretical" (predicted) output. He raised the possibility that the good fits were due to spurious correlation between trending series, noting that "it has some times been charged that . . . equally good results would be secured by comparing the relative movement of hogs in Wisconsin, cattle in Wisconsin, with the physical product in manufacturing" but responded to this concern by arguing that there was an a priori theoretical connection between capital, labor, and output that did not exist for pigs, cattle, and output and by showing that deviations of $P$ and $P'$ from their three-year moving averages were also highly correlated (Cobb and Douglas 1928, p. 160).

In an interesting twist on the fit argument, Douglas pointed out that observations (years) showing large differences between actual $P$ and estimated $P'$, because they were abnormal in some way, actually strengthened the case that the Cobb–Douglas regression reflected the normal relationship between capital, labor, and product. Years in which $P$ was below $P'$ were recession years, and those in which $P$ was above $P'$ were years of prosperity. Since the capital index measured existing capital rather than capital utilized, and the labor index was men rather than man-hours, one would expect this pattern: in a recession, when plants were idled and overtime eliminated, the capital and labor indexes overstated the amount of the inputs actually employed, and so the equation produced a predicted output that was too high. Likewise, prosperities were periods of full capital utilization and long, intense hours for workers, leading the indexes to understate true input use (Douglas 1934, pp. 160–61).

Douglas also made much of evidence that various measures of labor's compensation tended to correspond to measures of marginal productivity derived from his estimates. This included the striking similarity between the .75 estimate for $k$ produced by the Cobb–Douglas regression and the estimate from the National Bureau of Economic Research of labor's share of the total value added in manufacturing, but Douglas added to this by showing positive correlations between moving averages of real wages in various industries and moving averages of the measures of labor's marginal product in those industries implied by his function. There was some ambiguity, however, as to how readers were to regard this evidence. At times, Douglas seemed to be arguing that the regression provided a means of testing the marginal productivity theory, and that a close concurrence between wages and estimated marginal products represented a verification of this theory. More often, however, Douglas pointed to the consistency of his results with the predictions of marginal productivity theory as an additional proof of the plausibility of his procedure, a line of argument that essentially assumed the truth of the marginal productivity theory.[6]

## Critiques and Responses

Douglas made bold claims in "A Theory of Production" and *A Theory of Wages*: Using generally available data and accessible statistical techniques, he had shown that the actual relationship between capital, labor, and output in manufacturing could be closely approximated by a simple function, one which embodied and allowed quantification of the hypothesis of diminishing marginal productivity. He had demonstrated a relationship between the characteristics of this "law of production" and the distribution of income between labor and capital, a relationship posited by a well-known but still-contested theory of distribution. Unsurprisingly, the Cobb–Douglas regression attracted considerable attention from Douglas's fellow economists.

[6] This ambiguity about whether the Cobb–Douglas procedure should be regarded as a means of testing various predictions of marginal productivity theory, as a way of measuring the parameters of a theory maintained to be true, or both, was not clarified in Douglas's subsequent discussions of the procedure, and was at times a point of contention in the literature on production function estimation in subsequent decades. See, for example, Mendershausen (1941), and the exchange between Shaikh (1974) and Solow (1974).

Sumner Slichter, assigned to discuss Cobb and Douglas's paper at the 1927 AEA meetings, was a decidedly unfriendly critic of the work (Slichter 1928). The bulk of his remarks were devoted to detailing problems with the index of fixed capital constructed by Douglas, but his complaints went beyond issues of data quality as he believed the entire project to be wrong-headed. Despite the fact that the marginal productivity theory was not explicitly mentioned in the paper, Slichter thought he could see a hidden agenda, and he did not approve. "Professors Cobb and Douglas conclude that it has been statistically demonstrated that the relationship between the agents of production on one hand and the volume of output on the other meets the requirements of the marginal productivity hypothesis." Slichter disputed this specific claim and argued more generally that marginal productivity theory had little to offer as a framework for thinking about distribution. He closed his comments with a final indictment of the research (p. 170):

> There is probably no more important single cause for our meagre knowledge of the distributive process than the fact that the subject has been so largely studied within the narrow limits imposed by the assumptions of static economics. . . . Quantitative economics, by helping to provide the raw materials for a realistic theory, can be of great use in liberating the study of distribution from the tyranny of economic statics. But it can be of little assistance if statisticians and mathematical economists are too completely preoccupied with verifying the propositions of static doctrine.

In *The Theory of Wages,* Douglas (1934) responded to critics like Slichter who complained of the lack of realism in the marginal productivity theory. He included a chapter on the assumptions of the theory, both explicit and implicit, with long discussions of the extent to which each was valid for the United States. After arranging the implicit assumptions on a scale ranging from "largely valid but not wholly so" to "partially true but on the whole not true," Douglas (pp. 94–95) commented:

> Many, who have seen the degree of variance between real life and the assumptions of the productivity school, have in their impatience declared that because of this defective basis, the conclusions which have been drawn from the productivity theory are not worthy of credence and hence should be disregarded. But such an attitude as this ignores the fact that the assumptions do represent real tendencies which in the aggregate are more powerful than those of a conflicting nature.

Such critics, Douglas argued, seemed ignorant of the fact that the method of deduction and abstraction used to build the marginal productivity theory was also the method that had achieved great success in the natural sciences.

Another type of unfriendly critic was exemplified by Douglas's colleague Frank Knight, who argued that the key concepts of economic theory were essentially static and abstract, while historical data was dynamic, reflecting the action of forces that were assumed away in static theory. Thus, statistical methods could never quantify theoretical concepts.[7] Douglas (1934, pp. 106, 107) dismissed such arguments rather undiplomatically in *The Theory of Wages*:

> [T]he high priests of "pure" theory are never tired of pointing out that they are dealing only with static conditions—as of one moment in time for one community. When statistical series dealing with time sequences or even relative distributions in space are brought forward, the armchair theorists brush these aside on the ground that they may include shiftings of the curves or different curves. These series are then dismissed as being merely historical or empirical.
>
> Now it is of course true that one of the aims of statistical economics . . . should be to approximate as far as possible the static concepts and to give concrete meaning and definite values to them. But if this cannot be completely carried out . . . [s]hould we abandon all efforts at the inductive determination of economic theory and remain in the ivory tower of "pure" theory[?] If this is what is done, we may as well abandon all hope of further developing the science of economics and content ourselves with merely the elaboration of hypothetical assumptions which will be of little aid in solving problems since we will not know the values. Or shall we try to make economics a progressive science?

There were also friendly critics of the Cobb–Douglas paper who, while finding fault with various details of what Douglas had done, expressed considerable enthusiasm for the "method of attack" represented by the research and offered constructive suggestions for pushing the research program forward. One such friendly critic was Douglas's colleague and prominent empirical economist Henry Schultz. Schultz accepted the point that Cobb and Douglas's statistical procedure, which employed time series data but made no adjustment for secular changes, could not result in a verification of a static theory like the marginal productivity theory.[8] He argued, however, that that the Cobb–Douglas method of estimating production relations should be "modified, not abandoned." Referring to his own approach to estimating the supply and demand curves implied by neoclassical theory, Schultz (1929) described the strategy of adjusting the data to remove long-term trends, then correlating deviations from those trends in order to isolate relationships that were closer in principle to the concepts of static theory.

Another prominent friendly critic was J. M. Clark, who within a few months of the appearance of "A Theory of Production" published an article devoted solely to discussing issues raised by the Cobb–Douglas paper (Clark 1928). His criticisms were

[7] Letter from F. Knight to Douglas, 10/12/1932; see also Douglas (1976, p. 905). Knight, whose opinion carried considerable weight with a number of younger Chicago faculty members at the time, was generally hostile to empirical work in economics, and was highly critical of Henry Schultz's work as well (Reder 1982).

[8] As Morgan (1990, sec 5.2) explains, possible approaches to testing a static theory with time series data were much discussed in the literature on estimating supply and demand relationships, to which Schultz was a major contributor.

numerous, but most were constructive, aimed at improving the Cobb–Douglas analysis of marginal productivity rather than discrediting it. Like Slichter, he questioned the accuracy of Douglas's capital and labor series, but took it for granted that "they will be improved and refined as the authors continue their researches." A central feature of the article was a suggestion for an augmented version of the Cobb–Douglas regression. Clark believed that the Cobb–Douglas equation offered a good account of the "normal" or long–run relationship between labor, capital, and output, but did a poor job of representing the effect of cyclical fluctuations in labor and capital utilization. He proposed altering the function by including a factor representing cyclical swings, and suggested $P = L^k C^{1-k}(L/L_n)^m$, where $L_n$ represented the "normal" level of employment, or that which the capital stock was designed to accommodate. This formula allows the marginal increases in labor input to have a magnified impact on output if they are allowing idled capital to be put back into use. Moreover, Clark fit this function to Douglas's data, although not by the method of least squares. Clark was also troubled by the fact that the Cobb–Douglas regression left no room for improvements in technology to affect productivity, but believed that this problem could be solved, if not with data based on historical aggregates, then with comparative studies of "simultaneous" data from different industries. The Cobb–Douglas study was "a bold and significant piece of pioneer work in a hitherto neglected field," which Clark clearly hoped would be followed up by others.

In *The Theory of Wages*, Douglas responded positively to many of Clark's comments and suggestions, even reporting some of his own experiments with Clark's modified formula. He also devoted several pages to the question of how the substantial technological progress over his sample period affected the meaning of his estimates, a concern raised by other commentators in addition to Clark. He did not go so far as to develop a modification of his regression procedure that would accommodate technological change, nor would he ever, as he would soon change the focus of his research program to the estimation of the regression with cross-section datasets, for which the issue of accounting for technological progress was no longer important.[9]

Modern economists may sense an air of the familiar in the initial reactions to Douglas's new empirical procedure. While few economists today would agree with Sumner Slichter's rejection of the marginal productivity theory as a framework for the inductive study of distribution, it is not unusual to see an empirical research program being criticized on the grounds that it is guided—or the critic might say constrained—by a too-literal reliance on the assumptions of an oversimplified theoretical framework and will thus miss important aspects of the phenomenon under study and/or produce meaningless results. It is also not uncommon to see the argument that the quality of existing data is not adequate to the needs of a

[9] Moving to cross sections of industry-level data introduced the new concern that different industries might have different production functions, a problem that Douglas recognized but never addressed empirically. A solution to the problem of estimating an aggregate time series production function while also accounting for and even measuring technological improvement was introduced in Solow's (1957) seminal article on growth accounting.

proposed new empirical technique. Such arguments may be offered as a reason to reject the new technique, as Slichter seemed to be doing in his discussion of "A Theory of Production," but some who raise issues of data quality take the attitude of J. M. Clark that the eventual development of better data is to be expected, making it worthwhile to experiment with and develop a better understanding of the technique in the meantime.

Finally, the differing reactions of Frank Knight and Henry Schultz to the Cobb–Douglas regression reflect a difference of opinion that persisted, and arguably still persists, among economists. Both men were strong believers in the usefulness for economic research of neoclassical equilibrium models and the abstract deductive method that produced them, but they disagreed on whether it was worthwhile or even feasible to attempt to use data from a messy, dynamic world to test these models or to extract empirical measures of their key components. Schultz believed that the bridging of the theory–data gap through the further development of statistical techniques and data was both possible and one of the most important tasks of modern economics, and he applauded Douglas's work; Knight believed that economic theoretical concepts like the market demand curve for a good were essentially ideal types that could never manifest themselves in data generated by a dynamic real world economy, and saw work like that of Douglas and Schultz as a waste of time at best, and perniciously misleading at worst.

After the appearance of *The Theory of Wages*, a significant set of friendly criticisms came from mathematical economists who embraced marginal productivity theory and who were trying to make sense of the relationship between Douglas's regression equation and the equations of their theoretical systems. These included Wassily Leontief (1934), Jacob Marschak (1936), and David Durand (1937). It was in these articles, as well as a very critical article by Mendershausen (1938) that the phrase "production function" was first consistently applied to the relationship that Douglas was attempting to estimate, although Douglas quickly adopted it himself. At the time, the phrase was rare in the economics literature, used almost exclusively by those concerned with mathematical formalization of neoclassical theory and/or the statistical estimation of the components of the resulting models. As mentioned earlier, Douglas saw his work as a part of the program to estimate the relationships of neoclassical theory, so the young econometricians who referred to Douglas's regression as a production function were affirming Douglas's conception.

However, while these neoclassically oriented econometricians were embracing Douglas's program as complementary to their own, they were also redefining the objectives of the program and developing criteria for evaluating Douglas's methods and results that Douglas himself would not have accepted. This outcome arose partly because they thought about marginal productivity theory within the context of Walrasian general equilibrium theory, in which the production function was a characteristic of a firm. As a result, they were crucially concerned with whether the estimated coefficients revealed anything useful about the true parameters of firm-level production functions, or perhaps were in some sense averages of those parameters. But this question, which became increasingly important to economists engaged in production function estimation as time passed, was never crucial for Douglas. As Douglas (1971, p. 29) noted in his autobiography, he was taught theory by J. B. Clark, and received "a thorough drilling in (the marginal productivity) principle, which served me well . . . when I started my own inductive work in the theory." But Clark's formal analysis ran in terms of aggregates: the basic wage rate and the interest rate depended on the marginal products of "social" capital and "social" labor (Stigler 1941, p. 307). A student of Clark would have had no trouble thinking of an aggregate production function as a primal entity to be estimated, and its parameters as significant theoretical quantities in themselves.

## Aftermath

Douglas's work with the Cobb–Douglas production function continued for another 14 years after *The Theory of Wages*, as he and various coauthors estimated Cobb–Douglas regressions using both aggregate time series data and cross-section industry-level data, while developing a series of arguments defending the procedure as a way of obtaining important knowledge about the economy from empirical data. Douglas's research career ended in 1948 with his election to the U.S. Senate, but in the ensuing decades the procedure of regressing the log of a measure of output on the logs of measures of various inputs became a standard and accepted empirical procedure in a number of areas of microeconomics and macroeconomics, while spawning several other, more complex approaches to estimating empirical production functions. There remain open questions about the scientific value of this procedure in each of the contexts in which it is applied, some of which are variations of the friendly and unfriendly questions raised by Douglas's initial critics. However, measured by the extent to which it has been embraced, applied, and elaborated upon by subsequent economists, Douglas's innovative 1927 idea that one could use statistical analysis to uncover meaningful empirical relationships between inputs and outputs, as well as his specific implementation of that idea using the Cobb–Douglas functional form and least squares regression, was an overwhelming success.

■ *The author would like to thank Ross Emmett, David Autor, Timothy Taylor, Charles Johnson, and John List for helpful comments.*

## References

**Ayres, Leonard.** 1927. "The Dilemma of the New Statistics." *Journal of the American Statistical Association* 22(157): 1–8.

**Biddle, Jeff E.** 1999. "Statistical Economics, 1900–1950." *History of Political Economy* 31(4): 607–51.

**Clark, J. M.** 1928. "Inductive Evidence on Marginal Productivity." *American Economic Review* 18(3): 450–67.

**Cobb, Charles W.** 1930. "Production in Massachusetts Manufacturing, 1890–1928." *Journal of Political Economy* 38(6): 705–07.

**Cobb, Charles W.** 1939. "Note on Frisch's Diagonal Regression." *Econometrica* 7(1): 77–80.

**Cobb, Charles W., and Paul H. Douglas.** 1928. "A Theory of Production." *American Economic Review,* Papers and Proceedings of the Fortieth Annual Meeting of the American Economic Association, 18(1, Supplement): 139–65.

**Douglas, Paul H.** 1918. "The Problem of Labor Turnover." *American Economic Review* 8(2): 306–16.

**Douglas, Paul H.** 1919. "Is the New Immigration More Unskilled than the Old?" *Publications of the American Statistical Association* 16(126): 393–403.

**Douglas, Paul H.** 1930. *Real Wages in the United States, 1890–1926.* New York: Houghton Mifflin.

**Douglas, Paul H.** 1934. *The Theory of Wages.* New York: MacMillan.

**Douglas, Paul H.** 1948. "Are There Laws of Production?" *American Economic Review* 38(1): i–ii, 1–41.

**Douglas, Paul H.** 1971. *In the Fullness of Time: The Memoirs of Paul H. Douglas.* New York: Harcourt Brace Jovanovich.

**Douglas, Paul H.** 1976. "The Cobb–Douglas Production Function Once Again: Its History, Its Testing, and Some New Empirical Values." *Journal of Political Economy* 84(5): 903–15.

**Douglas, Paul H., and Francis Lamberson.** 1921. "The Movement of Real Wages, 1890–1918." *American Economic Review* 11(3): 409–26.

**Durand, David.** 1937. "Some Thoughts on Marginal Productivity, with Special Reference to Professor Douglas' Analysis." *Journal of Political Economy* 45(6): 740–58.

**Knight, Frank.** 1932. Letter to Paul Douglas, 10/12/1932. Frank Knight papers, Box 58, Folder 16. Housed in the Special Collections of the University of Chicago Library.

**Leontief, Wassily.** 1934. "Interest on Capital and Distribution: A Problem in the Theory of Marginal Productivity." *Quarterly Journal of Economics* 49(1): 147–61.

**Marschak, Jacob.** 1936. "An Empirical Analysis of the Laws of Distribution." *Economica,* New Series, 3(10): 221–26.

**Mendershausen, Horst.** 1941. "A Rejoinder" to "A Reply to Dr. Menderhausen's Criticism" in comments labeled "On the Significance of Another Production Function." *American Economic Review* 31(3): 567–69.

**Mendershausen, Horst.** 1938. "On the Significance of Professor Douglas' Production Function." *Econometrica* 6(2): 143–53.

**Morgan, Mary S.** 1990. *The History of Econometric Ideas.* Cambridge, UK: Cambridge University Press.

**Reder, Melvin W.** 1982. "Chicago Economics: Permanence and Change." *Journal of Economic Literature* 20(1): 1–38.

**Samuelson, Paul A.** 1979. "Paul Douglas's Measurement of Production Functions and Marginal Productivities." *Journal of Political Economy* 87(5, Part 1): 923–39.

**Schultz, Henry.** 1929. "Marginal Productivity and the General Pricing Process." *Journal of Political Economy* 37(5): 505–51.

**Shaikh, Anwar.** 1974. "Laws of Production and Laws of Algebra: The Humbug Production Function." *Review of Economics and Statistics* 56(1): 115–20.

**Slichter, Sumner.** 1928. "Economic and Social Aspects of Increased Productive Efficiency—Discussion." *American Economic Review,* Papers and Proceedings of the Fortieth Annual Meeting of the American Economic Association, 18(1, Supplement): 166–170.

**Solow, Robert.** 1957. "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics* 39(3): 312–20.

**Solow, Robert.** 1974. "Law of Production and Laws of Algebra: The Humbug Production Function: A Comment." *Review of Economics and Statistics* 56(1): 121.

**Stigler, George J.** 1941. *Production and Distribution Theories: The Formative Period.* New York: MacMillan.

**Wicksell, Knut.** 1896. "Ein neues Prinzip der gerechen Besteurung." In *Finanz theoretische Unterschungen,* Jena. (Translated as "A New Principle of Just Taxation," in *Classics in the Theory of Public Finance,* edited by R.A. Musgrave and A.T. Peacock. London: MacMillan, 1958.)

Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 237–244

# 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 e-mail at ⟨taylort@macalester.edu⟩, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., Saint Paul, Minnesota, 55105.

## Smorgasbord

C. Randall Henning and Martin Kessler discuss "Fiscal Federalism: US History for Architects of Europe's Fiscal Union." They discuss how Alexander Hamilton pushed for a plan under which the federal government took over state debts in 1790—a pattern that would then be maintained for a half-century. "[T]he debt assumption of 1790 set a precedent that endured for several decades. The federal government assumed the debt of states again after the War of 1812 and then for the District of Columbia in 1836. . . . This pattern was broken in the 1840s, when eight states plus Florida, then a territory, defaulted. . . . The indebted states petitioned Congress to assume their debts, citing the multiple precedents. British and Dutch creditors, who held 70 percent of the debt on which states later defaulted, pressed the federal government to cover the obligations of the states. They argued that the federal government's guarantee, while not explicit, had been implied. Prices of

■ *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.26.2.237. doi=10.1257/jep.26.2.237

the bonds of even financially sound states fell and the federal government was cut off from European financiers in 1842. . . . John Quincy Adams evidently believed that another war with Britain was likely if state debts were not assumed by the federal government. . . . However, on this occasion Congress rejected the assumption petition . . . Eventually, most states repaid all or most of their debt as a condition for returning to the markets. . . . The rejection of debt assumption established a "no bailout" norm on the part of the federal government. The norm is neither a 'clause' in the US Constitution nor a provision of federal law. Nevertheless, whereas no bailout request had been denied by the federal government prior to 1840, no such request has been granted since, with one special exception [the District of Columbia in the 1970s] . . . During the 1840s and 1850s, states adopted balanced budget amendments to their constitutions or other provisions in state law requiring balanced budgets. This was true even of financially sound states that had not defaulted and their adoption continued over the course of subsequent decades, so that eventually three-fourths of the states had adopted such restrictions." Available here as Working Paper 12-1 from the Peterson Institute for International Economics, January 2012, at ⟨http://www.iie.com/publications/wp/wp12-1.pdf⟩ and the Breugel Essay and Lecture Series, January 2012, ⟨http://www.bruegel.org/download/parent/669-fiscal-federalism-us-history-for-architects-of-europes-fiscal-union/file/1537-fiscal-federalism-us-history-for-architects-of-europes-fiscal-union/⟩.

Edward Glaeser and Jacob Vigdor document and discuss "The End of the Segregated Century: Racial Separation in America's Neighborhoods, 1890–2010." "Segregation has declined steadily from its mid-century peak, with significant drops in every decade since 1970. As of 2010, the separation of African-Americans from individuals of other races stood at its lowest level in nearly a century. Fifty years ago, nearly half the black population lived in what might be termed a "ghetto" neighborhood, with an African-American share above 80 percent. Today, that proportion has fallen to 20 percent." "There is every reason to relish the fact that there is more freedom in housing today than 50 years ago and to applaud those who fought to create that change. Yet we now know that eliminating segregation was not a magic bullet. Residential segregation has declined pervasively, as ghettos depopulate and the nation's population center shifts toward the less segregated Sun Belt. At the same time, there has been only limited progress in closing achievement and employment gaps between blacks and whites." Manhattan Institute for Policy Research, Civic Report No. 66, January 2012. At ⟨http://www.manhattan-institute.org/pdf/cr_66.pdf⟩.

Cass B. Sunstein, now serving as Administrator of the Office of Information and Regulatory Affairs within the Office of Management and Budget, provides an overview of "Empirically Informed Regulation." "In recent years, a number of social scientists have been incorporating empirical findings about human behavior into economic models. These findings offer useful insights for thinking about regulation and its likely consequences. . . . Relevant research suggests that four such approaches have particular promise: (1) using disclosure as a regulatory tool, especially if disclosure policies are designed with an appreciation of how people process information; (2) simplifying and easing choices through appropriate default rules, reduction of complexity and paperwork requirements, and related strategies; (3) increasing the salience of certain factors or variables; and (4) promoting social norms through private–public partnerships and other approaches that operate in the service of agreed-upon public goals. . . . But even if the standard accounts of potential market failures are supplemented, it does not necessarily follow that more regulation is justified. . . . Indeed, some of the findings might argue in favor of less rather than more regulation. . . . It should not be necessary to acknowledge that public officials are subject to error as well. Indeed, errors may result from one or more of the findings traced above; officials are human and may also err. The dynamics of the political process may or may not lead in the right direction. It would be absurd to say that empirically informed regulation is more aggressive than regulation that is not so informed, or that an understanding of recent empirical findings calls for more regulation rather than less. The argument is instead that such an understanding can help to inform the design of regulatory programs." *University of Chicago Law Review*, Fall 2011, vol. 78, no. 4, pp. 1349–1429. At ⟨http://lawreview.uchicago.edu/sites/lawreview.uchicago.edu/files/uploads/78_4/Sunstein_Essay.pdf⟩.

Alan B. Krueger, now serving as Chairman of the Council of Economic Advisers, has given a speech on "The Rise and Consequences of Inequality in the United States." "Not since the Roaring Twenties has the share of income going to the very top reached such high levels. The magnitude of these shifts is mindboggling. The share of all income accruing to the top 1% increased by 13.5 percentage points from 1979 to 2007. This is the equivalent of shifting $1.1 trillion of *annual* income to the top 1 percent of families. Put another way, the increase in the share of income going to the top 1% over this period exceeds the total amount of income that the entire bottom 40 percent of households receives." "Countries that have a high degree of inequality also tend to have less economic mobility across generations. . . . While we will not know for sure whether, and how much, income mobility across generations has been exacerbated by the rise in inequality in the U.S. until today's children have grown up and completed their careers . . . the persistence in the advantages and disadvantages of income passed from parents to the children is predicted to rise by about a quarter for the next generation as a result of the rise in inequality that the U.S. has seen in the last 25 years." January 12, 2012. At ⟨http://www.whitehouse.gov/sites/default/files/krueger_cap_speech_final_remarks.pdf⟩.

Ruth Towse provides an overview of "What We Know, What We Don't Know, and What Policy-makers Would Like Us to Know about the Economics of Copyright." "Almost all economists are agreed that the copyright term is now inefficiently long with the result that costs of compliance most likely exceed any financial benefits from extensions (and it is worth remembering that the term of protection for a work in the 1709 Statute of Anne was 14 years with the possibility of renewal as compared to 70 years plus life for authors in most developed countries in the present, which means a work could be protected for well over 150 years)." "Copyright could become more similar to a patent by having an initial term of protection of a work, say of 20 years, renewable for further terms. . . . The advantage of this is twofold: it enables a 'use it or lose it' regime to function and, more relevant to the economics of copyright, it enables the market to function better in valuing a work (the vast majority of works, as we know, are anyway out of print because they are deemed to have no commercial value while the copyright is still valid); knowing that renewal would be necessary would also alter contractual terms between creators and intermediaries, thereby improving the efficiency of contracting and the prospect of fairer contracts." *Review of Economic Research on Copyright Issues*, December 2011, vol. 8, no. 2, pp. 101–120. At ⟨http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2024588⟩.

Philip Turner asks "Is the Long-term Interest Rate a Policy Victim, a Policy Variable or a Policy Lodestar?" "From the mid-1950s to the early 1980s, this aggregate [debt of domestic US non-financial borrowers—governments, corporations, and households] was remarkably stable—at about 130% of GDP. It was even described as the great constant of the US financial system. The subcomponents moved about quite a bit—for instance, with lower public sector debt being compensated by higher private debt. But the aggregate itself seemed very stable. During the 1980s, however, this stability ended. Aggregate debt rose to a new plateau of about 180% of GDP in the United States. At the time, this led to some consternation in policy circles about the burden of too much debt. It is now about 240% of GDP. Leverage thus measured—that is, as a ratio of debt to income—has increased. Very many observers worry about this. Whatever the worries, lower rates do make leveraged positions easier to finance. Once account has been taken of lower real interest rates, debt servicing costs currently are actually rather modest . . . "The concluding note of caution is this: beware of the consequences of sudden movements in yields when long-term rates are very low. . . . A change of 48 basis points in one month . . . would have a larger impact when yields are 2% than when they are 6%. With government debt/GDP ratios set to be very high for years, there is a significant risk of instability in bond markets." Bank of International Settlements, BIS Working Papers No. 367, December 2011. At ⟨http://www.bis.org/publ/work367.pdf⟩.

## Complements to JEP Articles

Adeel Malik and Bassem Awadallah discuss "The Economics of the Arab Spring." "During the period 1996–2006, labour force in Middle East and North Africa has grown three times as much annually as in the rest of the developing world, resulting in one of the largest rates of youth unemployment in the world." "The state in most Arab economies is the most important economic actor, eclipsing all independent productive sectors. . . . The state-centred development paradigm rests on an uninterrupted flow of external windfalls. In fact, many of the region's pathologies—whether it is a weak private sector, segmented labor markets, or limited regional trade—are ultimately rooted in an economic structure that relies overwhelmingly on rents derived from fuel exports, foreign aid or remittances. Reliance on such unearned income streams is the 'original sin' for Arab economies." Center for the Study of

African Economies, University of Oxford, CSAE WPS/2011-23, December 2011. ⟨http://www.csae.ox.ac.uk/workingpapers/pdfs/csae-wps-2011-23.pdf⟩. This article makes a useful complement to the article by Filipe R. Campante and Davin Chor in this issue.

The Congressional Budget Office has published "Comparing the Compensation of Federal and Private-Sector Employees." "Overall, the federal government paid 2 percent more in total wages than it would have if average wages had been comparable with those in the private sector, after accounting for certain observable characteristics of workers. . . . On average for workers at all levels of education, the cost of hourly benefits was 48 percent higher for federal civilian employees than for private-sector employees with certain similar observable characteristics . . . The most important factor contributing to differences between the two sectors in the costs of benefits is the defined-benefit pension plan that is available to most federal employees. . . . Overall, the federal government paid 16 percent more in total compensation than it would have if average compensation had been comparable with that in the private sector, after accounting for certain observable characteristics of workers." January 2012. At ⟨http://www.cbo.gov/sites/default/files/cbofiles/attachments/01-30-FedPay.pdf⟩. The CBO results and discussion with regard to federal workers are qualitatively similar, and complementary, to the article by Maury Gittleman and Brooks Pierce in the Winter 2012 issue of this journal, "Compensation for State and Local Government Workers."

The Federation of American Scientists have published a collection of essays under the title *The Future of Nuclear Power,* edited by Charles D. Ferguson and Frank A. Settle. In Chapter 2, Stephen Maloney offers "A Critical Examination of Nuclear Power's Costs." "Since the nuclear industry's inception more than 50 years ago, its forecasts for costs have been consistently unreliable. The 'first generation' plants, comprising both prototype reactors and the standard designs of the 1950s–1960s, failed to live up to promised economics. This trend continued with the construction of Generation II plants completed in the 1970s, which make up the present nuclear fleet. . . . Nuclear plant construction costs escalated approximately 24 percent per calendar year compared to 6 percent annual escalation for coal plants. . . . The scale-up of nuclear plants brought less than half the economic efficiencies projected." More recently, "In June 2006, a consortium of companies announced plans to build two more reactors at the South Texas Project site for an estimated cost of $5.2 billion. NRG, the lead company, made history by becoming the first company to file an application with the NRC. CPS Energy, a municipal utility, was one of its partners. In October 2007, CPS Energy's board approved $206 million for preliminary design and engineering. In June 2009, NRG revised the estimate to $10 billion for the two reactors, including finance charges. A few weeks later, this estimate rose to $13 billion, including finance charges. Later that year, the estimate reached $18.2 billion . . ." February 2012. At ⟨http://www.fas.org/pubs/_docs/Nuclear_Energy_Report-lowres.pdf⟩. The report offers a useful supplement to the paper by Lucas W. Davis, "Prospects for Nuclear Power," in the Winter 2012 issue of this journal.

## Inflation and Hyperinflation

Janet Koech gives chapter and verse on "Hyperinflation in Zimbabwe." "From 2007 to 2008, the local legal tender lost more than 99.9 percent of its value." "Hyperinflation and economic troubles were so profound that by 2008, they wiped out the wealth of citizens and set the country back more than a half century. In 1954, the average GDP per capita for Southern Rhodesia was US$151 per year (based on constant 2005 U.S.-dollar purchasing power-parity rates). In 2008, that average declined to US$136, eliminating gains over the preceding 53 years . . ." "The *Economic Times* newspaper noted on June 13, 2008, that "a loaf of bread now costs what 12 new cars did a decade ago," and "a small pack of locally produced coffee beans costs just short of 1 billion Zimbabwe dollars. A decade ago, that sum would have bought 60 new cars." In early 2009, the economy dollarized: "While the South African rand, Botswana pula and the U.S. dollar were granted official status, the U.S. dollar became the principal currency. Budget revenue estimates and planned expenditures for 2009 were denominated in U.S. dollars, and the subsequent budget for 2010 was also set in U.S. dollars. An estimated four-fifths of all transactions in 2010 took place in U.S. dollars, including most wage payments . . ." *Annual Report of the Globalization and Monetary Policy Institute*, Federal Reserve Bank of Dallas. At ⟨http://www.dallasfed.org/assets/documents/institute/annual/2011/annual11b.pdf⟩.

Phil Davies writes about "Taking the Measure of Prices and Inflation," subtitled "A century of evolution—and near-constant criticism—has greatly improved price indexes. But work continues to perfect these closely-watched economic indicators." "In 1904, the federal Bureau of Labor (forerunner of the BLS) published a monthly index of retail food prices gleaned from 800 merchants in large industrial centers. The index, covering the past 13 years, priced 30 principal food items and weighted them according to average consumption. Within a few years, the food price index reflected data gathered from over 1,000 retail establishments in 40 states. . . . Intent on setting equitable wages for factory workers vital to the war effort, the National War Labor Board in 1918 called upon the BLS to produce nationwide data on the 'cost of living'; changes over time in this index would indicate how much household income would have to change to maintain roughly the same standard of living. The stated goal of this exercise: '[I]nsure the subsistence of the worker and his family in health and reasonable comfort.'. . . Over the next two years, BLS agents fanned out across the country to collect prices for about 145 consumer products and services. Price takers carefully specified items to make pricing of identical or similar items easier in future surveys and surveyed about 12,000 working-class families in 42 states to gather information about income and consumption patterns. In 1919, the bureau released the first comprehensive set of cost-of-living indexes for 31 major industrial and shipbuilding centers. Thereafter, updated indexes were issued semiannually for individual cities (Washington, D.C., was added in 1921) and the nation as a whole." The essay also discusses how price indexes have sought to address "quality change, consumer substitution and technological innovation," and the differences between the Consumer Price Index and the Personal Consumption Expenditure as measures of inflation. *The Region*, Federal Reserve Bank of Minneapolis, December 2011, pp. 28–38. At ⟨http://www.minneapolisfed.org/publications_papers/pub_display.cfm?id=4792⟩.

## Collections on a Theme

The Fall 2011 issue of *The Future of Children* features nine article on the theme "Work and Family." From a description of the paper by Suzanne M. Bianchi, "Changing Families, Changing Workplaces": "The share of married mothers in the labor force has risen from a little over a quarter in 1960 to more than 70 percent today. During the 1960s, only 10 percent of mothers were at work within three months of giving birth; by the early years of the twenty-first century that figure had risen to over 40 percent, with 64 percent of women back at work within twelve months after a birth. Labor force participation rates are now nearly as high among women with preschool-aged children as they are among those with school-aged children. Over the same period, the share of children living with a single parent has grown sharply. Today about one-quarter of families with children are headed by single parents . . ." Alison Earle, Zitha Mokomane, and Jody Heymann discuss "International Perspectives on Work-Family Policies: Lessons from the World's Most Competitive Economies." From the "Summary": "Using indicators of competitiveness gathered by the World Economic Forum, the authors identify fifteen countries, including the United States, that have been among the top twenty countries in competitiveness rankings for at least eight of ten years. . . . They find that every one of these countries, except the United States, guarantees some form of paid leave for new mothers as well as annual leave. And all but Switzerland and the United States guarantee paid leave for new fathers. . . . The majority of these countries provide paid leave for new mothers, paid leave for new fathers, paid leave to care for children's health care needs, breast-feeding breaks, paid vacation leave, and a weekly day of rest. Of these, the United States guarantees only breast-feeding breaks (part of the recently passed health care legislation)." The journal is at ⟨http://futureofchildren.org/⟩.

The Winter 2012 issue of the *Cato Journal* has a dozen articles on the subject "Is Immigration Good for America?" For example, Giovianni Peri presents the case for why immigration of low-wage labor has little effect on wages of natives. "In summary, an economy will respond to immigration along several margins—through increased investment by firms, specialization of natives, complementarities between natives and immigrants, technological response by firms, and job creation. . . . This explains why a long tradition of empirical economic studies has found very small to no effect of U.S. immigration on native wages and employment at the national and at the local level." The issue is available at ⟨http://www.cato.org/pubs/journal/cj32n1/cj32n1.html⟩.

## Discussion Starters

Juha Siikamäki inquires into "State Parks: Assessing Their Benefits." "Valuing recreation time monetarily requires determining the opportunity cost of time. To illustrate the potential magnitude of recreation's time value, I used a conventional and commonly adopted approach where recreation time is valued at one-third the wage rate. . . . Extrapolating from the above results, I estimate about 33 percent of current time use for nature recreation can be attributed to the U.S. state park system. This equals annually about 9.7 hours of nature recreation per capita, or about 2.2 billion hours of nature recreation in total in the United States. The estimated time value of nature recreation generated by the entire U.S. state park system is about $14 billion annually (about $62 per person annually, on average). . . . That value is considerably larger than the annual operation and management costs of state parks." *Resources*, from Resources for the Future, 2012, no. 179, pp. 28–33. At ⟨http://www.rff.org/Publications/Resources/Pages/179-Parks.aspx⟩.

Charles Morris asks "What Should Banks Be Allowed To Do?" "Under the criterion for permissible activities stated above, banking organizations would be able to conduct the following activities: commercial banking, investment banking . . . and asset and wealth management. Investment banking and asset and wealth management are mostly fee-based services that do not put much of a firm's capital at risk. In addition, asset and wealth management are similar to the trust services that always have been allowed for banks. In contrast, the other three categories—dealing and market making, brokerage services, and proprietary trading—have little in common with core banking services and create risks that are difficult to assess, monitor, and control. Banking organizations would be restricted from activities that involve trading, including customer trading. While allowing customer trading might seem reasonable, it would make restrictions on proprietary trading difficult to enforce because the securities inventory used to facilitate customer trading cannot be easily distinguished from proprietary assets." *Economic Review*, Federal Reserve Bank of Kansas City, Fourth Quarter 2011, pp. 55–80. At ⟨http://www.kc.frb.org/publicat/econrev/pdf/11q4Morris.pdf⟩.

# Notes

***For additional announcements, check out the continuously updated* JEP *online Bulletin Board,*** *⟨http://www.aeaweb.org/bulletinboard.php⟩. Calls for papers, notices of professional meetings, and other announcements of interest to economists should be submitted to Ann Norman at ⟨jep@jepjournal.org⟩ in one or two paragraphs containing the relevant information. These will be posted at the* JEP *online Bulletin Board. Given sufficient lead time, we will also print a shorter, one-paragraph version of your notice in the "Notes" section of the* Journal of Economic Perspectives. ***Deadlines for "Notes":*** *March 20 for the* JEP *Spring issue, which mails the end of May; June 20 for the* JEP *Summer issue, which mails the end of August; September 20 for the* JEP *Fall issue, which mails the end of November; and December 10 for the* JEP *Winter issue, which mails the end of February. We reserve the right to edit material received.*

**Webcasts are online.** Four 2012 Annual Meeting sessions are available to Association members online: "The Euro: Challenges to Improve a Currency Union," "The Political Economy of U.S. Debt and Deficits," "Markets with Frictions," and "What Happened to the U.S. Employment Miracle?" Your AEA web user ID and password provide access at ⟨http://www.aeaweb.org/webcasts/⟩. Webcasts from 2009 through 2011 and Continuing Education Programs are available for members at the same website.

**Retired faculty available for part-time or temporary teaching.** JOE now lists retired economists interested in teaching on either a part-time or temporary basis at ⟨http://www.aeaweb.org/joe/available_faculty/⟩. Individuals can add or delete their name any time. Listings are deleted on November 30; the service is closed during December and January, reopening February 12.

**2012 Summer Training Program and the AEA Minority Scholarships.** The University of New Mexico (UNM) hosts the 2012 AEA Summer Training and Minority Scholarship Program. The six-week program runs from June 17 through July 28 on the Albuquerque UNM campus. The Program provides undergraduates with study and research opportunities that prepare them with a better understanding of what doctoral-level study of economics entails. Additional information, application, and nomination information is at ⟨http://healthpolicy.unm.edu⟩ or email ⟨center@unm.edu⟩.

**Call for papers and Training Seminar.** The **Society of Government Economists (SGE)** is pleased to announce its annual conference to be held in Washington, DC, on November 5–6, 2012, cosponsored by the George Washington University Economics Department. **Deadline for paper and session proposals: September 30, 2012.** There will be **Special Sessions on Indian Country (Native American) Economics.** Anyone interested in this area of research is encouraged to respond to SGE's call for papers. SGE Conference participants may also register for a special, one-day **Training Seminar on "Integrity, Ethics, and Responsible Leadership in Economics."** The seminar will be led by Deirdre McCloskey (author of *The Cult of Statistical Significance*), George DeMartino (author of *The Economist's Oath*), Susan Offutt (Chief Economist, GAO), and Bryan Roberts (Senior Economist, Nathan Associates). See the Society's website, ⟨http://www.sge-econ.org/⟩, for more information.

**Call for papers.** The annual meeting of the **Transportation and Public Utilities Group (TPUG)** will be held January 4–6, 2013, in conjunction with the AEA annual meetings in San Diego, California. The TPUG invites papers in areas related to transportation, energy, telecommunications, and water and wastewater. Paper proposals should include an electronic copy of the paper title, one-page abstract, and author contact information. For transportation papers, send these items to Professor Peter Loeb at ⟨ploeb@andromeda.rutgers.edu⟩; for public utility papers, Professor Carolyn Freidman at ⟨carolyn.gideon@tufts.edu⟩. All program participants must register for the AEA convention and become TPUG members ($20 TPUG membership fee).

**Call for papers.** The 14th annual conference of the **National Business & Economics Society** will be held March 6–9, 2013, at the Marriott Los Suenos Resort in Herradura, Costa Rica. NBES is a multidisciplinary academic association that focuses on promoting research of both a theoretical and practical nature. See the NBES website at ⟨www.nbesonline.com⟩. Interested authors should submit the full paper or a 1–2 page abstract of the research in question to info@nbesonline.com. **Submission deadline: July 1st 2012.**

**Call for papers.** The **January 2013 Business & Economics Society International (B&ESI) Conference** will be held January 7–10, 2013, in Perth, Australia. **Deadline for abstract submission and participation: December 10, 2012.** You may participate as panel organizer, presenter of one or two papers, chair, moderator, discussant, or observer. All papers will pass a blind peer review process for publication consideration in an anthology of selected papers titled "Global Business & Economics Anthology." For more information, please contact Helen Kantarelis. Phone: 508-852-3937; fax: 508-595-0089; e-mail: ⟨hkan@besiweb.com⟩. Website: ⟨www.besiweb.com⟩.

**Call for papers.** The **July 2013 Business & Economics Society International (B&ESI) Conference** will be held July 6-9, 2013 in Monte Carlo, Monaco (French Riviera.) **Deadline for abstract submission and participation: March 30, 2013.** You may participate as panel organizer, presenter of one or two papers, chair, moderator, discussant, or observer. All papers will pass a blind peer review process for publication consideration in an anthology of selected papers titled "Global Business & Economics Anthology." For more information, please contact Helen Kantarelis. Phone: 508-852-3937; fax: 508-595-0089; e-mail: ⟨hkan@besiweb.com⟩. Website: ⟨www.besiweb.com⟩.

**Call for papers.** The soon-to-be-launched *International Journal of Economics and Business Studies (IJEBS)* will hold its first international conference in Techno India Campus in Salt Lake City, Kolkata (formerly Calcutta) India, on December 20–21, 2012. **Deadline for paper abstracts is July 31, 2012**, and for the completed paper is November 15, 2012. Registration fee is $350.00. Please submit the abstracts to Prof. Debasish Chakraborty, ⟨chakra1d@cmich.edu⟩ (or ⟨dchak@yahoo.com⟩), Phone: 989-859-1499, or to Prof. Kishore G. Kulkarni, (⟨kulkarnk@mscd.edu⟩), Phone: 303-556-2675. More information is available at ⟨www.ijebs.com⟩.

**Call for papers:** In 2013, Elsevier's ***International Review of Economics & Finance*** will feature a **special issue on "The Dynamics of International Migration."** The issue will be guest edited by Oded Stark, Universities of Bonn and Tuebingen. Manuscripts, double-spaced and preferably not longer than 30 pages, should be submitted electronically **on or before October 1, 2012** to Wilhelm Kohler, Professor of International Economics, University of Tuebingen, e-mail: ⟨wilhelm.kohler@uni-tuebingen.de⟩.

**Call for papers.** The *International Journal of Happiness and Development* is soliciting papers for the inaugural as well as subsequent issues. *IJHD* seeks to broaden our understanding of "happiness" and how it may relate to development from economic, political, psychological, and/or sociological perspectives. The Journal entertains all definitions of happiness and interprets development at both micro and macro levels. For additional information, see the Journal website: ⟨www.inderscience.com/ijhd⟩.

**The National Institutes of Health (NIA) adds genetic data to Health and Retirement Study (HRS).** A 20-year nationwide survey of the health, economic, and social status of older Americans (50+), has added genetic information from consenting participants to its database. The HRS is a database for studying retirement and the baby boom generation. On March 15, genetic data from approximately 13,000 individuals were posted to dbGAP, the NIH's online genetics database. The data consist of approximately 2.5 million genetic markers from each person and are available for analysis by qualified researchers. Data were obtained from saliva samples collected from HRS participants since 2006. Specific information on the data can be found at ⟨http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000428.v1.p1⟩ and ⟨http://hrsonline.isr.umich.edu⟩. The NIA's scientific program seeks to understand the nature of aging and to extend the healthy, active years of life. For more information on research, aging and health, go to ⟨www.nia.nih.gov⟩.

**CNA Award for Operational Analysis.** This competitive award is open to all researchers and analysts outside of CNA and will be presented to the individual or team whose submitted work is judged as providing the most creative, empirically-based solution to a real-world problem or in support of a real-world decision. The winner of the CNA Award will be recognized at the CNA 70th Anniversary banquet in November 2012 and will receive $10,000. **Submissions must be received at ⟨CNAaward@cna.org⟩ by June 29, 2012.** To request instructions for submitting DoD classified work, or for answers to other questions, send an email to that same address, or call Ms. Patricia Sanders at 703-824-2038.

# CONSIDERATIONS FOR THOSE PROPOSING TOPICS AND PAPERS FOR JEP

Articles appearing in the journal are primarily solicited by the editors and associate editors. However, we do look at all unsolicited material. Due to the volume of submissions received, proposals that do not meet *JEP*'s editorial criteria will receive only a brief reply. Proposals that appear to have *JEP* potential receive more detailed feedback. Historically, about 10–15 percent of the articles appearing in our pages originate as unsolicited proposals.

## *Philosophy and Style*

The *Journal of Economic Perspectives* attempts to fill part of the gap between refereed economics research journals and the popular press, while falling considerably closer to the former than the latter. **The focus of *JEP* articles should be on understanding the central economic ideas of a question, what is fundamentally at issue, why the question is particularly important, what the latest advances are, and what facets remain to be examined.** In every case, articles should argue for the author's point of view, explain how recent theoretical or empirical work has affected that view, and lay out the points of departure from other views.

We hope that most *JEP* articles will offer a kind of intellectual arbitrage that will be useful for every economist. For many, the articles will present insights and issues from a specialty outside the readers' usual field of work. For specialists, the articles will lead to thoughts about the questions underlying their research, which directions have been most productive, and what the key questions are.

Articles in many other economics journals are addressed to the author's peers in a subspecialty; thus, they use tools and terminology of that specialty and presume that readers know the context and general direction of the inquiry. By contrast, **this journal is aimed at all economists, including those not conversant with recent work in the subspecialty of the author.** The goal is to have articles that can be read by 90 percent or more of the AEA membership, as opposed to articles that can only be mastered with abundant time and energy. Articles should be as complex as they need to be, but not more so. Moreover, the necessary complexity should be explained in terms appropriate to an audience presumed to have an understanding of economics generally, but not a specialized knowledge of the author's methods or previous work in this area.

The *Journal of Economic Perspectives* is intended to be scholarly without relying too heavily on mathematical notation or mathematical insights. In some cases, it will be appropriate for an author to offer a mathematical derivation of an economic relationship, but in most cases it will be more important that an author explain why a key formula makes sense and tie it to economic intuition, while leaving the actual derivation to another publication or to an appendix.

*JEP* does not publish book reviews or literature reviews. Highly mathematical papers, papers exploring issues specific to one non-U.S. country (like the state of agriculture in Ukraine), and papers that address an economic subspecialty in a manner inaccessible to the general AEA membership are not appropriate for the *Journal of Economic Perspectives*. Our stock in trade is original, opinionated perspectives on economic topics that are grounded in frontier scholarship. If you are not familiar with this journal, it is freely available on-line at <http://e-*JEP*.org>.

## *Guidelines for Preparing JEP Proposals*

Almost all *JEP* articles begin life as a two- or three-page proposal crafted by the authors. If there is already an existing paper, that paper can be sent to us as a proposal for *JEP*. However, given the low chances that an unsolicited manuscript will be published in *JEP*, no one should write an unsolicited manuscript intended for the pages of *JEP*. **Indeed, we prefer to receive article proposals rather than completed manuscripts.** The following features of a proposal seek to make the initial review process as productive as possible while minimizing the time burden on prospective authors:

- Outlines should begin with a paragraph or two that precisely states the main thesis of the paper.
- After that overview, an explicit outline structure (I., II., III.) is appreciated.
- The outline should lay out the expository or factual components of the paper and indicate what evidence, models, historical examples, and so on will be used to support the main points of the paper. The more specific this information, the better.
- The outline should provide a conclusion
- Figures or tables that support the article's main points are often extremely helpful.
- The specifics of fonts, formatting, margins, and so forth do not matter at the proposal stage. (This applies for outlines and unsolicited manuscripts).
- Sample proposals for (subsequently) published *JEP* articles are available on request.
- For proposals and manuscripts whose main purpose is to present an original empirical result, please see the specific guidelines for such papers below.

The proposal provides the editors and authors an opportunity to preview the substance and flow of the article. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article. After the editors and author(s) have reached agreement on the shape of the article (which may take one or more iterations), the author(s) are given several months to submit a completed first draft by an agreed date. This draft will receive detailed comments from the editors as well as a full set of suggested edits from *JEP*'s Managing Editor. Articles may undergo more than one round of comment and revision prior to publication.

Readers are also welcome to send e-mails suggesting topics for *JEP* articles and symposia and to propose authors for these topics. If the proposed topic is a good fit for *JEP*, the *JEP* editors will work to solicit paper(s) and author(s).

Correspondence regarding possible future articles for *JEP* may be sent (electronically please) to the assistant editor, Ann Norman, at <anorman@JEPjournal.org>. Papers and paper proposals should be sent as Word or pdf e-mail attachments.

## *Guidelines for Empirical Papers Submitted to JEP*

The *JEP* is not primarily an outlet for original, frontier empirical contributions; that's what refereed journals are for! Nevertheless, *JEP* occasionally publishes original empirical analyses that appear uniquely suited to the journal. In considering such proposals, the editors apply the following guidelines (in addition to considering the paper's overall suitability):

1. The paper's main topic and question must not already have found fertile soil in refereed journals. *JEP* can serve as a catalyst or incubator for the refereed literature, but it is not a competitor.
2. In addition to being intriguing, the empirical findings must suggest their own explanations. If the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. The empirical findings must be robust and thought provoking, but their interpretation should not be portrayed as the definitive word on their subject.
3. The empirical work must meet high standards of transparency. *JEP* strives to only feature new empirical results that are apparent from a scatter plot or a simple table of means. Although *JEP* papers can occasionally include regressions, the main empirical inferences should not be regression-dependent. Findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*.

The Journal of
Economic Perspectives
A journal of the American Economic Association

# 2012 Application/Renewal for Membership

**AMERICAN ECONOMIC ASSOCIATION**
2014 Broadway, Suite 305
Nashville, TN 37203
Ph. 615-322-2595 fax: 615-343-7590
Federal ID No. 36-2166945
www.vanderbilt.edu/AEA

| | |
|---|---|
| **RENEWING MEMBERS, ENTER ACCT. NUMBER & EXP. DATE** | **IF PAYING BY CREDIT CARD, PLEASE FILL OUT BELOW** |
| ACCOUNT NUMBER: | CARD NUMBER: |
| EXPIRATION DATE: | EXP DATE: CSC CODE: |

| | | |
|---|---|---|
| FIRST NAME: | MI: | LAST NAME: |
| ADDRESS: | | |
| CITY: | | STATE/PROVINCE: ZIP: |
| COUNTRY: | | □ Check here if non-US |
| PHONE: | | FAX: |
| PRIMARY FIELD OF SPECIALIZATION: | | |
| SECONDARY FIELD OF SPECIALIZATION: | | |
| EMAIL: | | □ Check here to exclude your email address from the public directory |

Please include my email address to receive:
□ Announcements about public policy affecting economists or the economics profession
□ Surveys of economists for research purposes
□ Commercial advertising

**MEMBERSHIP DUES** — Based on annual income. Please select one below.

| | | |
|---|---|---|
| □ Annual income of $70,000 or less | $20 | $ |
| □ Annual income of $70,000 to $105,000 | $30 | $ |
| □ Annual income over $105,000 | $40 | $ |

The AEA dues above include online access to all seven AEA journals.
For print or CD subscription(s) indicate preference below and add appropriate charge(s).

| **Journal** | | **Print** | **Int'l Postage*** | **CD*** | |
|---|---|---|---|---|---|
| AER | (7 issues, incl. P&P) | □ Add $20 | □ Add $25 | □ Add $15 | $ |
| AER Papers & Proceedings Only* | | □ Add $10 | n/a | n/a | $ |
| JEL | (4 quarterly issues) | □ Add $15 | □ Add $15 | □ Add $15 | $ |
| JEP | (4 quarterly issues) | □ Add $15 | □ Add $15 | □ Add $15 | $ |
| AEJ: Applied | (4 quarterly issues) | □ Add $15 | □ Add $15 | n/a | $ |
| AEJ: Policy | (4 quarterly issues) | □ Add $15 | □ Add $15 | n/a | $ |
| AEJ: Macro | (4 quarterly issues) | □ Add $15 | □ Add $15 | n/a | $ |
| AEJ: Micro | (4 quarterly issues) | □ Add $15 | □ Add $15 | n/a | $ |

* Int'l postage applies only to print journals mailed outside of the U.S. No additional postage is required for CDs or the AER Papers and Proceedings.

## AEA Journals via JSTOR online

| | | |
|---|---|---|
| JSTOR | □ Add $16 | $ |

| | | | | | |
|---|---|---|---|---|---|
| | | | | Sub Total | $ |
| Check One: | □ 1 Year | □ 2 Years | □ 3 Years | TOTAL AMOUNT | $ |

Make checks payable to: American Economic Association.
Must be drawn on a US bank.
Apply online at http://www.aeaweb.org/membership.php

Payments must be made in advance. We accept checks (in US dollars only, with correct coding for processing in US banks) and credit cards; online or by faxing or mailing the application. Please choose one method; it is the Association's policy NOT TO REFUND dues.

# 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 $15.00, you can receive this journal in print. Change of address notice must be received at least six weeks prior to the publication month.

*The Journal of*

# *Economic Perspectives*

Spring 2012, Volume 26, Number 2

## Symposia

### *100 Issues of JEP*

### *International Trade*

## Articles

## Features

AMERICAN
ECONOMIC
ASSOCIATION