PAPERS AND PROCEEDINGS OF THE
ANNUAL BANK CONFERENCE
ON
DEVELOPMENT ECONOMICS

Supplement to THE WORLD BANK ECONOMIC REVIEW

2015 Annual Bank Conference on Development Economics
“Productivity, Growth, and the Law”
June 15–16, 2015 Mexico City

Invited Papers
Gains from Foreign Direct Investment: Macro and Micro Approaches
Laura Alfaro
Should Human Rights Law Play a Role in Development?
Eric A. Posner

Shorter Papers on:
Growth and Productivity
Entrepreneurship and Access to Credit
Infrastructure, Competition, and Growth
Labor Laws, Jobs, and Education
Leadership, Law, and Development
Erratum

Editors Kaushik Basu and Andrew Foster
THE WORLD BANK
ECONOMIC REVIEW

EDITORS
Eric Edmonds, Dartmouth College
Nina Pavcnik, Dartmouth College

CO-EDITORS
Francisco H. G. Ferreira, World Bank
Karl Hoff, World Bank
Leora Klappe, World Bank
Aart C. Kraay, World Bank
David J. McKenzie, World Bank
Luis Servén, World Bank

The editorial team would like to thank former editor Andrew Foster for overseeing the review process of the articles in this issue.

ASSISTANT TO THE EDITOR
Marja Kuiper

EDITORIAL BOARD
Harold H. Alderman, International Food Policy Research Institute
Chong-En Bai, Tsinghua University, China
Pranab K. Bardhan, University of California, Berkeley, USA
Kashish Basu, World Bank
Thorsten Beck, Cass Business School, City University London, UK
Johannes van Biesebroeck, K.U. Leuven, Belgium
Maureen Cropper, University of Maryland, USA
Asli Demirguc-Kunt, World Bank
Quy-Toan Do, World Bank
Frédéric Docquier, Catholic University of Louvain, Belgium
Eliana La Ferrara, Università Bocconi, Italy
Francisco H. G. Ferreira, World Bank
Augustin Kwasi Fou, United Nations University, WIDER, Finland
Paul Glewwe, University of Minnesota, USA
Jeremy Magnud, University of California, Berkeley, USA
William F. Maloney, World Bank
David J. McKenzie, World Bank
Jaime de Melo, University of Geneva, Switzerland
Ugo Panizza, The Graduate Institute Geneva
Nina Pavcnik, Dartmouth College, USA
Vijayendra Ran, World Bank
Martin Ravallion, Georgetown University, USA
Jaime Saavedra-Chanduvi, World Bank
Claudia Sepúlveda, World Bank
Dominique Van De Walle, World Bank
Christopher M. Woodruff, University of California, San Diego, USA

The World Bank Economic Review is a professional journal used for the dissemination of research in development economics broadly relevant to the development profession and to the World Bank in pursuing its development mandate. It is directed to an international readership among economists and social scientists in government, international agencies, universities, and development research institutions. The Review seeks to provide the most current and best research in the field of quantitative development policy analysis, emphasizing policy relevance and operational aspects of economics, rather than primarily theoretical and methodological issues. Consistency with World Bank policy plays no role in the selection of articles.

The Review is managed by one or two independent editors selected for their academic excellence in the field of development economics and policy. The editors are assisted by an editorial board composed in equal parts of scholars internal and external to the World Bank, World Bank staff and outside researchers are equally invited to submit their research papers to the Review.


Instructions for authors wishing to submit articles are available online at www.wber.oxfordjournals.org.

Please direct all editorial correspondence to the Editor at wber@worldbank.org.
Editors’ Introduction

Invited Papers
Gains from Foreign Direct Investment: Macro and Micro Approaches
Laura Alfaro

Should Human Rights Law Play a Role in Development?
Eric A. Posner

Growth and Productivity
On the Origins and Development of Pakistan’s Soccer-Ball Cluster
David Atkin, Azam Chaudhry, Shamyla Chaudry,
Amit K. Khandelwal, Tariq Raza, and Eric Verhoogen

Cross-Sector Misallocation with Sector-Specific Distortions
Julio Leal

From Firm Productivity Dynamics to Aggregate Efficiency
Bernabe Lopez-Martín

Entrepreneurship and Access to Credit
Psychometrics as a Tool to Improve Credit Information
Irani Arráiz, Miriam Bruhn, and Rodolfo Stucchi

Unitary or Noncooperative Intrahousehold Model? Evidence from Couples in Uganda
Nathan Fiala and Xi He

Opportunity versus Necessity: Understanding the Heterogeneity of Female Micro-Entrepreneurs
Gabriela Calderon, Leonardo Iacovone, and Laura Juárez

Infrastructure, Competition, and Growth
Highways and Spatial Location within Cities: Evidence from India
Ejaz Ghani, Arti Grover Goswami, and William R. Kerr

ICT Use, Competitive Pressures, and Firm Performance in Mexico
Leonardo Iacovone, Mariana Pereira-López, and Marc Schiffbauer
Do Deep Trade Agreements Boost Vertical FDI?  
_Alberto Osnago, Nadia Rocha, and Michele Ruta_  
S119

**Labor Laws, Jobs, and Education**

Exporters, Engineers, and Blue-collar Workers  
_Irene Brambilla, Daniel Lederman, and Guido Porto_  
S126

The Causal Impacts of Child Labor Law in Brazil: Some Preliminary Findings  
_Caio Piza and André Portela Souza_  
S137

The Effect of Publicly Provided Health Insurance on Education Outcomes in Mexico  
_Carlo Alcaraz, Daniel Chiquiar, María José Orraca, and Alejandrina Salcedo_  
S145

**Leadership, Law, and Development**

The Importance of Legitimacy  
_Robert Akerlof_  
S157

Are Politically Connected Firms More Likely to Evade Taxes?  
Evidence from Tunisia  
_Boo Rijkers, Hassen Aroui, and Leila Baghdadi_  
S166

Courting Economic Development  
_James R. Brown, J. Anthony Cookson, and Rawley Z. Heimer_  
S176

Erratum  
S188
Editors’ Introduction

This volume contains the Papers and Proceedings of the 26th Annual Bank Conference on Development Economics (ABCDE) held on June 15–16, 2015, in Mexico City. The theme of the conference was “Productivity, Growth, and the Law,” and it was hosted jointly by the Bank of Mexico and the World Bank.

The 2015 ABCDE Organizing Committee was composed of Kaushik Basu (World Bank, Chair), Asli Demirgüç-Kunt and Indermit Gill (World Bank), Alberto Torres and Laura Juárez (Bank of Mexico), and Andrew Foster (Brown University).

The committee chose the topics of the ABCDE sessions and selected the papers to be presented at those sessions. A call for papers appeared on the World Bank’s Web site and other academic outlets six months in advance of the conference. In this volume, a subset of the selected papers are being published, along with the keynote addresses of two well-known academics. Comments and discussions are not being published, but most of the comments, in the form of PowerPoint presentations, can be downloaded from the ABCDE Web site (www.worldbank.org/abcde2015).

The guidelines for choosing which papers would be published in the World Bank Economic Review (WBER) Papers and Proceedings issue differ from those for the regular issues of the WBER. First, the papers are edited but are not subject to the formal referee process that regular submissions to the WBER must pass—although papers can be rejected if after a thorough reading it is decided that they lack adequate merit. Second, the length of the paper is controlled.

Third, papers are more exploratory and looser than formal articles on original research.

Finally, a great deal of teamwork was required to produce a volume of this size as part of the WBER series. The editors wish to thank the authors for their cooperation in meeting deadlines and all the other people who made this volume possible. In addition, we are deeply grateful to Gabriela Calderón and Laura Juárez (Bank of Mexico), and Claudia Sepúlveda (World Bank), who supervised and organized the conference program.
Gains from Foreign Direct Investment: Macro and Micro Approaches

Laura Alfaro

This paper discusses the importance of an “integrated approach” to the study of the effects of FDI on host countries. Macro-level work that examines countries at different stages of development and institutional capacity is needed to surface the role of local conditions and absorptive capacities; micro-level work, that is firm-level data in developed as well as developing nations, to understand the mechanisms that impart substance to the anticipated benefits; and theoretical work to guide the analyses. The paper summarizes likely motives for foreign direct investment and potential effects of FDI on local economies as well as recent findings from the macro literature on the role of complementarities between FDI and local policies, conditions, and institutions and summarizes new efforts to understand the micro mechanisms and channels by which host countries can benefit from multinational activity, within and between firm productivity increases.

I. Introduction

Foreign direct investment, as it embodies technology and know-how as well as foreign capital, has in recent decades come to be seen as an engine of growth and development. Understanding the effects of foreign direct investment (FDI) and multinational production has consequently become a critical topic in academic and policy circles. Knowledge spillover, backward and forward linkages with local firms, the potential for technology transfer, new processes, improved managerial skills, employee training, and access to international production networks and markets all have the potential to increase productivity and output, create employment, help diversify exports, and transform the production structure of the economy by fueling growth and fostering development.

Laura Alfaro is a professor at Harvard Business School, Morgan 263, Boston, MA 02163, USA. Her e-mail address is lalfaro@hbs.edu. The author was Minister of Planning and National Economic Policy for Costa Rica, 2010–12.

The academic literature on foreign direct investment has been surveyed many times. See, e.g., Alfaro et al. (2009), Kose et al. (2009), Harrison and Rodríguez-Clare (2010), Alfaro and Johnson (2012), Antrás and Yeaple (2014), Melitz and Redding (2014), and Alfaro (2015) for recent surveys of determinants, effects, spillover channels, and empirical findings.

© The Author 2016. Published by Oxford University Press on behalf of the International Bank for Reconstruction and Development / THE WORLD BANK. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
In the face of such a panoply of potential benefits, developed as well as developing countries have not only substantially reduced barriers to but also offered special incentives to attract, foreign direct investment. Yet the impact of the increasing role of FDI on host economies has been difficult to assess. Indeed, empirical evidence for FDI generating the expected positive effects is ambiguous at the micro- as well as macro-levels.

In this article, I discuss the value of an “integrated approach” to the study of the effects of FDI on host countries. Macro-level work that examines countries at different stages of development and institutional capacity is needed to surface the role of local conditions and absorptive capacities; micro-level work, that is firm-level data in developed as well as developing nations, to understand the mechanisms that impart substance to the anticipated benefits; and theoretical work to guide the analyses.

The macro-level data needed to reveal the effects on FDI of varying local conditions and institutional frameworks necessarily spans multiple countries and sufficiently long periods of time. Such data are essential to studies that would examine the broader environmental milieu in which the benefits of FDI gestate. But aggregate data cannot shed light on the channels through which capital controls affect the economy at the micro-level. To identify these, and the mechanisms that account for delivery of the benefits of FDI, requires marrying the firm- and plant-level data of developed and developing nations.

Availability of data thus becomes a critical issue. Census firm-level data are available in few countries and very few developing countries. Rarely, moreover, can data be accumulated over long periods or for similar periods across countries. Data may also entail usage restrictions (confidentiality, data restrictions, etc.) and lack desired variables. The workaround for data limitations is to exploit different sources of data, such as firm registries.

Work aimed at affording an understanding of the effects of FDI needs to be informed as well by theory, not only to guide empirical work but also to illuminate policy implications. For example, positive gains from multinational activity are often attributed to knowledge spillover and technology externalities transmitted from foreign to domestic firms (i.e., between-firm benefits). But multinational production can also precipitate more intense competition in product and factor markets and reallocation of resources from domestic to multinational, and from less productive to more productive domestic firms (i.e., between firm and market selection). Distinguishing spillover from market reallocation is critical to assessing the economic impact of FDI and setting corresponding economic policies. But the effect of multinational selection, knowledge spillover, and market reallocation on aggregate productivity and the welfare impact of multinational production cannot be disentangled simply by examining the relationship between multinational production and host-country average productivity, as all channels predict a positive relation. To separate these effects requires a theoretical framework that explicitly incorporates them (Alfaro and Chen 2013).

What has come to light thus far from the pursuit of an integrated approach to establishing the benefits of FDI is that they are conditional on the presence of
complementary conditions that facilitate their absorption by firms, regions, and countries. These benefits include the policy environment, quality of local institutions and financial markets, sector characteristics and market structure, and spatial co-location (Alfaro et al. 2004, 2010; Alfaro and Hammel 2007; Alfaro and Charlton 2013; Alfaro and Chen 2014). The sources of gains also differ within and between firms (Alfaro and Chen 2012a, 2012b, 2013). Implications of the findings of integrated macro- and micro-level studies that investigate the role of local conditions and sources of potential gains from multinational production, being critical to the design of economic policy, can usefully inform both policy and academic debates on FDI.

Section 2 summarizes likely motives for foreign direct investment and potential effects of FDI on local economies as well as recent findings from the macro literature on the role of complementarities between FDI and local policies, conditions, and institutions. Data issues and the challenges of micro-level work are discussed in section 3. Section 4 examines the role of theory in complementing recent work, and summarizes new efforts to understand the mechanisms and channels by which host countries can benefit from multinational activity, within and between firm productivity increases. Section 5 concludes.

II. Motivations for and Effects of Multinational Activity: The Role of Local Conditions

A multinational enterprise (MNE), or multinational or transnational corporation (MNC and TNC, respectively), owns and controls production facilities or other income-generating assets in at least two countries. Inaugurating a green-field operation (i.e., constructing a new production facility) or acquiring control of an existing local firm constitutes direct investment by a foreign entity. In such relationships, source country parents control productive facilities of host country affiliates. It is the former’s partial or complete control over the latter’s activities that distinguishes FDI from portfolio investment.²

Figure 1 shows the increase in multinational activity in developed and developing countries over the past three decades. MNEs vary in size, sector, and structure. Firms variously invest abroad to serve a market directly or to export; to access inputs, raw materials, or labor; to improve operational efficiency; or to impede competitors’ acquisition of strategic assets (see Desai 2009). But do these ends offset such exigencies of doing business in another country as costs associated with communication and transport and stationing personnel abroad, customs and language barriers, and exclusion from local business and government networks? How can a foreign firm offset the local firm’s advantage of superior knowledge of the market, legal and political systems, language, and culture?

Hymer (1960) proposed a broadly accepted framework, derived from the industrial organization literature, in which real (as opposed to financial) factors

². See Alfaro et al. (2014) for definitions and stylized facts regarding foreign capital flows over the past four decades.
explain the location decisions of multinational firms. After all, if lower cost of capital were the only advantage, why wouldn’t a foreign investor simply make a portfolio investment and forego the headaches of operating in a different political, legal, and cultural milieu? Firms engage in FDI not per se because of differences in the cost of capital—as in the case of portfolio investment—but because certain assets are worth more under foreign control, which allows firms to compete in foreign environments.  

This view credits the genesis of FDI to the possession of some asset, such as technology or know-how, that constitutes a significant gain for the host country, which, in turn, suggests that FDI can play an important role in accelerating and modernizing a country’s economic growth.  

It is FDI’s embodiment of capital, technology, and know-how that affords the potential for host countries to benefit from direct knowledge transfer through partnerships with, and opportunities to learn from the innovation and experience of, parent firms, and interaction and movement in labor markets.  

Here, the findings of the macro literature on the effects of FDI on growth offer important insights. Evidence not of an exogenous positive effect of FDI on

---

3. Hymer’s approach was subsequently refined by, among other authors, Kindleberger (1969) and Caves (1974), and eventually rechristened the OLI (ownership advantage, location factor, internalization) framework in Dunning’s (1981) work. For a description, see Antrás and Yeaple (2014).

4. Note, however, that there might be offsetting costs to the host country. The proprietary asset or technology provides its owner with some market power or cost advantage over indigenous producers.
economic growth but rather of positive effects conditional on local conditions and policies\(^5\) reflects Rodrik and Rosenzweig’s (2009) observation that “Appropriate development policies typically exhibit high degrees of complementarity.” The FDI literature has identified important roles for the policy and institutional environments (Alfaro et al. 2007, 2008; Balasubramanayam et al. 1996); human capital (Borensztein et al. 1998); local financial markets (Alfaro et al. 2004, 2010); sector characteristics (Alfaro and Charlton 2013); and market structure (Alfaro et al. 2010). Harrison and Rodriguez-Clare (2010) similarly emphasize the relevance of complementary aspects of a trade policy regime, such as labor-market policies and ease of entry and exit, and Kose et al. (2009), who identify macroeconomic and structural policies that need to be in place for countries to reap the benefits of financial globalization, emphasize that capital account policies need to be viewed as part of a much broader set of policies.

Borensztein et al. (1998), using a data set of FDI flows from industrialized countries to 69 developing countries, finds FDI to be an important vehicle for transferring technology and promoting growth only when the host country has a minimum threshold of human capital, and Xu (2000), using data on US multinationals, finds that most developing countries do not meet such a threshold. These results suggest strong complementarities between FDI and human capital, and that FDI is more productive than domestic investment and important as a vehicle for technology transfer only when the host country has a minimum threshold stock of human capital.

Alfaro et al. (2004) argue, on the basis of studying the role local financial institutions play in channeling the contributions of FDI to economic growth, that underdevelopment of local financial markets can limit an economy’s ability to exploit the potential of FDI spillovers. The results of the analysis of growth on FDI to GDP (and various controls variables) indicate that FDI, on its own, does not exert a robust positive impact on growth. When the interaction term is included, however, the regression results become positive and significant, leading the authors to posit that realization of the positive benefits of FDI is contingent on a country’s possession of a strong financial sector.

Alfaro and Charlton (2013), using industry-level evidence derived from data on OECD countries, show the relation between FDI and growth to be stronger for industries that are more reliant on external financing. These results, as well as being consistent with the existing macro literature on the hypothesized benefits of FDI, are further evidence of important cross-industry differences in the effects of FDI.

Scarcity of capital for new investment is among the reasons policy makers advance for promoting foreign investment in developing countries. This argument is based on the assumption that additional capital is made available by foreign investors that establish new enterprises in local markets. Kindleberger (1969),

\(^5\) That firm-level panel studies tend to cover specific and quite different types of countries (transition, developing, emerging, industrialized) as well as different periods makes it difficult to understand the role of country specific conditions.
Graham and Krugman (1995), and Lipsey (2002), however, present evidence that investors often fail to fully transfer capital on taking control of a foreign company, tending instead to finance a significant share of their investment in the local market. Bilir, Chor, and Manova (2014) show that host countries with more developed financial markets attract more multinational entry. Many foreign investors, moreover, have hedged against rising exchange-rate volatility by borrowing in local capital markets. Foreign firms that borrow heavily from local banks instead of bringing scarce capital from abroad may exacerbate local firms’ financing constraints by crowding them out of domestic capital markets.

The importance of well-functioning financial institutions to economic development has been examined extensively in the literature. Researchers have shown that such markets, by reducing transaction costs, ensure that capital is allocated to the projects that yield the highest returns, thereby enhancing growth rates. As McKinnon (1973) observed, the development of capital markets is “necessary and sufficient” to foster “adoption of best-practice technologies and learning by doing,” which is to say, that limited access to credit markets restricts entrepreneurial development. If entrepreneurship fosters greater assimilation and adoption of best technological practice as made available by FDI, then absence of well-developed financial markets limits FDI’s potential positive externalities. Some local firms might be able to finance new requirements with internal financing, but the greater the technology-knowledge gap between current practice and new technology, the greater the need for external financing, which, in most cases, is restricted to domestic sources. The creation of linkages may also depend on the access of funds by potential suppliers to enter the market or upgrade their products.

In order to formalize one mechanism through which the trickle down effect of FDI depends on the extent of local conditions, Alfaro et al. (2010) develop a theoretical model in which the presence of positive linkages depends on the local financial sector’s degree of development. A small open economy is modeled in which final goods production is carried out by foreign and domestic firms that compete for skilled and unskilled labor and intermediate products. An entrepreneur that would operate a firm in the intermediate goods sector must debut a new variety of intermediate good, development of which requires upfront capital investment. The more developed local financial markets, the easier it is for credit-constrained entrepreneurs to start firms. As the increasing variety of intermediate

6. The industrial organization literature, as noted above, suggests that firms engage in FDI not because of differences in cost of capital, but because certain assets are worth more under foreign than under local control. Were lower cost of capital the only advantage a foreign firm enjoyed over domestic firms, why a foreign investor would take the trouble to operate a firm in a different political, legal, and cultural milieu rather than simply make a portfolio investment would remain unexplained.

7. Hirschman (1958) argues that linkage effects are realized when one industry, by easing conditions of production, facilitates the development of another, thereby accelerating the pace of industrialization. He further argues that in the absence of linkages, foreign investments could have limited or even negative effects on (so-called enclave) economies. For further discussion and evidence, see Alfaro and Rodriguez-Clare (2004).
goods precipitates positive spillovers to the final goods sector, financial markets allow backward linkages between foreign and domestic firms to transform into FDI spillovers. Illustrating the implication of the model for realistic parameters, the authors find that for the same share of foreign production in total output, countries with more developed financial markets exhibit higher growth rates; increases in the amount of FDI (or the technology gap between foreign-owned firms and domestically owned firms) imply also in additional growth effects generated in the financially well-developed countries close to three times higher those financially poorly developed countries. Moreover, the authors find that differences in growth rates increase when domestic firms and MNEs are substitutes rather complements and by varying the relative skill ratios while assuming that MNEs use skilled labor more intensively.

That the literature on complementarities has found some countries to lack the preconditions requisite to reaping the potential benefits of FDI may help explain the ambiguity in findings on the relationship between FDI and growth. Spillovers from foreign to domestic firms depend on the domestic firms’ ability to respond successfully to new entrants, new technology, and new competition. That success is, to some extent, determined by local characteristics such as levels of human capital and development of local financial markets as well as by the overall institutional level of the country. 8 Weaknesses in these areas may reduce the capacity of domestic industries to absorb new technologies and respond to the challenges and opportunities presented by foreign entrants.

III. Micro Patterns and Implications

A central challenge facing the FDI literature has been the absence of a global source of firm-level data. Researchers have instead used multinational (MNC) activity at the industry level or aggregate FDI flows from balance-of-payments statistics as a proxy for foreign firm activity or country-level firm data. Cross-country empirical investigations at the firm level are impeded by the lack of high-quality data sets, which tend to be limited to advanced, and not be comparable across countries, and infrequency of economic censuses owing, especially in poor countries, to their high cost and institutional restrictions. Many sources, moreover, are encumbered by confidentiality restrictions that make it difficult to compile the data. No single institution having the capacity or resources to undertake the primary collection of census or census-like data over a wide range of countries and periods, researchers have taken recourse to other sources, such as business compilations (e.g., registries and tax sources) or surveys.

The data in WorldBase compiled by Dun and Bradstreet, relative to other international datasets, represent a quite large number and variety of sources (e.g.,

8. Examining, in an empirical framework, different explanations for the dearth of capital flows, particularly of FDI, from rich to poor countries, termed the Lucas paradox, Alfaro et al. (2007, 2008) find institutional quality to be the most important explanatory variable.
partner firms, telephone directory records, Web sites, self-registration). Using the WorldBase firm-level data set that includes location, ownership, and sector (at the four-digit level) for each of more than 650,000 multinational subsidiaries in 400 industries and 90 countries, Alfaro and Charlton (2009), consistent with the existing literature, find the bulk of multinational activity to occur between the rich nations of the world, but other patterns to differ, some plant-level findings of particular significance warranting reconsideration of the conventional wisdom in certain areas.

For analytical simplicity, FDI is usually classified as horizontal or vertical. Firms engage in horizontal FDI when they replicate a subset of their activities or processes in another country, in other words, when production is duplicated in an offshore venue (Markusen 1984; Markusen and Venables 2000). Firms engage in vertical FDI when they fragment production by function, that is, when, often motivated by cost considerations arising from factor cost differences, they break up the value-added chain (Helpman 1984).10 Models that assume low transport costs and comparative advantage are consistently rejected in favor of models that treat market access issues. Our results suggest that data limitations have led the literature to systematically underestimate vertical FDI, which our data set reveals to be far more prevalent than previously thought.

Combining four-digit sector level information and input-output tables to distinguish horizontal from vertical FDI, Alfaro and Charlton (2009) classify a horizontal subsidiary as a plant in the same sector code as the foreign owner parent, a vertical subsidiary as a plant that produces in sectors that input to the foreign parent’s product.

At the four-digit level, many foreign subsidiaries in the same two-digit industry as their parents occupy sectors related to highly specialized inputs to the parents’ production, which is to say, that most vertical activity is in sectors close to the parent firm (i.e., intra-firm FDI). Differences from previous findings emerge because much vertical FDI, being north-north, has been assumed to be market seeking (horizontal) when, in fact, firm-level data indicates the relationships to be vertical, their true nature being visible only at the four-digit level, and hence being missed at the two-digit level. Intra-industry vertical subsidiaries, being generally located in sectors related to high-skill inputs, largely tend also to be located in high-skill countries.

These differences matter for how FDI may impact countries. In particular, different motivations for FDI differ on how multinational activity affects factor incomes within and across countries. Horizontal FDI, being a substitute for trade, multinational activity may raise income in each country without necessarily changing its distribution; while for vertical FDI, a complement to trade, multinational activity may reduce absolute wage differences across countries and alter

9. See Alfaro and Charlton (2009) for a detailed discussion of WorldBase data and comparisons with other data sources.

10. For empirical evidence on the different types, see Alfaro and Charlton (2009).
How MNEs respond, relative to local firms, to crisis, and how MNEs’ performance is linked across the countries in which they operate, have been little studied. Alfaro and Chen (2012a,b) addressed this deficiency by investigating the effects of foreign ownership on resilience to negative shocks. The authors exploited the global scope and considerable heterogeneity of the recent global financial meltdown to explain the role of FDI in microeconomic performance. Specifically, they examined firms’ micro responses to, and differences in performance during, that crisis. The effects of foreign ownership were disentangled from other effects using the D&B data set, and observable and unobservable differences between foreign subsidiaries and local establishments controlled for by matching each foreign subsidiary with a local establishment with similar characteristics operating in the same country and industry. The effect of foreign ownership is inferred from divergences in performance. Comparing the effect of foreign ownership between the noncrisis (2005–2007) and crisis (2007–2008) periods enabled the authors to identify the role played by vertical and financial links in increasing the resilience of foreign subsidiaries to negative demand and financial shocks.

Foreign subsidiaries performed better than locally controlled plants with similar characteristics during the crisis, but not during the noncrisis. Locally controlled establishments were also outperformed during the crisis period by foreign subsidiaries with strong vertical, but not by those with horizontal, production links with their parent firms. This pattern as well was not observed during the noncrisis years. Superior performance by foreign subsidiaries operating in industries with greater intra-firm financial links was also observed only during the crisis period, and especially in host countries in which credit conditions had worsened.

These findings have important implications for academic and policy debates on the role of foreign direct investment. Notwithstanding concerns, prompted by a number of studies that have found FDI to have a significant negative effect on plant survival and stability (see, e.g., Bernard and Jensen 2007), that FDI is more volatile and occasions greater vulnerability, especially during crises, than domestic investment, Alfaro and Chen (2013) find vertical production and financial links between foreign subsidiaries and parent firms to have the potential to mitigate the impact of crises on host countries.

IV. THEORY AND EVIDENCE: SPILLOVERS, SELECTION, AND REALLOCATION

Alfaro and Chen (2013), exploring the mechanisms by which an economy responds to multinational production, disentangle the roles of knowledge

spillover roles, selection, and reallocation (within- and between-firm effects) in determining the aggregate impact of multinational production on host-country productivity. 12 Although research abounds on the knowledge spillover effect of multinational firms, the role of market reallocation, and the different ways in which market reallocation and knowledge spillover influence potential gains from multinational competition, have been little studied, and the relative importance of spillovers and market selection seldom analyzed.

Although all imply a positive relation between multinational production and host-country productivity, the implications and economic causalities of these mechanisms differ sharply. Positive externalities that accrue to knowledge spillover enhance the productivity of individual domestic firms, consequent market reallocation can contribute to the contraction of domestic industries. If, moreover, knowledge spillover is a significant source of productivity gains from multinational production, special treatment of foreign firms, often tendered by host countries in the form of tax breaks and financial incentives, may be justified. But if productivity increases are instead attributable to market reallocation, public resources may be better spent improving domestic market conditions including conditions of labor supply and credit access, and eliminating regulatory barriers to facilitate gains from competition and reallocation of resources.

It is difficult, however, to separate these effects in the absence of a theoretical framework that explicitly incorporates them. With this purpose, Alfaro and Chen (2013) develop an empirical framework based on a standard model of monopolistic competition and heterogeneous firms augmented to include potential externalities of multinational production. Using this micro theoretical foundation that captures these distinct aspects of multinational production, the authors develop an empirical strategy to distinguish their relative importance, while accounting for the self-selection of multinational firms.

The authors empirically evaluate the predictions using a large, cross-country, panel data set drawn from Orbis that contains comprehensive financial, operating, and ownership information for more than one million public and private manufacturing companies for the 2002–2007 period. The data set’s two notable strengths, broad cross-country coverage and detailed ownership (majority and wholly owned) information, enable us to identify multinational production, and explore the heterogeneous effect of foreign investment, across countries.

The empirical analysis suggests that domestic markets experience both knowledge spillover and factor reallocation as a consequence of multinational production. As entry by multinational firms raises the cutoff productivity of domestic firms, the least productive domestic firms exit. New multinational production

12. An emerging literature emphasizes the productivity effect of resource allocation across establishments (see Hsieh and Klenow [2009] and Alfaro, Charlton, and Kanczuk [2009]; Alfaro and Chari [2014]). The role of reallocation, although under-emphasized in evaluating gains from multinational production, is well established in determining the productivity gains from trade liberalization (see Melitz 2003; Melitz and Redding 2014).
also occasions an increase in the minimum revenue of continuing domestic firms, which implies an increase in fixed production cost and capital price. That the estimates show a significant decrease in aggregate price suggests increased competition and market reallocation. These results suggest that sensible policy should aim to facilitate gains from competition and the reallocation of resources by improving domestic conditions including credit access and labor supply (particularly skilled labor), while eliminating regulatory barriers.

Externalities and spillovers are, of course, by their very nature, difficult to measure. Quality upgrades, worker training, and improvements in the business environment and organizational practices are among other factors that can positively affect host economies.\textsuperscript{13} MNEs may also cluster, worldwide, to benefit from interaction with one another.

Regional and urban economists and economic historians have long recognized the agglomeration of economic activity as one of the salient features of economic development. An extensive body of research examines the distribution of population and production across space and economic characteristics and effects of spatial concentrations. Understanding the emerging spatial concentrations of multinational production around the world, and the driving forces behind these new concentrations relative to those of their domestic counterparts, is crucial to designing and improving policies.

Alfaro and Chen's (2014) investigation of the patterns and determinants of the global economic geography of multinational firms suggests that emerging offshore clusters of multinationals are not simply a reflection of domestic industrial clusters. That is, within a host country, multinationals and their domestic counterparts follow different agglomeration patterns. The location decisions of MNEs reflect not only such location fundamentals as market access (to avoid trade costs) and comparative advantage (to acquire abundant factors at lower cost) but also agglomeration economies that emphasize the benefits of geographic proximity between firms including lower transport costs between input suppliers and final good producers (vertical linkages), labor-market and capital-good-market externalities that reflect MNEs’ high capital- and innovation-intensity, and technology diffusion. Multinational entrants also exhibit a stronger propensity to cluster with incumbent multinationals than with incumbent local plants, especially in the presence of strong capital-good-market externalities and technology diffusion benefits. In particular, the authors find multinational foreign subsidiaries to be more agglomerative than domestic plants in capital-intensive, skilled-labor-intensive, and R&D-intensive industries.

These results are consistent with increasing segmentation of activities within the boundaries of multinational firms, in particular, the market-seeking and input-sourcing focuses of offshore production and emphasis of headquarters on such knowledge-intensive activities as R&D, management, and services.

\textsuperscript{13} See Bao and Chen (2013) for evidence of upgrading in anticipation of foreign competition.
V. Concluding Comments

This paper combines macro- and micro-level approaches, informed by theory, to examine the effects of foreign direct investment (FDI) on host (particularly developing) countries’ economies. An understanding is sought of the role of complementary local conditions conducive to reaping the benefits of FDI, and of the mechanisms, namely, within- and between-firm selection and market reallocation, by means of which the desired effects can be realized.

Insights gleaned from new research into the role of complementarities and the mechanisms by which FDI induces growth begin to reconcile the ambiguous evidence on FDI’s ability to generate growth in host countries. The research on complementarities has shown FDI’s positive impacts to not be exogenous but rather conditional on certain local conditions. Research into the mechanisms and channels by which FDI can generate positive externalities goes a step further in illustrating how complementarities—a competitive environment that assures that market share is allocated to the most productive firms, or developed financial markets that ensure that vertical supply relations develop into meaningful linkages—can act as “absorptive capacities” to facilitate realization of the benefits of FDI.

If knowledge spillover is the primary source of productivity gains, special treatment for foreign firms, often in the form of tax breaks and financial incentives, may be justified and sufficient. But if productivity gains also arise from market reallocation, improving domestic market conditions including labor supply and credit access, and eliminating barriers to reallocation may also be important. Improving local conditions and accelerating growth are, in turn, likely to attract foreign firms.

What are the policy implications of the “integrated approach”? FDI can play an important role in economic growth, most likely via suppliers. But local conditions matter and can limit the extent to which the benefits of FDI materialize. Sensible policies might eliminate barriers that prevent local firms from establishing adequate linkages; improve local firms’ access to inputs, technology, and financing; and streamline the procedures associated with selling inputs. Research suggests that improving domestic conditions as well should have the dual effect of attracting foreign investment and enabling host economies to maximize the benefits of foreign investment.

References


Many human rights advocates believe that development agencies—agencies that define their mission as providing economic and technical aid to impoverished countries—should be required to respect and promote human rights law. This style of human rights imperialism should be resisted. While development agencies should obviously comply with domestic law and try to promote good rather than bad outcomes, there is no benefit in holding them to human rights law. Human rights law was designed for states, not for NGOs, and how it would be applied to NGOs is far from obvious. Because of the ambiguity and vast scope of human rights law, the practical effect of these proposals would be to add another layer of bureaucracy to development projects while subjecting those projects to scrutiny by lawyers with little to guide them but their intuitive notions of right and wrong. JEL Codes: O15, O20

Many human rights advocates believe that development agencies should evaluate their projects from the standpoint of human rights law. The most recent flare-up occurred in connection with the World Bank’s draft Environment and Social Framework (2014). The purpose of this document is to ensure that development projects do not cause excessive harm to people or the environment. A project to build a road, for example, may displace people who live in its path and cause harm to the surrounding environment. The document encourages the bank and borrowers to ensure that the affected people are compensated and consulted; and that harm to the environment is mitigated or repaired.

Human rights advocates complain that the Framework does not recognize human rights law as a binding constraint on development projects. In a letter to the World Bank (Alston et al. 2014), a number of UN officials argue that the Bank should do more for human rights.

As the Bank seeks to revise and adapt its Safeguards approach [in its draft Environmental and Social Framework] to the challenges of the twenty-first century, we believe that it is imperative that the standards should be premised on a recognition of the central importance of respecting and promoting human rights. But there is no such provision in the current draft.

Human Rights Watch (2015a) expresses a similar view:

Eric A. Posner is a professor at the University of Chicago Law School. His e-mail address is eric_posner@law.uchicago.edu.
Human Rights Watch is disappointed that the primary reference to human rights in the draft framework is in the non-binding vision statement and urges you to remedy this. To suggest that human rights are merely visionary suggests that the World Bank views human rights as non-binding, undermines the international human rights framework, and does not remedy the long-standing problem that the bank does not analyze and address potential adverse human rights impacts of its activities.

The critics argue both that the Bank is legally obligated to incorporate human rights law into the Framework and that, by failing to do so, the Bank has missed an opportunity to help advance the cause of human rights.\(^1\)

The debate raises a broader issue of whether development agencies—including government agencies, international organizations, and NGOs—should promote human rights law in the course of financing development projects. There are actually two intertwined issues. The first is whether development agencies are legally obligated to respect or promote human rights law. The second is whether it is good policy for them to do so.

I argue that while development agencies should consider the impact of their projects on the well-being of the people affected by them, human rights law offers an unhelpful basis for evaluating projects. The problem with human rights law is that it is too ambiguous, contested, and politically charged to provide useful guidance. It is easy to anticipate that a commitment to human rights law will result in the erection of large bureaucratic hurdles that will interfere with development rather than promote it. The evidence suggests that economic growth is the most reliable means for advancing human rights, while there is little evidence that efforts to promote human rights directly by NGOs, governments, and others, actually succeed.

At the outset, I want to make a distinction between what I will call “maximal” and “minimal” human-rights promotion. When critics argue that development agencies should respect human rights they are rarely clear about what they mean. The maximal interpretation is that agencies should take it as part of their mission that of promoting human rights—improving governments’ compliance with human rights. The minimal interpretation is that agencies should avoid violating human rights directly or being complicit in the violation of human rights by governments and other entities that they cooperate with. In practice, these distinctions are hard to make, but they are important to keep in mind.

I. **Legal Issues**

Most states have ratified a number of human rights treaties, including the International Covenant on Civil and Political Rights (ICCPR), the International Covenant on Economic, Social, and Cultural Rights (ICESCR), the Convention on the Elimination of Discrimination Against Women (CEDAW), the

---

1. For an extensive discussion of efforts to persuade the World Bank to incorporate human rights into its mission, see Sarfaty (2012).
Convention on the Elimination of all Forms of Racial Discrimination (CERD), the Convention Against Torture (CAT), and the Convention on the Rights of Persons with Disabilities (CRPRD). The treaties declare that certain conditions and activities are protected as rights from government interference. For example, the ICCPR recognizes rights to freedom of expression and religious toleration, and rights not to be subject to arbitrary police actions, imprisonment without trial, and so on. The ICESCR guarantees rights to work, health care, pensions, and fair pay. CEDAW bans discrimination against women, while CERD bans discrimination against racial and ethnic minorities. CAT bans torture. CRPRD recognizes a range of rights and accommodations for people with disabilities. Table 1 lists the major treaties.

Table 1. The Major Human Rights Treaties

<table>
<thead>
<tr>
<th>Treaty</th>
<th>Entry into force</th>
<th>Ratifying states as of Jan. 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>International Convention on the Elimination of All Forms of Racial Discrimination (CERD)</td>
<td>1969</td>
<td>176</td>
</tr>
<tr>
<td>International Covenant on Civil and Political Rights (ICCPR)</td>
<td>1976</td>
<td>167</td>
</tr>
<tr>
<td>International Covenant on Economic, Social and Cultural Rights (ICESCR)</td>
<td>1976</td>
<td>160</td>
</tr>
<tr>
<td>Convention on the Elimination of All Forms of Discrimination against Women (CEDAW)</td>
<td>1981</td>
<td>187</td>
</tr>
<tr>
<td>Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT)</td>
<td>1987</td>
<td>153</td>
</tr>
<tr>
<td>Convention on the Rights of the Child (CRC)</td>
<td>1990</td>
<td>193</td>
</tr>
<tr>
<td>International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (CRMW)</td>
<td>2003</td>
<td>46</td>
</tr>
<tr>
<td>Convention on the Rights of Persons with Disabilities (CRPD)</td>
<td>2008</td>
<td>132</td>
</tr>
<tr>
<td>International Convention for the Protection of All Persons from Enforced Disappearance (CPPED)</td>
<td>2010</td>
<td>39</td>
</tr>
</tbody>
</table>

The treaties are, by their terms, binding on states, not on private individuals or corporations and other institutions. States offer various guarantees that they will respect and advance the rights. States comply with the treaties by passing laws that protect the rights of their citizens and other people on their territory, or providing administrative guidance to executive agencies. If a private individual murders someone else, the human rights treaties are not implicated. The murderer has violated domestic law only. By contrast, political assassinations undertaken by the state could be human rights violations. And if the government failed to prosecute murders, that too could be a human rights violation. Thus, there is nothing in the treaties that indicate that an NGO is directly bound by them.

The UN letter argues that because the Bank is an international organization with what lawyers call “international legal personality,” it is bound by “general rules of international law.” Moreover, the states that borrow from the Bank are

bound by human rights law, and the Bank has an obligation to avoid complicity in their human rights violations. But while there is a consensus that nongovernmental international organizations can undertake legal obligations by entering into treaties and agreements with states and are subject to certain customary norms of international law, there is no consensus that the general rules of international law include human rights obligations. Because most states have not ratified one or more human rights treaties, it is clear that “human rights law” as such is not a source of the “general rules of international law.” International law is based on the consent of states, and until states provide by treaty that NGOs are subject to human rights law, or certain parts of human rights law, NGOs must look to relevant domestic law for the source of their human rights-related obligations.

One reason for concern here—and I will return to this issue in part III—is that the treaties are obviously not written with NGOs in mind. All of the treaties contemplate a state that either abuses citizens (torturing or detaining them) or neglects them (fails to provide health care or schooling). NGOs do not try to exercise force over people, and we do not think that a group of people who start an NGO for a specific purpose—advancing religious ideals, providing malaria nets, documenting climate change, whatever—must also provide health care and schooling to people. If NGOs as a class are to be subject to international legal obligations, these must be worked out through negotiation among the affected parties. They cannot be inferred from the obligations that the treaties impose on states.

The UN letter is on stronger ground when it worries about complicity of NGOs in human rights violations by states. An NGO is morally complicit in human rights violations if it provides advice and resources to a government so that the government can harass a minority group. But beyond obvious cases like this one, complicity is an extremely delicate and difficult area of the law. Consider that governments in most (possibly all) poor countries commit human rights violations. An NGO that lends money to governments so that they can build infrastructure will almost certainly facilitate those violations by helping the government retain popularity or freeing up money for detentions and torture. Indeed, almost all development aid—including lending and cash aid by rich countries—would implicate those countries in human rights violations, because almost all poor countries engage in human rights violations. Governments that took seriously the complicity argument would be required to cut off aid to the countries that most need it.

Thus, a law against complicity, if broad enough, could block virtually all development projects. In domestic law, the rules of complicity are defined in a narrow fashion in order to avoid holding everyone liable for the legal violations of everyone else. Exactly how and whether complicity rules might be developed for NGOs is an important question, but one that needs to be answered—and implemented into law—before NGOs should be held to international human rights law.

The difficulties can be seen in the elaborate rules governing complicity under domestic law, where similar questions arise. If a bank lends money to a criminal
syndicate in order to finance extortion and prostitution, the bank or its employees will be held criminally liable. But if a bank lends money to a reputable corporation, and that money is then used for criminal purposes without the bank’s knowledge, it will not be held criminally liable. There are many intermediary cases—suppose the bank lends money to another bank with knowledge that the government suspects that some of the employees of the other bank engage in illegal underwriting practices. Numerous laws and judicial decisions have determined the scope of liability for lenders, usually requiring that the lender have some degree of knowledge or reason to know of the criminal activity and that the loans and the criminal activity be causally connected. It’s far from clear that these laws can be imported into the human rights context. A bank that lends to, say, Russia, China, Venezuela, or even the United States will typically know that government officials violate human rights, and since money is fungible, the loan might contribute to those violations in some way. Do countries violate human rights law by failing to provide medical care to poor people? Arguably yes, but if we made development agencies liable for these violations, then there will be less lending to poor countries rather than fewer human rights violations.

The heart of the UN letter, however, is not a legal argument but a policy argument. The letter argues that the World Bank should recognize itself as bound by human rights law because that is the right thing to do. It would improve development outcomes and promote respect for human rights law among states. The letter seems to make both the minimalist and maximalist arguments, perhaps in the alternative—the World Bank should itself not violate human rights, and it should promote respect for human rights as well. To that argument, I now turn.

II. Development as Freedom

The major intellectual source for the argument that development agencies should advance human rights law is Amartya Sen’s (1999) book, *Development as Freedom*. Sen’s book is remembered for his argument that development policy should advance “freedom,” which he defined in terms of the flourishing of human capabilities and “functionings.” Sen argues that freedom is a good in itself, by contrast to wealth, which is a means to further ends. For that reason, development policy should advance freedom rather than (or rather than merely) economic growth. He also argues that development agencies should advance freedom because freedom itself advances economic growth. So even if development agencies should consider their mission limited to that of advancing economic growth, they should promote freedom as an instrumental or intermediate good that contributes to wealth.

Sen’s capabilities approach is a version of what philosophers call “objective-list welfarism.”3 *Development as Freedom*, along with earlier work

3. For a discussion of the various philosophical approaches to understanding well-being, see Adler and Posner (2006).
by Sen, plays an important role in the intellectual history of welfare economics. Before the 1970s—in the period when development policy was initially debated and formulated—economists who believed that their discipline should be used to guide public policy tended to have a philosophically crude notion of the public good. As a descriptive science, economics takes people’s “preferences” as given, where preferences are just a description of how they rank states of the world that affect them. These preferences can frequently be derived simply by observing people’s behavior. When giving advice to policy makers, economists would retain this assumption. Implicit in their advice, then, was the assumption that the goal of governments is to maximize aggregate welfare as measured by people’s actual preferences. This, in turn, led to the view that governments should maximize economic growth, typically measured by GDP. Development policy focused on infrastructure. Countries need roads, bridges, power plants, and dams. The assumption was that if they were not building these things, it must be because they did not have access to capital markets. Development agencies could help countries by lending against the revenues generated by these projects, or by providing other types of subsidies. Investment in infrastructure should promote economic growth.

Philosophers had long argued that the “preference-based welfarism,” as I will call it, is not an attractive normative goal. Sen joined in this critique. Welfarism, as philosophers use this term, is the view that the public good is advanced when a population’s well-being is maximized. But what is well-being? Jeremy Bentham was a welfarist but of a particular kind. He believed that well-being referred to a positive mental state. Most modern philosophers reject Bentham because Bentham’s view implies that there are no higher or lower pleasures—that people who are coerced into taking happy pills are as well-off as people who derive happiness from family, work, and public commitment. A similar complaint has been leveled at preference-based welfarism. People’s preferences can be poorly informed, or distorted by circumstances; if so, satisfying those preferences will not necessarily make them better off. Sen argued, for example, that people who are accustomed to poverty may come to believe that poverty is good for them, or that women brought up in a sexist society might come to endorse their subordinate positions. Sen insisted that those people are not actually made better off when those preferences are satisfied.

Sen’s objective-list welfarism holds that certain goods are universal. People are made better off if they are educated, fed, given medical care, allowed to participate in politics, and so on—regardless of whether they think they are made better off by these things. When Sen uses the term “freedom,” he often seems to be defining it broadly to encompass all of these goods, so advancing people’s freedom means advancing their well-being. The idea is that a person cannot really be free unless he is adequately fed and clothed, has work, and so on; freedom is not just a matter of having choices.

4. See also Nussbaum (2011).
We can round out the discussion by mentioning that some philosophers disagree with Sen that welfare can be separated from preferences but disagree with economists that preferences necessarily reflect people’s well-being. On their view, satisfaction of preferences advances a person’s welfare only when those preferences are informed and undistorted. Thus, we can distinguish four views of welfare: objective; mental-state; preference-based; and “undistorted” preference-based. Most philosophers reject the preference-based approach of the economists, but debate about the other three continues.

Sen’s book was a watershed because it suggested that development policy should not necessarily be focused on maximizing GDP. Maybe infrastructure projects are not the best use of development money; or maybe, if they are, they should be used to advance people’s well-being in a philosophically proper sense. But what Sen’s view meant for development policy was actually quite obscure. What does it mean for development policy to advance “freedom” or “human capabilities” or objective well-being?

The major approach of his followers has been to try to use a metric that encompasses Sen’s list of objective goods. This metric would be superior to GDP because GDP reflects preference-based welfarism—just the total amount of money that people pay for goods and services, whether their preferences are good ones or not. The most famous such metric is the United Nation’s human development index (HDI). The HDI gives countries a score based on an algorithm that takes into account GDP per capita, literacy, and life expectancy.

Why do these three factors go into the HDI? The immediate problem faced with an objective-list approach is that of arbitrariness. If the good is not derived from people’s preferences, then what is it derived from? Why not use environmental quality, unemployment rate, and availability of contraception? Philosophers have proposed numerous approaches for answering these questions, but they cannot agree among themselves, and there is no obvious way to answer these questions.

But these controversies may not matter much. The reason is that people tend to have either all the objective goods or none of them. Wealthier people are more educated and healthier, and have more access to contraception, and are more likely to live in unpolluted localities, than poorer people are. One way to understand this phenomenon is that when philosophers argue about the list of objective goods, they look around at what people seem to enjoy, and what people seem to enjoy are the goods that they voluntarily buy or obtain through the political process or through migration. Another way is to recognize that if some goods really are objective, then people will buy them unless they are incredibly deluded.

The upshot is that it makes little difference if one uses HDI or GDP per capita as one’s metric for measuring the well-being of populations. They are highly correlated, in part because GDP per capita is an input in HDI, and partly because wealthier countries tend to have healthier and more educated populations.

Figures 1 and 2 illustrate this point. Figure 1 shows the relationship between GDP per capita (logged) and expected years of schooling, one of the inputs in HDI. Figure 2 shows the relationship between logged GDP per capita and
another such input—life expectancy. The figures show a very strong correlation in both cases.

So Sen’s philosophical point, while important, was of little practical consequence. Whether development agencies try to help countries improve their HDI
or their GDP per capita, they will probably do the same thing. They will fund infrastructure projects—because infrastructure projects, when successful, enrich a population and thus enable it to spend more money on health and education. And development agencies will fund health and education—because healthier and more educated populations make more money.

Still, the significance of the HDI for present purposes is that it shows that if one agrees with Sen, one might endorse a metric that approximates the objective goods that Sen endorses without having to rely on human rights or human rights law. The metric could be the HDI, or another metric that accounts for objective goods—or it could be GDP per capita, as I have argued. These metrics could support traditional development policy or minor variations on it. One could imagine, for example, a development policy that emphasizes support for schools and medical clinics rather than transportation infrastructure. Or the funneling of development resources to countries that score badly on health and education even though they are wealthier than countries that score well on these variables. These approaches may well be justified, but they do not entail a commitment to human rights as it has been embodied in international human rights law.

Sen does not argue that development agencies should advance human rights law or should otherwise be subject to or constrained by human rights law, though he does not rule out this view, either. He does not address it. In a chapter on human rights law, Sen argues that many non-Western cultures recognize the values that he sees as the source of human rights law—and therefore, that human rights law is not a foreign imposition on these non-Western cultures. The argument was a response to the so-called Asian-values position that people in Asian countries care about order and harmony rather than individualistic values. Be that as it may, Sen does not take a position on whether human rights law embodies his capabilities approach.

The UN letter and the HRW document, in contrast, argue that development policy should respect and promote human rights law. What that might mean is the subject of the next part.

### III. Human Rights Law

What would it mean for development policy to comply with human rights law? Human Rights Watch and the UN experts do not make clear their thinking on this issue. One possible argument is what I have called the “minimal” approach—that a development project must comply with human rights law; another possibility is the maximal approach—that a development project must promote human rights law. In the first case, a proposed project might be screened by human rights lawyers and blocked if they believe that it would cause a violation of human rights law. In the second case, development experts might look for projects that advance the values underlying human rights. To get a sense what either approach would entail, one must know something about human rights law.
A. What Is International Human Rights Law?

The treaties described above lay out hundreds of different rights or interests that states are required to protect. A partial list includes rights associated with criminal procedure (to a lawyer, not to be detained without charges, to a trial, to appeal, not to be tortured); with political and civil freedom (to speech, to association, to travel, to privacy in communications); with employment and other economic activity (to fair wages, to a job, to unionize, to own property); with family life (to marry whoever one wants, to bequeath property, to establish a family, for children to live with their parents); and with social insurance (to welfare, to health care, to education, to food). There are also rights to cultural freedom, to self-determination, to religious freedom, and to scientific progress. Numerous rights protect children, women, ethnic and racial minorities, and disabled people from discrimination, and offer various types of accommodations. There are other rights that are not found in the treaties but have been asserted by numerous states, and are sometimes thought to be part of the law. These include a right to security (protection from crime), the right to development (the right of poor countries to economic growth), and, most controversially, the right to be protected from defamation of religion, which has been asserted mainly by Islamic countries.

The rights are thus extraordinarily numerous, covering nearly all aspects of human life, and effectively requiring governments to advance all human interests of any importance. The rights are thus not only numerous but very generous. However, the rights are also often ambiguous. The right to freedom of expression, for example, contains exceptions for expression that disrupts public order or violates public morality. The treaties do not explain how the value of expression should be weighed against the values of order and morality; nor do they describe the relative weights of different rights when they conflict. For example, certain types of expression may violate religious sensibilities; the idea of defamation of religion reflects the view of many states that the religious sensibilities should take priority. The treaties also provide no indication how a government should devote resources to enhancing health, education, and welfare; how much money should be devoted to training police so they do not violate people’s rights; and so on.

Countries have created numerous international institutions that are given varying levels of authority for interpreting treaties, setting priorities, and monitoring compliance. These institutions include the Human Rights Council, the UN Office of the High Commissioner on Human Rights, nine treaty committees that are connected to the nine core treaties, and various regional courts, like the European Court of Human Rights. But none of these institutions possess the authority to issue final interpretations of the international treaties, and hence to resolve conflicts among states, which frequently interpret the rights in different ways. The international institutions also are given meager resources to enforce their views.

Probably for these reasons, states’ ratifications of human rights treaties seem to have had little impact on their behavior. The picture is complicated. A few studies have found that respect for human rights increases in some countries after
they ratify treaties, but the magnitudes are small, and the results hold for only some countries and only for some of the rights. The usual bugaboos of regression analysis—omitted variables, reverse causation problems, and noisy data—also require one to suspend judgment. For example, the evidence suggests that the amount of torture in the world—a human rights violation of exceptional importance—has not declined since the human rights treaties were negotiated. Figure 3 shows some additional evidence. If countries took ratification of the ICCPR seriously, then one would have expected extrajudicial killings to decline after they ratified it, and that freedom of speech, freedom of religion, and independence of the judiciary to increase. Yet there is no evidence of such effects. While some progress occurred in the 1990s—probably because of the collapse of communism, not the rise of human rights law—over the last fifteen years progress in human rights has ground to a halt.

One of the reasons that human rights has not advanced is that countries have not created strong enforcement mechanisms. The councils, offices, commissions, and courts that I mentioned above have few formal powers to compel states to comply with human rights, and are starved of funds and staff. Countries are simply unwilling to delegate to independent institutions the power to compel them to act against their will.

5. See Chilton and Posner (2016) for a survey of the literature, which includes Hathaway (2002), Simmons (2009), and Fariss (2014).
This state of affairs can be contrasted with domestic rights enforcement. In advanced democracies, constitutions usually contain a limited number of judicially enforceable rights—for example, rights to fair trials. Courts are powerful and independent and usually enforce the rights by ordering the release of people detained in violation of them, ordering the government to pay money to victims of rights violations, and so on. Even in countries with constitutional rights to work, housing, and the like, courts rarely enforce these rights; instead, they are left to politics. So, as a practical matter, in advanced countries, the courts focus on political and civil rights. Because the courts are respected, and because they hear numerous cases, they are able to resolve conflicts among rights, and provide interpretations so that rights can be applied in specific contexts.

Imagine, then, a bank in an advanced country that seeks to make a loan and does not want to violate the law. The bank’s lawyers can advise it as to what domestic law is, and in this way the bank can easily avoid loans that, for example, are used by criminal organizations. Banks are under no obligation to respect constitutional rights, which are enforced against governments, not private actors.

B. International Human Rights Law and Development Policy

With this picture of international human rights law in mind, let us turn back to development policy. Development agencies face a huge array of choices as to what projects to fund. Old-style development projects include power plants, dams, and canals. Today, many development agencies fund vaccination programs, malaria net distributions, and educational programs. Any of these projects could be good or bad from a human rights standpoint, but as we have seen, it’s hard to know what the human rights standpoint requires.

Consider, for example, a simple project like the construction of a road. A road can enhance economic activity by reducing transportation costs, but it can also harm people. People whose dwellings block the planned route will be displaced, of course, but roads can also have more complicated impacts. A limited-access highway could convert settlements not connected to exits into ghost towns. The government may need to use force to move people, to protect materials from scavengers, and to protect roadbuilders from criminals. The road construction may harm the environment, and interfere with farming by causing diversion of waterways and blockage of paths for farm machinery. Government officials may engage in reprisals against people who criticize the project. And then often there will be corruption, incompetence, and needless violence.

6. In most poorer countries, courts are too weak or corrupt to enforce rights vigorously, though often rights are respected because public opinion approves of them, especially in democracies.
No development agency will deliberately cause harm, but all development agencies must cooperate with governments, including local governments and police. The flipside of the hard-earned lesson that locals must be “stakeholders” in development projects—that a development project will fail unless the agency consults the people affected by it and obtains their approval and cooperation—is that the development agency must also defer as much as possible to their interests and judgments. In developing countries, where corruption and violence is endemic, this means that projects also will be subject to these scourges.

What would it mean to bind the development agency to human rights law for a project like this one? Many of the harms I have described can be easily recast as human rights violations—if people are displaced without compensation or with inadequate compensation, for example, or are forced to pay bribes, or if environmental harms occur. The project might therefore be blocked, despite the good it can bring. Or, alternatively, the development agency might be required to take measures to eliminate the human rights violations.

But it is impossible to eliminate all human rights violations. The best that could be done would be mitigation measures. For example, the development agency could be required to survey affected people periodically and ensure that their rights are respected. But this introduces three costs.

First, the cost of monitoring for human rights violations is not trivial. Staff and other resources must be used to interview people, evaluate their statements, seek official responses, compile data, and write reports. Except when abuses are egregious, it will be often be difficult to evaluate people’s claims. People who oppose projects for reasons having nothing to do with human rights or who seek to pressure the government to increase payouts may exaggerate the damage done by a project.

Second, because violence and corruption are widespread in poor countries—the countries that are more desperately in need of development aid—there will almost always be a plausible argument that a development project causes human rights violations. All development projects require the permission and cooperation of government officials, and all of them can be used by government officials for political ends—to help favored constituents, to harm disfavored groups. Where government officials are corrupt and resort to violence, the agency will find itself implicated in these human rights violations, almost unavoidably. This means that pressure to avoid complicity in human rights violations can push development agencies out of countries that need their help the most.

Third, because of the extraordinary diversity of countries, it will often be difficult to determine whether human rights violations that accompany projects can be controlled or not. In one country, it may be reasonable for a development agency to assume that people who are displaced from housing will eventually obtain redress in courts or through the political process, while in another country such an assumption may not be reasonable. Countries differ greatly with respect to their institutional capacity and cultural tolerance for corruption and violence. Indeed, variation will occur across regions within countries.
Even if these costs can be tolerated, another problem looms above all the rest—the ambiguity of human rights law, which I discussed above. Consider a development project to build a power plant that will reduce the cost of electricity in a poor rural area, but will also cause harm to the environment, interfere with traditional hunting grounds, require displacement of some people, and create opportunities for local corruption. It is easy to say that the power plant will advance some rights and retard others. Some people will be better off, others worse off. How is the development agency to determine whether the project complies with human rights law?

It seems very likely that the question will be shunted off to lawyers in a bureaucracy somewhere, who will develop rules of thumbs and presumptions based mainly on their intuitions. If the lawyers are ignorant of local conditions, as they are likely to be, and so unable to make the tradeoffs that best advances the public good, they will block some good projects while approving some bad projects. The additional layer of bureaucratic decision making will cause delay and increase cost, with very limited benefits, if any at all.

And this creates a significant problem. The places with the worst human rights violations—and the worst institutions—are the places that need development aid the most. But if agencies face substantial costs and risks because of their exposure to human rights law, then they will most likely respond by abandoning the poorest people for those who are not so poor.

C. Lessons from the World Bank’s Rule of Law Projects

The current push for the World Bank and other development agencies to comply with human rights law echoes an earlier movement from the 1980s and 1990s. At that time, the view arose that development agencies should advance the “rule of law” in developing countries. Rule-of-law projects took many different forms, including: training judges; building courthouses; drafting bankruptcy legislation; introducing title registries; and developing arbitration programs. The theory was that many countries failed to develop, despite receiving a great deal of aid, because they lacked secure property rights, transparent legal procedures, and other features of the rule of law. The absence of these protections deterred foreigners from investing in countries, and allowed many of the benefits of development to be lost to corruption and government inefficiency.

“Human rights” and the “rule of law” are not the same concepts, and indeed it is hard to compare them because they are both so vague, but they clearly overlap. One can think of the rule of law as a particular way of institutionalizing the values underlying human rights. Some human rights—like the right to a fair trial—are inseparable from the rule of law, which emphasizes procedural fairness and the independence of the judiciary. Other human rights—like the right to vote—are vulnerable in countries where the rule of law is absent or weak. Elections can be undermined by corruption, and disputes can explode into violence; a state with strong legal institutions can address both of these problems. At the same time, some human rights go beyond what the rule of law requires. The
right to health care, for example, is not entailed by the rule of law, as it is normally understood. Still, the concepts are close enough that the history of the rule-of-law projects is instructive.

As Carothers (2003) discusses in an influential paper, the rationales for rule-of-law promotion seemed self-evident, but scrutiny revealed some serious ambiguities. One argument for the rule of law is that it is necessary to attract foreign investment. Investors will not put money in a country unless they know that their property interests will be protected by courts from expropriation. However, China stood as an obvious counterexample to this thesis. Recent research explains that foreign and domestic investors often obtain security for their property rights from complicated personal and political relationships—with other businesses, with government officials, with influential members of the public—rather than through the rule of law. An additional point, overlooked by researchers as far as I know, is that foreign investors can benefit if a country violates the rule of law. Most obviously, the host government may expropriate land from farmers and hand it over or sell it to a foreign investor because the host government seeks economic growth. Host governments might also use their security forces to protect foreign-owned pipelines, mines, and factories, and security forces in poor countries frequently violate human rights. And governments may need the option to expropriate in order to appease a public demand for wealth redistribution and in this way forestall revolutionary movements and insurrections that would be more damaging to foreign investors than a limited and orderly round of expropriation by the incumbent government.

Another argument for rule-of-law promotion mentioned by Carothers is that the modern market economy depends on the rule of law. It is a core idea among western thinkers that a market economy depends on the enforcement of property and contract rights by an independent judiciary. Yet this is an altogether too static understanding of markets. In developed countries, this may well be true. But in developing countries, the major problem with development is often the barriers—institutional, customary, traditional, religious, political—that stand in the way of development of a market economy. China again is exemplary. The key point is that while China encouraged some types of economic activity by protecting (often de facto) property rights, it also made way for this economic activity by violating customary property norms—echoing the enclosure movement in Great Britain, and the dispersal of Indians in the United States. All of these activities violated the rule of law; all of them may well have been necessary for development.

A third argument is that the rule of law promotes democracy. Again, it is not clear whether this is true. Revolutions that inaugurate democracies often begin with massive rule-of-law violations; rigid insistence on the rule of law can also block the formation of democracy in countries with authoritarian traditions and

8. See also Carothers (2006); Dam (2006).
institutions. But the more important point is that many developing nations repudi-ate democracy—or, at least, democracy in the full-blown sense. Again, China above all. A cliché of the development literature is that development projects work best when the local government supports them. There must be “buy-in.” A country that sees rule-of-law projects as Trojan Horses from which democracy will spring will do all it can to undermine them.

All of these lessons from the rule-of-law movement have clear application to the goal of promoting human rights in developing countries. The relationship between human rights, development, and democracy is complex. Promoting human rights will not necessarily accelerate development or democratization. Countries that have advanced toward the endpoint envisioned by the human rights regime have done so unevenly. Many countries have gotten rich first and then supplied political rights. Others have respected some political rights (e.g., freedom of expression) but not others (e.g., freedom of religion). Many countries banned discrimination against various minorities only after advancing along other lines. The human rights regime assumes a kind of “recipe” for economic and political development, but no such recipe exists.

But there is an even more serious problem—for both rule-of-law promotion and human rights—and that is the “problem of knowledge,” as Carothers puts it. Rule-of-law projects foundered because of the complexity of the legal systems to which they were applied. To laymen, law is just a set of rules. Lawyers know better. In any society, the rules listed in books of statutes are little more than a starting point for understanding the law. (Indeed, many countries do not have those books, or do not enforce the laws in those books.) In practice, law is a highly complex undertaking involving numerous actors—judges, lawyers, religious figures, traditional mediators, informal go-betweens, trade groups, families—who often themselves have only a vague notion of what the rules are, and discover them only in the context of a specific dispute during which arguments that appeal to precedent and authority are hashed out. A development agency which asks local experts what the “law” is will likely receive many different answers, usually starting with “it depends . . . .” Grafting a well-developed statute from a Western country into the law books of a developing country—say, a bankruptcy code—can have unpredictable results. What matters is not that the law is in the books but how legal actors react to it, and it is often easy for them to ignore it completely.

Moreover, because the law varies tremendously from country to country—and even from place to place within countries—a rule-of-law project that achieves success in one place may founder in another. Law is different from other development projects that rely on universal scientific and engineering principles. Taking into account the differences in the environment, a dam will work as well in Cambodia as in Guyana. A bankruptcy law that works in Cambodia will not necessarily work in Guyana because legal principles and practices are not universal or even very similar across jurisdictions.

Human rights advocates echo rule-of-law promoters in a way that should make them uncomfortable. Human rights, like the rule of law, are taken to be
self-evident goods, hardly in need of any debate. As a result, there is no clear sense of what the rationale of human rights is, just as there was none for the rule of law. In both cases, advocates end up trying to apply a formula or recipe to diverse countries with diverse histories, traditions, and cultural and political values—without understanding that countries’ ability to absorb western institutions and change their way of doing things varies tremendously, and in a way that demands a pragmatic approach that cannot be formulated in treaties or UN documents.

**Conclusion**

Development agencies, like doctors, should do no harm. The dilemma that they face is that if they provide loans or other aid to governments in poor countries, they will almost always be required to cooperate with corrupt and dishonest officials, who may use their money to bad ends. But if they keep their hands clean under all circumstances, it will be impossible to provide aid to the people who need it most. What should they do?

Sen was right to criticize economists who believed that development agencies should focus on economic growth and nothing else. Certainly as a matter of theory, this is wrong. But the implications of Sen’s critique are more complicated than people have realized. It is not at all clear what the goal of development agencies should be if not economic growth; it is most unlikely that alternative measures like HDI are an improvement.

The idea that development agencies should promote human rights is superficially appealing but does not escape the problem. Human rights law is both too vague and too contestable to provide guidance to agencies. So what should agencies do? If they avoid breaking the laws of the country that they work in, and the laws of the western countries from which they obtain donations, and otherwise use common sense, they will probably do 90 percent of what any reasonable human rights approach would ask them to do. They would certainly avoid complicity in any kind of serious violence. Demanding that development agencies also promote human rights, by contrast, would ensnarl them in debates and controversies that have no end.

**References**


Sialkot, Pakistan, is the world center of hand-stitched soccer-ball production, home to roughly 130 firms, which produce for all major brands (Atkin et al. 2015a, 2015b). At first blush, the existence of this cluster is puzzling. Pakistanis’ sporting passion is cricket; soccer has always been of marginal interest in the country.¹ One might be tempted to dismiss the existence of a cluster where there is no apparent local demand as an anomaly, curious in itself but not indicative of a deeper pattern. But the phenomenon is not uncommon. For example, Sialkot is also a leading world producer of another not-particularly-Pakistani product: bagpipes (BBC 2012).

The existence of the soccer-ball cluster seems to argue against a keystone of modern thinking about trade and development: what Krugman (1980) called the “home market effect,” the idea that countries will tend to export goods for which

---

¹ As one indication, Google Trends indicates that since 2010 web searches for “cricket” have outnumbered searches for the names of all other sports combined by a factor of five in Pakistan.
they have large domestic markets. But a closer examination of the history of the cluster reveals a subtler story. In this short paper, we draw on historical sources and interviews with firms to trace the development of the cluster from its origins in the late nineteenth century. Our main message is that the home market effect appears to have been important in creating the cluster under British rule and that agglomeration effects—as well as some effective industrial policy—led to continued growth of the sector even after the colonists (the main local source of demand) left the subcontinent. In the presence of agglomeration effects, in other words, a lack of contemporary local demand is not in itself an argument against home market effects. The case study underlines the importance of understanding longer-term historical dynamics for understanding the current pattern of specialization in the world economy—and industrial development more generally.

The case of the soccer-ball industry is also consistent with the idea of Artopoulos, Friel, and Hallak (2013) that close exposure to final consumers is a key to manufacturing success in developing countries, at least in the early stages. In several of the Argentinian industries that Artopoulos et al. (2013) study, the exposure occurred because the pioneering entrepreneurs were themselves embedded in the richer-country markets before returning to Argentina. In the soccer-ball case, the British consumers were present locally. But the cases have in common the idea that initial knowledge of end-consumer tastes (which may not be easy to come by) appears to have been key to the success of the respective industries.

I. Origins of the Cluster

Sialkot is located close to the border between present-day Pakistan and India. The city historically had a large Muslim population. Because it straddled the Ravi and Chenab Rivers, the land was fertile and attracted many traders of agricultural goods and livestock and the city held multiple major cattle fairs each year (Government of the Punjab 1895). Because of its strategic location and natural resources, the Mughals established a fort in Sialkot at the end of the twelfth century. Later, the British established a large military garrison there in 1852. Drawn by the demands for goods and services generated by a large, well-funded military presence, skilled workers came to the city (Government of the Punjab 1889). The influx of soccer-playing foreigners, the availability of leather, and the presence of skilled Muslims able and willing to work with leather would all be important for development of the soccer ball cluster.

In the late nineteenth century, production of soccer balls arose as a way for the local artisans to use their skills with leather to cater to the British in colonial India. Iqbal Sandal, member of a well-known family in the industry and author of a self-published history of soccer-ball production in Sialkot, describes the origins of the sector as follows (Sandal, undated). In 1889, a British sergeant approached a Muslim saddle-maker in Sialkot named Fazal Elahi and asked him to repair a ball. Elahi repaired the ball, but the leather of the ball was dried out and the ball quickly needed another repair. Elahi cut open the existing ball and made...
a replica following the same pattern from softened leather. The sergeant, happy not to have to wait the typical eight months for a new ball to arrive from England, ordered six more balls for his regiments. The sergeant also introduced Elahi to colleagues in other regiments around Sialkot and elsewhere on the subcontinent. After producing on his own for 15 years, in 1904 Elahi took on Alla Ditta Sandal, the son of a cousin, as an apprentice. Alla Ditta Sandal and his father later established their own firm. Elahi subsequently took in a number of other apprentices, who also subsequently launched new businesses.2

In this early phase, it appears that the crucial factors for the success of the industry were the local demand from British army personnel and knowledge spillovers due to spinoffs by apprentices from the first firms.

II. EXPANSION OF THE CLUSTER

Sialkoti traders began traveling to British garrison towns across India, first repairing and then selling new balls produced in Sialkot (Government of the Punjab 1921). During World War I, when English factories were producing few balls, Sialkoti firms began exporting to other colonies (Government of the Punjab 1921). Exports subsequently expanded to a wider range of countries. Khalid Gundra of Taj Mahal, a soccer ball firm founded in 1904, explains how his family’s firm grew over time:

Between 1914 and 1918, the demand for footballs was on the rise in Egypt because of the British presence there and the firm’s first export was to Egypt in around late 1910s. By the 1930s the firm started to supply sports goods to Australia, Canada, London and South Africa . . . During World War II (1939–1945), word starting spreading internationally about Sialkot’s sports industry and exports increased. By the mid 1960’s the Sialkot sports goods sector was expanding rapidly since Sialkot was recognized as a destination for supplying quality sports goods (including football) and customer orders exceeded the firm’s production capacity.3

This diversification meant that when most of the approximately 155,000 British subjects4 left the subcontinent with Independence and Partition in 1947, Sialkoti firms had other sources of demand beyond the local market. Export data for sporting goods starting in 1954–55 shows that already by that time, Pakistan was exporting over 10 million US dollars of sporting goods in 2013 US dollars,5

2. Spinoffs of this type appear to have been important among firms producing other sporting goods in Sialkot in the same period. The 1920 British Gazetteer for Sialkot reports: “The two Uberoi brothers . . . started the manufacture of cricket bats with only half a dozen workmen . . . In 1899 the Uberoi brothers separated and took to trade rivalry . . . The first grade workmen able to turn out work equal to the English model . . . are extremely difficult to retain. Many workmen have learnt in [one of the brother’s] shops and have gone away to set up their own.” Mostafa and Klepper (2011) note that, more recently, Bangladesh’s apparel industry also expanded through spinoffs from the initial firms that entered the industry.


4. The 1931 Census of India (Hutton 1933) reports 155,555 British subjects living in India. The 1941 census was deemed unreliable due to the war.

5. Source: Economic Survey of Pakistan (various issue) and State Bank of Pakistan. Export values in rupees are converted to US dollars using prevailing exchange rates and are then deflated by US Urban CPI (Dec. 2012 = 100).
the majority of which were soccer balls. As shown in figure 1, exports grew rapidly in subsequent decades, increasing to over US$20 million in 1964–65, over US$90 million in 1974–75, and peaking at US$547 million in 1997–98. Destination-level export data spanning the entire period are unavailable, but starting in 1972, the UN Comtrade database reports destinations for Pakistan’s exports. By 1972, Pakistani sporting goods had already penetrated 98 export markets and reached 135 markets by 1997. In more recent years, Sialkot has been challenged by East Asian producers, particularly at the low end of the market, but remains the dominant producer of hand-stitched soccer balls.

In the following subsections we consider two forces that appear to have been key to the industry’s post–WWII expansion: agglomeration forces and industrial policy.

**Agglomeration Forces**

As is well known, identifying the determinants of agglomeration econometrically is challenging in the best of circumstances and nearly impossible without detailed

---

6. The coarseness of product categories in export data during this period prevents us from calculating soccer balls’ share of sporting goods exports. However, they can be separately identified in US import data using the 10-digit HS classification system. In 1989, the earliest date for which we observe US imports at this level, Pakistani soccer-balls (HS 9506.62.40.80) accounted for 62 percent of the country’s exports of sporting goods (HS 9606) to the US, potentially a lower bound for global exports since the US is not noted for its love of soccer.

7. The relevant SITC code (Revision 1) in the UN Comtrade database is 8944 “Other sporting goods.”
data on many locations and industries. In this short paper, our approach is to rely on reports of participants in the industry about the agglomeration forces they themselves believe to be important.

Haji Abdul Rehman, a firm owner, gave an overview of the benefits to his firm from operating in the cluster:

Labor pooling is the most important factor in influencing the working of the firm . . . It would at least take 3 years, provided that everything goes well, to train the novices for the job . . . Also, cooperation among firms is an extremely important benefit . . . The small makers are largely dependent on big firms for the orders . . . Besides this, availability of specialized inputs and services is also a highly crucial outcome of clusters . . . [D]ie-makers and technicians for machine repairs are a call away . . . Finally, the cluster allows firms to keep themselves updated on any new developments. For instance, it takes a lot of effort, money, and time to figure out what type of designs are more preferred by the clients. By just following the designs demanded by famous brands in Sialkot, one can get an idea of global customers’ tastes and preferences.8

In November 2015 we asked five soccer-ball firms which agglomeration forces they believe were most important to the success of the Sialkot soccer-ball cluster. Table 1 reports their responses and reveals that firms believe labor pooling and knowledge spillovers were most crucial. Also important were access to specialized inputs and the ability to find subcontractors (which can be thought of as a specialized input service). In contrast, firms thought being able to negotiate with the government as a sector and the benefits of buyers coming to the cluster due to it being the global center of hand-stitched soccer balls were less important.

### Table 1. Survey of Soccer Ball Manufacturers about Growth Factors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor Pooling</td>
<td>7</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Knowledge Spillovers</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>8.8</td>
<td>2</td>
</tr>
<tr>
<td>Access to Specialized Inputs</td>
<td>8</td>
<td>10</td>
<td>–</td>
<td>10</td>
<td>7</td>
<td>8.75</td>
<td>3</td>
</tr>
<tr>
<td>Access to Subcontractors</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>7</td>
<td>8.4</td>
<td>4</td>
</tr>
<tr>
<td>Negotiating with Government as a</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td>7</td>
<td>7.2</td>
<td>5</td>
</tr>
<tr>
<td>united sector</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to Buyers</td>
<td>7</td>
<td>–</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>6.5</td>
<td>6</td>
</tr>
<tr>
<td>Shared Use of Facilities</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>6.4</td>
<td>7</td>
</tr>
</tbody>
</table>

Notes: Each firm was asked to rate the agglomeration force on a scale of 1–10. Then the firm average was used to rank each agglomeration force in order of importance for the development of Sialkot’s soccer-ball sector. Dashes represent nonresponse.


### Industrial Policy

Producers also partially credit government industrial policy for the post-war expansion. Gundra of Taj Mahal cites two programs in particular. First, in 1960–61, the government created a specialized industrial area in Sialkot that
offered land to firms at 50 percent of the actual land value. Second, between the early 1970s and mid-1990s, the government operated an export rebate scheme. The scheme was designed to spur nontraditional exports (it applied only to manufacturing products) by allowing firms to claim back customs duties, sales tax, and excise duties paid on raw materials that were used as inputs for exports (see Government of Pakistan 1978 for details).

Gundra notes: “The 1980’s was a golden period for the football sector . . . the government offered significant rebates on exports.” Abdul Rehman adds: “The government offered large rebates for exporters in 1970s until the 1990s. It was during this period that foreign clients used to visit and stay in the city for around 2–3 months searching for football manufacturers.” The rapid expansion of exports in the 1980s and 1990s shown in figure 1 is consistent with this perceived success of the export rebate policy, with soccer-ball exports booming in the 1980s and 1990s. The Punjab Industries Data (published by Department of Industries, Government of Punjab) allows us to supplement these figures with firm-level data. According to these data, 181 firms in Sialkot produced soccer balls as their primary product in 2010, and table 2 provides a tabulation of the firms’ year of incorporation. Of the surviving firms in 2010, more than half were established during the 1980s and 1990s when the export subsidy program was in place.

Since the mid-1990s, the Sialkot soccer-ball cluster has been undergoing a gradual contraction (see Atkin et al. (2015b), fig. A.1.) Although it has retained the dominant position in hand-stitched soccer balls, the hand-stitched segment has been losing market share to machine-stitched balls, which tend to be lower-quality and are typically manufactured in China. Despite the recent decline, the

<table>
<thead>
<tr>
<th>Decade</th>
<th>Year of Establishment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1890s</td>
<td>1</td>
</tr>
<tr>
<td>1900s</td>
<td>1</td>
</tr>
<tr>
<td>1910s</td>
<td>0</td>
</tr>
<tr>
<td>1920s</td>
<td>1</td>
</tr>
<tr>
<td>1930s</td>
<td>1</td>
</tr>
<tr>
<td>1940s</td>
<td>2</td>
</tr>
<tr>
<td>1950s</td>
<td>16</td>
</tr>
<tr>
<td>1960s</td>
<td>14</td>
</tr>
<tr>
<td>1970s</td>
<td>27</td>
</tr>
<tr>
<td>1980s</td>
<td>32</td>
</tr>
<tr>
<td>1990s</td>
<td>66</td>
</tr>
<tr>
<td>2000s</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>181</td>
</tr>
</tbody>
</table>

Notes: The source of this table is the 2010 Punjab Industries Data. The sample consists of 181 firms that report soccer balls as their primary product, and the table reports the year of establishment by decade.
cluster continues to employ tens of thousands of workers and remains an important export industry for Pakistan.

III. Conclusion

This short paper has examined the birth and growth of the Sialkot soccer ball cluster. The story that emerges is one of an initial home market effect, in this case due to the presence of British colonists, with later export success driven by the forces of agglomeration coupled with effective industrial policy. While only a single example, we think these three ingredients are likely to play important roles in industrial development more generally.

In addition, the case study is a useful reminder that the current location of industries does not necessarily reflect current economic fundamentals, neither in the classical sense of inherent productivity differences or differences in factor supplies nor in the “new trade theory” sense of the proximity of customers. Even the influential work on “self-discovery” by Hausmann and Rodrik (2003) tends to view the location of industries as responding to fundamentals: a pioneer firm may “discover” that it can be profitable producing a particular good in a particular location, and that may generate a knowledge externality for other firms, but the profitability of firms remains a function of exogenous characteristics of the location. In contrast, this case study suggests that, in the presence of agglomeration forces, once a sector reaches a certain critical mass, the persistence and growth of a sector can become entirely delinked from locational fundamentals. More work is required to examine how general this phenomenon is, but we suspect that it will not be difficult to identify other examples.

References


Government of Pakistan. Economic Survey of Pakistan (various years), Finance Division, Islamabad.


Cross-Sector Misallocation with Sector-Specific Distortions

Julio Leal

In this paper, I seek to sharpen the understanding of the economic channels through which sector-specific distortions affect aggregate outcomes. I pay special attention on how cross-sector misallocation might arrive as a result of these distortions, and clarify how models of resource misallocation across heterogeneous firms differ from models of resource misallocation across sectors. I present facts that suggest the presence of such distortions in Mexico, a noteworthy developing country. The paper highlights the importance of understanding the origin and nature of sector-specific distortions for policy design. JEL codes: O1, O11

One of the recent advances in the economic development literature is the idea that firm’s idiosyncratic distortions produce resource misallocation. The two seminal works are Restuccia and Rogerson (2008) and Hsieh and Klenow (2007). The key idea is that the presence of distortions prevents the equalization of marginal products across firms, and, as a result, the allocation of labor and capital across these units is affected. This, in turn, reduces aggregate productivity and output.

For example, figure 1 presents the allocation of labor across different establishment sizes for Mexico and the United States. Mexico allocates a large fraction of employment in small low-productive establishments. In Leal Ordóñez (2014), I argue that the origin of these contrasting allocations can be traced down to the presence of informality in developing countries, which constitutes an implicit subsidy for low-productive activities.¹

This type of misallocation can naturally be present across any groups in the economy. In particular, one can also think of the existence of sector-specific distortions that produce cross-sector resource misallocation. This was first recognized by Jones (2011a, 2011b). A simple example of these type of distortions would be fiscal policies that differ across sectors. Similarly, some industries might operate in more concentrated markets than others which also represents a

¹ Recent papers on this topic include: Prado (2011), D’Erasmo and Moscoso Boedo (2012), Lopez-Martín (2015), and Lopez (2015).
form of sector-specific distortion. Table 1 presents an index by the OECD that measures the presence of policies and regulations that reduce competition in economic sectors. The table compares the best practices of OECD countries to the case of Mexico. It shows that regulation in Mexico is high and quite heterogeneous across sectors.

Figure 2 compares the cross-sector allocation of labor for Mexico and the United States. While differences are certainly expected in some industries due to the different stages of development between the two countries (e.g., in agriculture); there are other sectors with differing labor allocations, which discrepancies
could be the result of bad policy. This might be the case for Construction, Education, and Transportation, for example.

Is cross-sector misallocation of the same nature as misallocation across heterogeneous firms? In the next few pages, I intend to study the nature of cross-sector misallocation using a simple model that allows me to clarify the economic channels through which this phenomenon occurs. As I will show, for cross-sector misallocation to occur, the presence of an income effect that stems from the change in the rents associated to the distortions will be crucial.

I will argue that heterogeneous firms and multisector models of resource misallocation differ in one important aspect. In heterogeneous firms models, such as the one used by Restuccia and Rogerson, the allocation of resources depends mostly on the heterogeneity of productivity levels across firms with the demand side playing a nonimportant role. In contrast, in multisector models such as the one used by Jones, the allocation of resources is closely linked to the demand structure with the heterogeneity of productivity levels playing a small role.

Finally, I will also argue that a policymaker should focus on reducing extremely high distortions (relative to the average distortion) in order to obtain large gains in aggregate output. In fact, the gains in aggregate output of equivalent
changes in relatively low distortions are small. This is due to the convexity of the marginal productivity of labor.

This short paper is closely related to my recent work on input-output linkages and cross-sector misallocation—Leal (2015a, 2015b). In Leal (2015b), in particular, I provide a quantitative assessment of the effect of reducing sector-specific distortions in Mexico. The key difference is that, for this short paper, I focus on the decentralized equilibrium properties of a model without input-output linkages. This allows me to emphasize the key economic channels through which misallocation actually occurs. First, I will describe a baseline multisector model without distortions and its equilibrium properties. Then I will study how allocations are affected in the model when distortions are introduced. It turns out that a simple two-sector model is enough to capture the main economic channels that I seek to study.

A SIMPLE MODEL

There are two sectors in the economy. There is a representative household that consumes a composite of the two goods, and is endowed with $H$ units of labor. Preferences are given by $u(C)$, and leisure is not valued. Labor can be allocated to either sector without relocation costs. There is a producer of the composite that uses goods from the two sectors to produce, and has access to technology $Y(c_1, c_2) = c_1^{\beta_1} c_2^{\beta_2}$, with $\beta_1 + \beta_2 = 1$. In each sector $i$, there is a representative firm that has access to the following production function: $Q_i = A_i h_i^{a_i}$, where $h_i$ is hours, $A_i$ is labor productivity, and $a_i$ is the labor income share.

The problem of the household is to maximize utility subject to the budget constraint:

$$ C = wH + T, \text{ where } w \text{ is the wage rate, and } T \text{ are government transfers.}$$

Since no savings are allowed, the household simply consumes all her income. The composite producer maximizes profits taking good prices as given. This problem gives the optimal demand functions for the goods:

$$ p_i = \beta_i Y / c_i, \forall i $$

Note that these optimal demands exhibit a zero cross-price elasticity, that is, changes in $p_1$, do not affect the demand for $c_2$, and vice versa. Remember that when the price of sector $i$ changes, it produces both an income effect (IE), and a substitution effect (SE) in both demanded quantities $c_1$ and $c_2$. Suppose that, $p_1$ goes down. By the SE, the quantity consumed of $c_1$ increases, while the quantity consumed of $c_2$ decreases. The IE indicates that demand for both goods should go up.

2. Other papers on input-output linkages and aggregate outcomes include: Long Jr. and Plosser (1987); Horvath (1998, 2000); Acemoglu et al. (2012).

3. In contrast to the work in Leal (2015b), in this short paper I do not perform a quantitative assessment of the main arguments in it; instead, I use the paper to explore several different aspects of cross-sector misallocation, such as the asymmetry of effects on aggregate output when increasing or reducing a single distortion, while others remain constant (see the analysis below).
Thus, the IE reinforces the SE for the case of $c_1$, while it goes in the opposite direction for $c_2$. As is well known for the case of Cobb-Douglas preferences, the IE and the SE cancel each other for $c_2$, which renders the zero cross-price elasticity result.

Note that, by inspection of equation 1, the demand for $c_i$ depends linearly on income $Y$. In particular, note that, given the zero cross-price elasticity, an exogenous change in productivity of good 1, will only affect the demand curve for good 2 through its effects on aggregate income.

The profit maximization problems of the representative firms in each sector lead to the supply curves of the goods:

$$p_i = Q_i^{1-a} \left( \frac{1}{A_i} \right)^{\frac{1}{\alpha}} \frac{w}{a_i}, \forall i$$

Note that this curve depends on productivity and the wage rate. In particular, note that it depends linearly on the wage rate. The final piece of the equilibrium is market clearing. Markets for the composite good clear, $C = Y$; labor markets clear $h_1 + h_2 = H$; and markets in each sector clear. In the absence of distortions market clearing in sectors is given simply by $c_i = Q_i$, and, thus, aggregate income is given by $Y = wH$.

To keep the analysis as simple as possible and without loss of generality, I will focus on an economy where the coefficients of labor ($a_i$) in the sectoral production functions are set to 1. This simple model has an analytic solution. In particular, it can be shown that equilibrium labor for sector $i$ is:

$$h_i = \beta_i H,$$

which implies that output and consumption in each sector are given by $Q_i = c_i = A_i(\beta_i H)$. Thus aggregate output is:

$$Y = (A_1 \beta_1)^{\beta_1} (A_2 \beta_2)^{\beta_2} H,$$

in turn, wages and prices are given by: $w = Y/H$, and $p_i = w/A_i, i = 1,2$.

Numerical Analysis of Equilibrium

Column 1 of table 2 presents the equilibrium for a hypothetical baseline economy where the endowment of labor is normalized to 1. In this economy, the two sectors are symmetric, that is, both share the same productivity parameters ($A_1 = A_2 = 1$), and the same coefficients in the composite production function ($\beta_1 = \beta_2 = 1/2$). A few interesting features of the equilibrium emerge in this baseline economy. First note that since sectors are symmetric, they are allocated the same amount of labor in the equilibrium ($h_1 = h_2 = 1/2$). As a result, sectors produce the same quantities and have the same prices ($Q_1 = Q_2 = p_1 = p_2 = 1/2$). Note also that aggregate
income is equal to the wage rate ($Y = wH = w = 1/2$) because labor supply $H$ is 1. Finally note that there are no distortions and thus, no rents associated with such distortions ($T = 0$).

How do the equilibrium allocation and prices change when the economy experiences an increase in the productivity of sector 1? This is shown in the second column of the table. The exercise consists on doubling the productivity of sector 1. As expected, the price of sector 1 is reduced, and its production as well as aggregate output and wages increase. Interestingly, the allocation of labor is not affected across sectors by the change in productivity.

Understanding why the labor allocation is unaffected by changes in relative productivity across sectors is crucial to understand the economic channel through which misallocation actually occurs when distortions are introduced. From a social planner’s perspective, the allocation of labor is determined by the equalization of marginal products of labor across sectors:

$$
\left( \frac{\partial Y(c_1, c_2)}{\partial c_1} \right) \left( \frac{dc_1}{dh_1} \right) = \left( \frac{\partial Y(c_1, c_2)}{\partial c_2} \right) \left( \frac{dc_2}{dh_2} \right).
$$

The marginal product that is relevant in this case is how much more quantity of the composite do we obtain if we use an extra unit of labor in sector $i$. Since the total supply of labor is fixed, one has to reduce the amount of labor of sector 2 in order to increase the amount of labor in sector 1. Thus, increasing the amount of labor to sector 1 in one unit gives a gain which is the marginal product of $c_1$, but it also has a cost, which is the amount of output lost by diverting labor out from sector 2, that is, the marginal product of $c_2$. The quick answer to why a change in productivity of a single sector does not affect the allocation of labor is that the increase in the productivity, say, of sector 1 affects both sides of equation 3 proportionally. The increase in productivity of sector 1 also increases the marginal

### Table 2. Effect of duplicating the value of productivity $A_1$

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Increase in $A_1$</th>
<th>Symmetric distortions</th>
<th>No rents back</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>0.5</td>
<td>0.7071</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>$W$</td>
<td>0.5</td>
<td>0.7071</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>$b_1$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$b_2$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$Q_1$</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>$Q_2$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>$p_1$</td>
<td>0.5</td>
<td>0.3536</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.5</td>
<td>0.7071</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$T$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.25</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Notes:** The table shows equilibrium outcomes for different scenarios. The baseline corresponds to the case of no distortions with symmetric productivity parameters; the second column shows the effect of doubling the productivity in sector 1, while keeping everything else constant; the third column introduces symmetric distortions to the baseline assuming that rents are given back to the household; and finally, the fourth column assumes that the rents from distortions are lost.
product of sector 2 because $c_1$ and $c_2$ are complements; thus sector 2 has more $c_1$ to produce with! This result entirely depends on the specification of production functions, in this case, the Cobb-Douglas specification of both composite and sectoral production functions combine to produce this result.

But how can we understand this result from a decentralized equilibrium point of view? The key is in how the market for sector 2 is affected when the productivity of sector 1 increases. First, note that both demand and supply curves for the sector 2 market exhibit zero cross-price elasticities, that is, there is no direct transmission through the price of sector 1. Second, note that the demand curve depends linearly on aggregate income $Y$, and the supply curve depends linearly on the wage rate. Since the wage rate and aggregate income are equal in the case of no distortions, this means that both curves are shifted-up by the same factor. Thus, the equilibrium quantity of sector 2 is not altered by the increase in $A_1$. Consistent with this shift, the equilibrium price of sector 2 goes up. In summary, the reason why productivity does not alter the allocation of labor across sectors relies on two features of the supply and demand curves: that the cross-price elasticities are zero and that these depend linearly on income.

The introduction of distortions will disrupt this symmetry between demand and supply. The demand curve will respond to total income (which includes transfers), while the supply curve will continue to respond to wages. Thus, when transfers are present, the two curves won’t shift proportionally in response to changes in distortions, despite these are isomorphic to the productivity parameters.

**Distortions**

To understand this clearly, consider how the equilibrium is affected when distortions are introduced. In particular, consider distortions that act as output taxes. The sectoral firms first order conditions are affected in the following way (see the appendix for more details):

$$(1 - \tau_i)\hat{p}_iA_i = w. \tag{4}$$

Note that the marginal revenue goes down when $\tau_i$ increases. In column 3, table 2 presents the effect of introducing symmetric distortions in both sectors: $\tau_1 = \tau_2 = \tau = 1/2$. Note that the introduction of symmetric distortions does not alter the allocation of labor across sectors, neither the quantity of aggregate output, nor sector prices. There are several reason for this, the first one is that labor is inelastically supplied, thus, the introduction of taxes can not affect its supply. In fact, it is precisely this inelastic factor that bears all the tax burden: wages are reduced by 1/4, while the revenue from distortions $(T)$ increases in exactly that amount. In addition, note that the marginal productivity of labor in both sectors is affected by the same factor $\hat{f}(1 - \tau)$, thus keeping the equalization of marginal products, and, thus, the same labor allocation. In particular, note that both the marginal product of labor and the marginal cost are now half of what they were in the baseline.
One important, often neglected, feature of the model is that distortions produce rents \((T = \tau_1 p_1 Q_1 + \tau_2 p_2 Q_2)\), and the aggregate effects of distortions crucially depend on what we assume about the use of these rents in the economy. Note in particular, that one possibility is that, as we have assumed, the household gets the rents as transfers: \(C = wH + T\); but we could have also assumed that the rents are taken out of the country, in which case the household budget constraint would be \(C = wH\). The demand curves of sectors were not affected in the case of symmetric distortions because all the reduction in wages was compensated by the increase in rents, leaving aggregate income intact. However, this won’t be the case when the rents from distortions are lost. In this case, distortions are isomorphic to productivity, and aggregate output, goes down together with wages. This affects both demand and supply curves and the equilibrium effect is shown in column 4 of table 2. Note that in this case \(T = 0\), and the quantity produced in each sector is cut by half. Note that the allocation of labor remains unaltered. This will happen even if distortions are not symmetric across sectors. The reason is that distortions in this case act as reduction in productivity, and, as explained above, the allocation of labor is invariant to changes in relative productivity across sectors. The feasibility effect of distortions is nonetheless quite strong.

Finally, let’s analyze what would happen when distortions differ across sectors. It can be shown that in this case, the allocation of labor is altered in the following way (see the appendix):

\[
\frac{h_i}{H} = \frac{(1 - \tau_i)\beta_i}{(1 - \tau_1)\beta_1 + (1 - \tau_2)\beta_2}.
\]

That is, the higher the tax rate of sector \(i\), the less labor is allocated to that sector. When rents from distortions are given back to the household, there is an income effect along with changes in distortions. In particular, when distortions increase, there is more rents associated to them, and vice versa. When distortions differ across sectors, the allocation of labor is affected (one sector loses labor, the other one gains labor), and the equalization of marginal products is broken. As a result, aggregate output is lower than in the baseline because the reduction in output in the sector that loses labor is larger than the increase in output in the sector that gains labor (this is precisely a consequence of marginal products not being equalized).

Table 3 shows equilibrium output for different levels of distortions. Note that whenever tax rates are equal, the output level is undistorted, as in the baseline \((Y = 0.5)\). Invariably, every time the two tax rates differ, aggregate output is lower than in the baseline. Note that it could be the case that less output is obtained even when distortions are reduced. Start for example from a situation where distortions are symmetric but high, say, for example: \(\tau_1 = \tau_2 = 0.5\) (this corresponds to the cell right in the middle of the table). In this case, aggregate output is the same as in the baseline, because no misallocation is present. If one of the tax rates is reduced, aggregate output falls, because misallocation is
introduced. Remember that labor is in-elastically supplied, and, thus, it can not respond positively to lower distortions or taxes.

Figure 3 shows the pattern of equilibrium aggregate output for different levels of the distortion in sector 1, \( \tau_1 \), given the distortion in sector 2 fixed at \( \tau_2 = 0.5 \). As can be seen in the figure, the highest level of output is reached when the two distortions are equal to each other: \((\tau_1, \tau_2) = (0.5, 0.5)\). As the distortion in sector 1 increases, the loss in aggregate output increases as well. It is not easy for this model to exhibit substantial losses in aggregate output due to misallocation. For example, when distortions are the pair \((0.8, 0.5)\), aggregate output is still 90% of the value in the benchmark. When the distortion in sector 1 decreases, the fall in output is even smaller: when distortions are given by the pair \((0.2, 0.5)\), aggregate output is 97% of the baseline value. To obtain large deviations from the baseline level of aggregate

<table>
<thead>
<tr>
<th>( \tau_1 )</th>
<th>0</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
<td>0.4922</td>
<td>0.4714</td>
<td>0.4213</td>
<td>0.099</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4922</td>
<td>0.5</td>
<td>0.493</td>
<td>0.4583</td>
<td>0.1178</td>
</tr>
<tr>
<td>0.5</td>
<td>0.4714</td>
<td>0.493</td>
<td>0.5</td>
<td>0.4841</td>
<td>0.1386</td>
</tr>
<tr>
<td>0.7</td>
<td>0.4213</td>
<td>0.4583</td>
<td>0.4841</td>
<td>0.5</td>
<td>0.1767</td>
</tr>
<tr>
<td>0.99</td>
<td>0.099</td>
<td>0.1178</td>
<td>0.1386</td>
<td>0.1767</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Notes: The table shows equilibrium aggregate output for different combinations of distortion levels in the two sectors. When distortions are symmetric (the diagonal), aggregate output is unaffected, regardless of the distortion level.

**Figure 3. The Effect of Misallocation**

Notes: The figure shows how equilibrium aggregate output is affected when the distortion of one sector changes, while the one in the other sector remains constant. The figure suggests that reducing extremely high distortions brings large gains in aggregate output.
output, one of the distortions needs to be quite extreme (above .70 at least), and it has to be combined with a really low distortion in the other sector (lower than 0.3 at least). Still, not so big changes in distortions could lead to changes in aggregate output that are relevant at business cycles frequencies (changes of 1 or 2 percent).

Note when one of the tax rates is fixed, the effect of any two opposite equidistant pair of taxes from their baseline values—for example, (.5, .7) and (.5, .3)—will not be symmetric. This can be seen in figure 3: the negative effect on aggregate output of changes in $\tau_1$ is small for rates below 0.5 but is large for rates above 0.5. Why does this asymmetry occur? The reason relies on the convexity of the marginal productivity of labor: the slope is steeper for values of $h_2$ to the left of 1/2 than to the right. Given that figure 3 is constructed keeping the tax rate on sector 2 fixed, changing the tax rate in sector 1 is equivalent to moving along the marginal productivity curve of sector 2. When we move $\tau_1$ out of its baseline value of 0.5, we obtain that shifts to the right reduce the marginal productivity of $h_2$ less than what equivalent shifts to the left increase it. Thus, when $\tau_1$ increases, a substantial amount of labor is shifted out of sector 1 and into sector 2, without depressing much the marginal productivity of sector 2; in contrast, when $\tau_1$ decreases, the marginal productivity of sector 2 quickly goes up and the amount of labor allocated to this sector is lower. As a result, the amount of labor shifted and the associated misallocation effect turns out to be asymmetric in response to opposite changes in taxes.

**Conclusion**

So, what have we learned? This short paper sought to clarify the key economic channels through which sector-specific distortions affect aggregate outcomes. I have shown that these distortions can lead to resource misallocation across sectors if an income effect associated to the rents from distortions is present. In particular, if the rents from these distortions are “lost in the sea,” then no misallocation occurs. Nonetheless, in this last case, aggregate output is still undermined through a feasibility effect. For a quantitative assessment of the effect of reducing sector-specific distortions in an input-output model with cross-sector misallocation, see Leal (2015b).

The paper also clarifies the way in which misallocation across heterogeneous firms differs from misallocation across sectors. The allocation of labor across firms crucially depends on the heterogeneity in their productivity levels, while the demand structure plays a small role (see Restuccia and Rogerson (2008); Hsieh and Klenow (2009)); in contrast, for the allocation of labor across sectors is the opposite: it heavily depends on the demand structure, while the heterogeneity of productivity levels across sectors plays a small role.

Finally, the paper highlights the importance of understanding the origin of sector-specific distortions for policy design. For the case of Mexico, it suggests that improving the conditions of competition in Mexican industries could lead to important gains in aggregate output and productivity (see Leal 2015b).
Appendix

In this appendix, I focus on the decentralized equilibrium of this model. First, the problem of the household:

Household

$$\max u(C), \text{ s.t. } C = wH + T$$

Thus, provided that $u$ is increasing in $C$, the solution consists on consuming all income. Note that the income results from two sources: labor income and lump-sum transfers from the government. Note also that I am normalizing the price of the composite to 1.

Composite. The problem of the composite producer is as follows:

$$\max \{Y(c_1, c_2) - p_1 c_1 - p_2 c_2\}$$

The first order conditions are given by:

$$\beta \frac{Y}{c_1} - p_1 = 0$$

$$\beta \frac{Y}{c_2} - p_2 = 0$$

$$p_1 c_1 + p_2 c_2 = Y$$

These conditions lead to well-known relationships between, prices, and consumption:

$$\frac{c_i}{c_j} = \frac{\beta_j p_j}{\beta_i p_i},$$

also, consumption shares are constant:

$$\beta_i = \frac{p_i c_i}{Y}, \quad (5)$$

and optimal demands exhibit a zero cross-price elasticity:

$$c_i = \beta_i Y / p_i, \quad (6)$$

that is, changes in $p_1$, do not affect the demand for $c_2$. Note that, by inspection of equation 6, the demand for $c_i$ depends linearly on income.
Firm’s problem. The problem of the firm is as follows:

$$\max_{h_i} \{(1 - \tau_i)p_iA_i h_i - wh_i\},$$

thus, the first order conditions associated with this problem are:

$$(1 - \tau_i)p_iA_i = w. \tag{7}$$

Also note that if we divide the first order conditions for the two firms $i = 1, 2$, a relationship between relative prices and relative productivity emerges:

$$\frac{p_i}{p_j} = \frac{(1 - \tau_i)A_i}{(1 - \tau_j)A_j}.$$ 

Market clearing. The labor market, and the market for the composite, clear:

$$h_1 + h_2 = H,$$

$$C = Y$$

Putting together the market clearing condition for the composite, and the household’s budget constraint, I obtain: $Y = C = wH + T$.

However, for the case of the good’s market in each sector, the market clearing depends on whether the rents associated with the distortion are given back to the household or not. Let’s focus first on the case when the revenue is lost. In this case $T = 0$ in the budget constraint of the household problem, which implies that the revenue from taxes is not available for consumption, thus the market clearing should be:

$$p_i c_i = (1 - \tau_i)p_iQ_i, \quad \forall i = 1, 2. \tag{8}$$

Equilibrium. With this at hand we can analyze the equilibrium. As a reference point let’s start with the analysis of the undistorted economy where $\tau_i = 0$, $i = 1, 2$ and $T = 0$. Substituting the market clearing condition of sector $i$, in the first order conditions of the firm:

$$p_iA_i h_i = wh_i$$

$$\Leftrightarrow p_iQ_i = wh_i$$

$$\Leftrightarrow p_i c_i = wh_i,$$
substitute the first order condition 5, $p_i c_i = \beta_i Y$ into the above expression:

$$\iff \beta_i w H = w b_i$$

$$\iff h_i / H = \beta_i$$

This gives the allocation for labor in equilibrium, which is the same as the allocation in the social planner’s problem (not shown). Using this equilibrium allocation, I can obtain equilibrium quantities for production: $Q_i = A_i b_i = A_i \beta_i H$; for consumption: $c_i = Q_i = A_i \beta_i H$; and for aggregate output:

$$Y = (A_1 \beta_1) b_1 (A_2 \beta_2) b_2 H.$$  

This also implies that the wage is equal to: $w = Y / H = (A_1 \beta_1) b_1 (A_2 \beta_2) b_2$, and prices are given by $p_i = w / A_i$, $i = 1, 2$.

**The distorted economy** Now consider the distorted economy, with $\tau_i > 0$, $i = 1, 2$, and $T = 0$, that is, the case (a) above. In this case, tax rates are isomorphic to sectorial productivities $A_i$, and aggregate output is $Y = (\tilde{A}_1 \beta_1) (\tilde{A}_2 \beta_2) H$, where $\tilde{A}_i = (1 - \tau_i) A_i$. As a result, the equilibrium allocation of labor is not affected by changes in taxes, because of the same reasons described above. The result is quite intuitive, if the government grabs a fraction of output from the firms, and does not give it back, the effect is similar to a reduction in productivity.

Finally, consider the case in which taxes are positive, but the revenue is given back to the household. In this case, we have to take into account the extra income effect in the demand for sectorial goods that arises through lump-sum transfers. The allocation of labor in this case is given by:

$$\theta_i = h_i / H = \frac{(1 - \tau_i) \beta_i}{(1 - \tau_1) \beta_1 + (1 - \tau_2) \beta_2},$$

which is affected by changes in taxes. Note that the higher the tax rate of sector $i$, the less labor is allocated to it. Also note that when $\tau_1 = \tau_2$, $\theta_i = \beta_i$, the allocation of labor is efficient. This result is independent of the level of taxes. Thus, in contrast to case (a), the efficient allocation can be achieved even with positive tax rates as long as these do not differ across sectors.

**The role of misallocation** To understand the importance of misallocation of resources across sectors, note that in the distorted economy with transfers, aggregate output is given by:

$$Y = (A_1 \theta_1) b_1 (A_2 \theta_2) b_2 H.$$  

Notice that if we start with an economy with identical tax rates (an efficient one), a reduction in the tax rate of one of the sectors decreases aggregate output.
This counterintuitive result is present because as you increase the dispersion between the two tax rates, more misallocation of labor occurs. To see this clearly, consider the following example. Assume you start from a level $\tau$ of taxes in both sectors, then $\theta_i = \beta_i$, $i = 1,2$, and marginal productivities are being equalized across sectors. If you decrease the tax rate in sector 1, the equilibrium will allocate more labor to this sector ($\theta_1 > \beta_1$), and less labor to sector 2 ($\theta_2 < \beta_2$), which will break the equalization of marginal productivities. In consequence, the gain in output $Y$ associated to that extra units of labor in sector 1 will not be enough to compensate for the loss in output associated with the reduction of hours in sector 2.

The effect of misallocation can be captured by comparing the the log aggregate output in the social planner’s problem vs. the expression for log aggregate output in the distorted-economy case. Taking logs:

$$
\ln Y = \beta' a + \beta' \ln \theta + \ln H.
$$

And the effect of misallocation is captured by the term:

$$
Y - Y^e = \beta'[\ln \theta - \ln \beta] < 0.
$$

Note that when $\tau_1 < \tau_2$, $\frac{\beta_1}{\theta_1} < 1$, and $\frac{\beta_2}{\theta_2} > 1$, so the proof that $Y < Y^e$ is not trivial. It uses the fact that the arithmetic mean of a variable is greater than or equal to the geometric mean, which in turn is an application of Jensen’s inequality. Also useful for future reference is to have a general expression of the change in aggregate output in response to a change in one of the tax rates $\tau_i$. Let $Y^0$ be the aggregate output level before the tax change, and $Y^1$, the corresponding one after the change, then:

$$
\ln \left( \frac{Y^1}{Y^0} \right) = \beta' \ln \left( \frac{\theta^1}{\theta^0} \right) = \sum_{i=1}^{2} \beta_i \ln \left( \frac{\theta_i^1}{\theta_i^0} \right).
$$

Output will increase if the movement in $\tau_i$ is in the direction that brings the two tax rates closer to each other, otherwise, output will decrease.

**References**


From Firm Productivity Dynamics to Aggregate Efficiency

Bernabe Lopez-Martin

I construct a quantitative framework to evaluate how financial constraints can reduce productivity growth at the firm level and result in lower aggregate productivity. I consider a model where firms are able to invest in innovation in order to increase their productivity, or knowledge capital. This investment is a costly and uncertain enterprise. As the capacity to obtain external funds is diminished, resources allocated to this effort will be reduced due to different mechanisms at work. First, the return of this investment in the case of success may be diminished by the inability to quickly increase production capacity if the credit necessary to do so is scarce (i.e., if entrepreneurs cannot rent the optimal level of physical capital). Second, financial constraints reduce profits obtained by entrepreneurs and therefore the amount of assets they are able to accumulate in every period. Finally, financially underdeveloped economies will be characterized by a lower average ability of entrepreneurs. This is due to the lower equilibrium wage in the economy, which results in a larger mass of individuals opting to set up firms. In the margin, these individuals tend to have lower ability to manage a firm and relatively low prospects of generating firm productivity growth through innovation. JEL codes: O11, O16, O30, O4, E23

I. Quantitative Model

The model builds upon the frameworks of occupational choice and heterogeneous entrepreneurial ability of Lucas (1978) and industry dynamics of Hopenhayn.
There is a continuum of individuals who possess heterogeneous innate entrepreneurial ability, with some probability they have an opportunity to become entrepreneurs and establish a firm. All individuals earn the same wage as workers, since there is no heterogeneity in their effective units of labor and workers are perfectly mobile across firms. Entrepreneurs can allocate resources to investment in technology through a controlled stochastic process (Pakes and McGuire 1994; Klette and Kortum 2004; Farias et al. 2012; Doraszelski and Jaumandreu 2013): entrepreneurs decide every period the amount of resources invested in innovation, which determines the probability of an increase in firm productivity.

Preferences

Time is discrete and a period, indexed by $t$, represents a year. Individuals value the consumption of the final good, denoted $c_t$, through lifetime and intratemporal preferences represented as follows:

$$U = \mathbb{E}\left[ \sum_{t=0}^{\infty} (\beta(1 - \mu))^t u(c_t) \right] \quad \text{and} \quad u(c_t) = \frac{c_t^{1-\sigma}}{1-\sigma}$$

where $\beta$ is the discount factor, $\sigma$ is the coefficient that governs risk aversion. The probability that an individual dies in any given period is $\mu$, so that the effective discount factor is $\beta(1 - \mu)$. When an individual dies, he is immediately replaced by another individual with the same entrepreneurial ability $\phi$ so that the mass of individuals and their distribution over ability is constant.

Production Technology

Production of the final good is carried out by single establishment firms and each firm is managed by its owner. Individuals possess innate and permanent entrepreneurial ability $\phi$, received according to a distribution $h(\phi)$. Entrepreneurs are able to accumulate knowledge capital denominated $n$ (through a process described below) and have access to a decreasing returns to scale production technology with respect to capital and labor:

$$q = (\phi n)^{1-\nu} f(k, l)$$

with $f(k, l) = k^\alpha l^{1-\alpha}$,

where $k$ is capital and $l$ is labor used in production. Following Lucas (1978), $\nu \in (0, 1)$ is the span-of-control parameter that determines the decreasing returns to scale in the production technology.

1. Differences in management quality are an important determinant of productivity differences across firms (see Bartelsman and Doms 2000; Foster et al. 2001; Syverson 2011).
Entrepreneurs can invest in the innovation good $x$ to increase the stock of knowledge capital. Three outcomes are possible every period, depending on the amount of investment in the innovation good in the previous period: knowledge capital may increase by a proportion $\Delta$, it may remain constant, or decrease by $\Delta$. Knowledge capital is defined on the grid $\{n, n(1 + \Delta), n(1 + \Delta)^2, ..., \bar{n}\}$, where $n$ and $\bar{n}$ are the lowest and highest possible levels of knowledge capital, respectively. The probability of a successful outcome is given by:

$$P(n' = n(1 + \Delta) | n, x) = (1 - \gamma) \frac{(1 - \lambda)a(x/n)}{1 + a(x/n)} + \gamma. \quad (3a)$$

There are diminishing returns to innovation investment $x$. Fixing a probability of success in innovation, $P(n(1 + \Delta) | s, x)$, the necessary investment in innovation goods $x$ to increase the productivity of the firm by a fixed percentage is proportional to knowledge capital $n$. The probability of a negative outcome is given by:

$$P(n' = n/(1 + \Delta) | n, x) = \frac{(1 - \gamma)\lambda}{1 + a(x/n)}. \quad (3b)$$

Knowledge capital summarizes the history of investment and success in innovations and governs the size of the firm (Klette and Kortum 2004). Furthermore, it is lost when the firm closes, regardless of whether exit is due to an exogenous exit shock or the entrepreneur finds it optimal to close the firm. Finally, knowledge capital is assumed to be firm-specific and there is no market for its trade.

**Workers**

We group state variables by letting $s = (\varphi, n, b)$, where $\varphi$ is the individual’s permanent entrepreneurial ability, $b$ are financial assets and $n$ is knowledge capital. Additionally, $z \in \{e, w\}$ denotes whether the individual is an entrepreneur or a worker, respectively. The problem of the worker amounts to a savings decision.

---


3. The model can be extended to consider unbounded knowledge capital, which would require additional conditions to guarantee a well defined dynamic program and convergence in the stationary distribution (see Atkeson and Burstein, 2010). We select the upper bound so that a negligible amount of firms reach this level.

4. It can be verified that optimal labor and capital inputs, output and profit are proportional to knowledge capital $n$ under the production function previously specified in the case of no financial restrictions.
and determining the conditions under which it is optimal to establish a firm:

\[
v_w(s) = \max_{\{c, b' \geq 0\}} u(c) + \beta(1 - \mu) \sum_{\{e'\}} Q(e')v(s') \\
\text{s.t.} \quad c + b' = w + (1 + r)b \quad \text{and} \quad n = n_w
\]  

(4)

At the beginning of each period, workers face their occupational choice \(v(s) = \max\{v_e(s \cdot \varphi, n, b), v_w(s)\}\), where the value of becoming an entrepreneur is represented by \(v_e(s \cdot \varphi)\). The worker is free to continue in the labor market and earn a wage \(w\) every period, or become an entrepreneur if he is given an opportunity to start a firm with probability \(Q(s') = 1\) equal to \(\theta\). When individuals re-enter the labor market, their knowledge capital is reset to \(n_w\), which follows from the interpretation that knowledge capital represents an intangible asset embedded in the firm. All new-born individuals receive an entrepreneurial ability \(\varphi\) from the distribution \(h(\varphi)\) and initial assets \(b\). In this setup occupational choice depends on the ability of the individual as an entrepreneur but may also depend on financial wealth, necessary to reach a sufficiently profitable scale of production when financial constraints are present.

**Entrepreneurs**

Given the level of knowledge capital and financial wealth, profits for an entrepreneur are given by:

\[
\pi(s) = \max_{\{k, l\}} q - (r + \delta)k - wl \quad \text{s.t.} \quad k \leq \overline{k}(s),
\]

where the choice of capital input is restricted by an endogenous collateral constraint, which we describe below. Additionally, the entrepreneur may return to the labor market in any period:

\[
v_e(s) = \max_{\{c, x, b' \geq 0\}} u(c) + \beta(1 - \mu) \sum_{\{x'\}} P(n'\mid n, x) \max\{v_w(s'), v_e(s')\} \\
\text{s.t.} \quad c + b' + x = \pi(s) + (1 + r)
\]

(5)

The entrepreneur is able to invest in the knowledge capital of the firm as long as the firm is active, but is lost if the individual decides to return to the labor market.

**Financial Markets: Endogenous Collateral Constraints**

In specifying the collateral constraints we follow Amaral and Quintin (2010) and Buera et al. (2011). At the beginning of a period the entrepreneur makes a deposit \(b\) and rents capital \(k\) from a financial intermediary. At the end of the period, the entrepreneur receives his deposit earning interest rate \(r\) and pays the
cost of capital rental at the total rate of $r + \delta$. Borrowing and capital rental are realized within a given period and the assets of the individual are restricted to be positive $b \geq 0$ in all periods.

Entrepreneurs may renege on financial contracts after production has taken place. If this occurs, the entrepreneur keeps a fraction $(1 - \psi)$ of the undepreciated capital and the revenue net of labor payments. The punishment for default is the loss of the financial assets deposited with the financial intermediary $b$. Entrepreneurs regain access to financial markets in the following period. This implies that a static condition determines enforceable allocations, allowing for the consideration of financial constraints in a tractable manner. In this setup parameter $\psi$ indexes enforcement of financial contracts in the economy, which encompasses economies with no credit when $\psi = 0$ and perfect credit markets if $\psi = 1$.

The analysis is restricted to financial contracts that are incentive-compatible, there is no default in equilibrium. Effectively, imperfect enforcement of financial contracts determines an upper bound $\bar{k}(s)$ on the amount of capital that entrepreneurs are able to borrow. Mathematically the financial constraint can be described as follows. In the case of no-default the entrepreneur receives profits plus interest rate income from financial assets:

$$\max_{\{l\}} q - wl - (r + \delta)k - x + (1 + r)b,$$

while in the case of default the entrepreneur would receive (off-equilibrium):

$$\max_{\{l\}} (1 - \psi)(q - wl + (1 - \delta)k) - x.$$

Capital rental is said to be enforceable if and only if it satisfies (6) $\geq$ (7). Note that $x$ does not distort the bound of enforceable capital. The borrowing limit is increasing in financial wealth since the loss of collateral is greater in the case of default. It is also increasing in firm productivity, as only a share of output is kept in the case of default (see Amaral and Quintin 2010; Buera et al. 2011).

Equilibrium

The state space is given by $(\varphi, n, b, z)$, we previously defined $s = (\varphi, n, b)$ and $z \in \{e, w\}$. Given an interest rate $r$, an open economy stationary competitive equilibrium consists of: quantities and production inputs $\{q(s), l(s), k(s)\}$ and profits $\{\pi(s)\}$, policy functions for financial asset accumulation $\{b'(s, z)\}$, policy function for investment in the innovation good $\{x(s)\}$, a wage $w$, and an invariant measure $M(s, z)$ of individuals over the state space, such that: workers and entrepreneurs solve the dynamic problems $\{v(s), v_x(s), v_w(s)\}$ and profit maximization, the market clearing condition in the labor market holds (the demand of labor by entrepreneurs equals the mass of workers), the measure $M(s, z)$ is
consistent with policy functions, optimal decision rules and the stochastic process for knowledge capital.

II. Parameters

The parameters of the model are divided into two groups: a first group of standard parameters taken from the literature (see table 1), a second group of parameters that are set to match key features of the US economy (the baseline economy, table 2).

Predicted Parameters

The predetermined parameters are taken from different models of firm dynamics: Atkeson and Kehoe (2005), Amaral and Quintin (2010), Buera et al. (2011).

<table>
<thead>
<tr>
<th>Table 1. Predetermined Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of parameter</td>
</tr>
<tr>
<td>Effective discount factor</td>
</tr>
<tr>
<td>Risk aversion</td>
</tr>
<tr>
<td>Interest rate (open economy assumption)</td>
</tr>
<tr>
<td>Span-of-control</td>
</tr>
<tr>
<td>Income share of capital</td>
</tr>
<tr>
<td>Capital depreciation rate</td>
</tr>
<tr>
<td>Innovation technology</td>
</tr>
<tr>
<td>Innovation technology</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Calibrated Parameters and Target Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibrated parameters</td>
</tr>
<tr>
<td>Exogenous exit rate</td>
</tr>
<tr>
<td>Entrepreneurial opportunity probability</td>
</tr>
<tr>
<td>Pareto distribution of entrepreneurial ability</td>
</tr>
<tr>
<td>Innovation technology</td>
</tr>
<tr>
<td>Initial knowledge capital</td>
</tr>
<tr>
<td>Size innovation steps</td>
</tr>
<tr>
<td>Endogenous collateral constraint</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target statistics</th>
<th>data</th>
<th>model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total firm entry (and exit) rate</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>Exit rate firms 20+ workers</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Std. deviation growth rates</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>Relative size firms 25–30/1–5 years (survivors)</td>
<td>2.69</td>
<td>2.62</td>
</tr>
<tr>
<td>Employment at firms w/500+ workers</td>
<td>0.30</td>
<td>0.29</td>
</tr>
<tr>
<td>Knowledge capital investment/total output (%)</td>
<td>4.40</td>
<td>4.70</td>
</tr>
<tr>
<td>Private credit/output</td>
<td>2.30</td>
<td>2.34</td>
</tr>
</tbody>
</table>
Midrigan and Xu (2014), Restuccia and Rogerson (2008), among many others. The values of two of the innovation technology parameters, $a$ and $\lambda$, are taken from Farias, Saure, and Weintraub (2012).

**Calibrated Parameters**

In the model, the total exit rate includes the entrepreneurs that decide to close their firms and those that exit due to an exogenous shock. The exogenous exit rate $\mu$ is set to match the exit rate of firms with more than 20 employees for the US total business sector (see fig. 1.2 in Bartelsman et al. 2009). Entrepreneurial ability is distributed according to a discretized Pareto distribution with nine possible values; its parameter is closely related to the size distribution of firms in the model; our target is the share of employment in firms with more than 500 workers. Parameter $\vartheta$ represents the probability that a worker receives an opportunity to set up a firm; this parameter is linked to the total entry rate of firms (in a stationary equilibrium the total and entry rates are equalized).

The technology accumulation parameters mainly determine the life-cycle growth of firms. I target the average size of firms that are 25–30 years relative to firms that are younger than 5 years (Hsieh and Klenow 2014). Additionally, McGrattan and Prescott (2010) estimate the ratio of business intangible investment to *total adjusted domestic product* to be 0.044 for the United States. Parameter $\Delta$ governs the size of the fluctuations in productivity. We follow Atkeson and Burstein (2010), albeit with a different specification for the accumulation of knowledge capital, and target the standard deviation of the growth rate of employment of large firms in the model, taking the statistic computed for publicly traded firms in the United States for the period 1980–2001 from Davis et al. (2007).

Parameter $\psi$ represents financial development and the access to credit for firms. As is standard in the literature, we target the ratio of private credit provided by financial institutions and private bond markets over GDP (Beck et al. 2009). The value for the US results in an economy with perfect financial markets in the model: the average of the ratio for the years between 1993 and 2001 is 2.30 (this covers the period of the data used to impute firm life-cycle growth in Hsieh and Klenow 2014).

### III. Quantitative Analysis

The main purpose of our model is to help us understand the channels through which financial development affects aggregate productivity in a model with endogenous firm level productivity. In the presented framework, financial constraints affect the growth of productivity at the firm level and can potentially explain firm life-cycle productivity growth differences between developed and developing economies.

The quantitative exercise consists in reducing parameter $\psi$ to approximately match the volume of credit relative to output in a developing economy such as...
Mexico. The main results for aggregate statistics are presented in Table 3: the fall in total output is 36.4 percent, explained by a fall in total aggregate productivity of 10.3 percent and a decrease in capital to approximately one third of the developed economy.

To gain insight on the sources of the productivity loss we can decompose the weighted average firm level productivity as follows:

$$v = \frac{\sum \omega_i \cdot (q_i/Q)}{Q} = \bar{v} + \text{cov}(\omega, q/Q)$$

with $\omega = (\varphi n)^{1-v}$,

where $\hat{v}$ is average firm level productivity weighted by the output share of the firm $q_i/Q$ and $\bar{v}$ is the arithmetic mean of firm level productivity. This decomposition provides a measure of the efficiency of resource allocation in the economy: the covariance term reveals to what extent more productive firms in the economy have a larger share of output (see D’Erasmo et al. 2012). As can be seen in this model, both the weighted and the unweighted averages for firm productivity are lower in the economy with less access to credit as well as the covariance term (results in Table 4).

The impact on firm life-cycle productivity growth is considerable: in the developed economy the ratio of productivity for firms 25–30 years of age relative to the same firms when they were 1–5 years of age is 2.75, while this ratio is reduced to almost half to 1.31 in the less financially developed economy.

### Table 3. Main Quantitative Results: Aggregate Statistics

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>Mexico</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic credit to private sector/output</td>
<td>2.336</td>
<td>0.216</td>
</tr>
<tr>
<td>Output per worker</td>
<td>1.000</td>
<td>0.636</td>
</tr>
<tr>
<td>Capital per worker</td>
<td>1.000</td>
<td>0.314</td>
</tr>
<tr>
<td>Total factor productivity</td>
<td>1.000</td>
<td>0.897</td>
</tr>
<tr>
<td>Wage</td>
<td>1.000</td>
<td>0.642</td>
</tr>
</tbody>
</table>

### Table 4. Main Quantitative Results: Firm Level Statistics

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>Mexico</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted average firm productivity</td>
<td>1.745</td>
<td>1.425</td>
</tr>
<tr>
<td>Unweighted average firm productivity</td>
<td>1.191</td>
<td>1.105</td>
</tr>
<tr>
<td>Covariance: weights and productivity</td>
<td>0.554</td>
<td>0.320</td>
</tr>
<tr>
<td>Relative productivity firms 25–30/1–5 years (survivors)</td>
<td>2.748</td>
<td>1.313</td>
</tr>
<tr>
<td>Employment at firms w/500+ workers</td>
<td>0.289</td>
<td>0.182</td>
</tr>
</tbody>
</table>

5. It can be shown that in this type of models TFP is obtained from the equations $Q = A(K^a L^{1-a})^a$, where $L$ is the mass of workers (see Midrigan and Xu 2014).
IV. Conclusion

The stylized equilibrium framework proposed provides a link between financial constraints, productivity growth at the firm level, and aggregate productivity. This can contribute to our understanding of the lower life-cycle productivity growth of firms in developing economies as recently emphasized in the literature (Hsieh and Klenow 2014). The proposed framework can be extended along several dimensions. In ongoing research, I explore a model which introduces the informal sector. Through the general equilibrium mechanism described previously, lower wages lead to an increase in the number of low-productivity low-growth firms in the informal sector which can potentially amplify the negative impact of financial constraints on the aggregate economy.

References


Psychometrics as a Tool to Improve Credit Information

Irani Arraíz, Miriam Bruhn, and Rodolfo Stucchi

This paper studies the use of psychometric tests, designed by the Entrepreneurial Finance Lab (EFL), as a tool to screen out high credit risk and potentially increase access to credit for small business owners in Peru. We compare repayment behavior patterns across entrepreneurs who were offered a loan based on the traditional credit-scoring method versus the EFL tool. We find that the psychometric test can lower the risk of the loan portfolio when used as a secondary screening mechanism for already banked entrepreneurs—that is, those with a credit history. The EFL tool can also allow lenders to offer credit to some unbanked entrepreneurs—that is, those without a credit history—who were rejected based on their traditional credit scores, without increasing the risk of the portfolio. JEL Classification: D82, G21, G32

I. Introduction

Given the important role that small and medium enterprises (SMEs) play in a healthy and dynamic economy, many studies have attempted to understand the factors that affect their creation and performance.¹ These studies show that SMEs face greater financial constraints than large firms and that these constraints could be one of the factors that limit their growth (Beck, Demirgüç-Kunt, and Maksimovic 2008; Cavallo, Galindo, and Izquierdo 2010; Ibarraran, Maffioli, and Stucchi 2010; Mateev, Poutziouris, and Ivanov 2013). SMEs face greater financial constraints in part because they are subject to information asymmetries that are less salient for large firms. SMEs often lack audited financial statements and

¹ Numerous studies have documented the important role played by SMEs in the process of economic development (see, for example, Beck, Demirgüç-Kunt, and Levine 2005; Beck and Demirgüç-Kunt 2006; Ayyagari, Beck, and Demirgüç-Kunt 2007; Nichter and Goldmark 2009; Liedholm and Mead 2013).
other information about their operations, and as a result, financial institutions have difficulties assessing the risk of lending to them (de la Torre, Pería, and Schmukler 2009).

This paper studies a novel intervention that aims to enhance the information on SMEs useful to potential lenders. A large body of literature has examined the role of information sharing, credit bureaus, and credit scoring in increasing credit to SMEs (see, for example, Berger, Frame, and Miller 2005; Brown, Jappelli, and Pagano 2009; Love and Mylenko 2003; Martinez Peria and Singh 2014). Not all countries, however, have well-developed credit bureaus that gather the level of information needed to build a reliable credit-scoring model. For example, the average credit bureau in Latin America and the Caribbean complies with only half of best practices and covers only 39.3 percent of the adult population (Doing Business Report 2014).

Thus, even though credit scoring can potentially improve SMEs’ access to credit, it can take many years to pass legislation that will lead to improvements in the quality and depth of information recorded by credit bureaus. In addition, banks may be reluctant to share proprietary information with other banks (Bruhn, Farazi, and Kanz 2013), and even after credit bureaus are set up and are working well, building an accurate credit-scoring model often requires many years of credit history. In the meantime, credit markets in developing countries may have to rely on alternative lending technologies to screen potential clients.

The Entrepreneurial Finance Lab (EFL) has developed an alternative credit information tool that can potentially be used by lenders to better screen loan applicants. This tool uses a psychometric application to predict entrepreneurs’ repayment patterns. This study looks at the effectiveness of this tool in reducing the risk of lending to SMEs in the context of a pilot exercise conducted in Peru. The financial institution participating in the study, the fifth-largest commercial bank in Peru, piloted the EFL tool, with the goal of expanding its lending to SMEs. Loan applicants were screened by the EFL tool, and all applicants that achieved a score higher than a threshold set by the bank were offered a loan.

Peru has several private credit bureaus that, together, cover 100 percent of the adult population. Thus, all loan applicants have a traditional credit score. But for individuals who have not previously taken out a loan from a formal financial institution, this score is based primarily on demographic information rather than actual credit history. For the purposes of this study, these individuals are referred to as “unbanked.” Applicants with traditional credit scores in an acceptable range (as defined by the implementing institution) were offered a loan even if their EFL score was below the threshold.

This setup allowed the researchers to test two possible uses of the EFL tool: (i) as a secondary screening mechanism for entrepreneurs accepted under the traditional credit-scoring method, to lower the risk of the SME loan portfolio; and (ii) as a skimming mechanism for applicants rejected under the traditional credit-scoring method, to offer more loans without increasing the risk of the portfolio. We used data on formal credit repayment behavior patterns, as collected by the Superintendencia de Banca y Seguros (SBS) in Peru.
Our results show that the EFL tool can reduce portfolio risk for “banked” entrepreneurs (i.e., those who have taken out loans from a formal financial institution) when it is used to complement traditional credit scores. We did not find evidence that the EFL tool can reduce the risk of the portfolio for unbanked entrepreneurs who have been approved through the traditional screening process. We also found that the EFL tool can be used to extend credit to some unbanked entrepreneurs who were rejected based on their traditional credit scores, without increasing the risk of the portfolio. However, for banked entrepreneurs, the EFL tool does not perform well as a skimming mechanism in the context examined in this paper.

Related studies by EFL researchers have shown that entrepreneurs’ business profits and repayment behavior patterns are strongly correlated with their personality traits (Klinger, Khwaja, and del Carpio 2013; Klinger, Khwaja, and LaMonte 2013; and Klinger et al. 2013). In this paper, we go one step further and examine the potential of the psychometric credit application as a tool to manage portfolio risk compared with a traditional credit-scoring method. Our paper is the first external study examining the predictive power of psychometric credit scoring; that is, it uses independently collected data on repayment behavior patterns and is not coauthored by a person affiliated with EFL.

II. BACKGROUND AND ANALYTICAL FRAMEWORK

Innovative Screening Methods: The EFL Tool

EFL developed a psychometric credit-scoring tool by first quantifying the individual characteristics of people who had defaulted on a past loan versus those who had not and of people who owned small businesses with high versus low profits. The characteristics were put in three categories: personality, intelligence, and integrity (Klinger, Khwaja, and del Carpio 2013). The EFL researchers’ hypothesis was that these assessments would allow them to identify the two main determinants of an entrepreneur’s intrinsic risk: the ability to repay a loan and the willingness to do so. Entrepreneurial traits, measured via personality and intelligence tests, determine an entrepreneur’s ability to generate cash flows in the future—cash flows that can, in turn, be used to repay any debt owed. Honesty and integrity traits, measured via the integrity test, determine the entrepreneur’s willingness to pay, independent of the ability to do so.2

After identifying questions that could potentially predict credit risk and trying out a first prototype of their tool, the EFL researchers developed a commercial application based on the responses to their tool and subsequent default behavior. The commercial application contains psychometric questions developed

2. Several papers have documented links between personality or intelligence tests and entrepreneurship or business performance (Ciavarella et al. 2004; De Mel, McKemzie and Woodruff 2010; Djankov, McLiesh, and Shleifer 2007; Zhao and Seibert 2006). To date, the only evidence on integrity and willingness to repay loans comes from EFL itself (Klinger, Khwaja, and del Carpio 2013).
internally and licensed by third parties relating to individual attitudes, beliefs, integrity, and performance.

The Implementation of the EFL Tool

In March 2012, the implementing institution started to pilot EFL’s psychometric credit-scoring model, with the objective of expanding its commercial lending to SMEs. Entrepreneurs who applied for a working capital loan (up to 18 months in duration with an average loan size of $3,855) were screened by the EFL tool as part of the application process. The EFL credit application used at this time took about 45 minutes to complete (the current version takes 25 minutes). To be approved for a loan, the entrepreneur either had to score above the threshold (defined by the institution) on the EFL application or had to be approved under the institution’s traditional screening method. Only entrepreneurs who were rejected under both screening methods were not offered a loan (table 1).

Hypotheses

We considered two ways that banks can apply the EFL tool in their credit-risk management and lending decisions and tested two corresponding hypotheses by comparing the repayment patterns of the different groups listed in table 1, separately for banked and unbanked entrepreneurs.

Hypothesis 1: Risk reduction. Entrepreneurs who were accepted under the traditional method but rejected based on their EFL score display poorer loan repayment behavior patterns than entrepreneurs who were accepted under both methods. Looking at table 1, this hypothesis implies that entrepreneurs in quadrant 3 have poorer repayment patterns than entrepreneurs in quadrant 1. If this hypothesis is true, the EFL credit application can be used as a secondary screening mechanism to lower the risk of the SME loan portfolio.

Hypothesis 2: Credit to new borrowers. Entrepreneurs who were rejected under the traditional method but accepted based on their EFL score do not display poorer loan repayment behavior than entrepreneurs who were accepted under the traditional model. In terms of table 1, this hypothesis implies that entrepreneurs in quadrant 2 have no poorer repayment patterns than do entrepreneurs in quadrants 1 and 3. If this hypothesis is true, banks can rely on the EFL

| Table 1. Credit Decisions Based on the EFL Score and Traditional Credit Score |
|---------------------------|------------------------|------------------------|
|                         | Accept screening method | Reject screening method |
|                         | Accept                | Reject                |
| EFL decision             | Accept                | (1) Accepted 659 entrepreneurs (20.6% unbanked) | (2) Accepted 158 entrepreneurs (10.1% unbanked) |
|                         | Reject                | (3) Accepted 860 entrepreneurs (25.1% unbanked) | (4) Rejected 209 entrepreneurs (7.2% unbanked) |

Downloaded from https://academic.oup.com/wber/article-abstract/30/Supplement_1/S67/2897404 by World Bank and IMF user on 08 August 2019
tool to help them offer credit to applicants they would otherwise have rejected, without increasing the risk of their SME portfolio.

III. Data

We obtained data collected by an EFL questionnaire that the implementing bank administered to 1,993 loan applicants between March 2012 and August 2013. These data include the EFL score and the date when the entrepreneur was screened by the EFL tool, as well as the applicant’s age, gender, marital status, business sales, and sector of activity. The implementing institution shared with us the threshold EFL score it used to determine whether or not to offer a loan. For each applicant, the institution also let us know which decision it would have taken based on the score provided by the private credit bureau.

We also obtained credit history data from the public credit registry managed by the SBS. All financial institutions subject to credit risk have to provide data to this public credit registry. Each month the SBS reports the maximum number of days in arrears (across all financial institutions), total debt, and classifies debtors in one of five status categories: normal, with potential payment problems, poor payment, doubtful payment, and loss. Only banked entrepreneurs appear in the public credit registry data. About 76 percent of the entrepreneurs in our sample were banked at the time they were screened by the EFL tool.

To assess loan repayment behavior, we defined three binary variables based on the SBS data, as listed in tables 2 and 3 below. We also used the number of days in arrears 6 and 12 months after applicants were screened by the EFL tool. All variables are described in more detail in appendix table A1. Appendix table A2 provides summary statistics.

IV. Empirical Results

We estimate linear regression models of the following form:

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i \in S.$$
Where $y_i$ is a measure of loan repayment behavior, $x_i$ is an indicator defined differently depending on the hypothesis we are testing (as described below), and $\varepsilon_i$ is the regression error term. $S$ is the sample of interest; it varies according to the hypothesis we are testing.\(^6\) The estimates reported in tables 2 and 3 correspond to $\alpha$ and $\beta$ for the specification above.\(^7\)

---

6. Appendix table A4 reports specifications that control for characteristics of the entrepreneurs and that use Probit, along with Horrace and Oaxaca (2006) tests. The results are robust to using these alternative specifications.

7. Since the estimates depend on the arbitrary threshold levels chosen by the implementing institution to accept/reject clients, we ran a sensitivity analysis moving the threshold levels between the 25th and 75th percentile of each score. Appendix 3 presents this analysis. The results are overall similar to the ones in the main text.
Testing Hypothesis 1: Risk Reduction

Table 2 lists our results from testing hypothesis 1: Entrepreneurs who were accepted under the traditional method but rejected based on their EFL score displayed poorer loan repayment behavior than entrepreneurs who were accepted under both methods. The sample in Table 2 includes only entrepreneurs who were accepted under the traditional model. Each pair of columns presents regressions of our outcome variables on a dummy variable equal to one if the entrepreneur was rejected based on an EFL score and accepted under the traditional model, and equal to zero if the entrepreneur was accepted under both methods. The first column presents the constant coefficient—the average for entrepreneurs accepted under both methods—while the second column presents the dummy variable coefficient (the difference between entrepreneurs rejected and accepted based on their EFL score).
The evidence in Table 2 suggests that the EFL tool has the ability to screen out higher-risk borrowers from the sample of banked entrepreneurs accepted under the traditional method (column 4). Banked entrepreneurs accepted under the traditional screening method but rejected based on their EFL score exhibit much poorer repayment behavior patterns across most of our indicators than entrepreneurs accepted under both methods. For example, banked entrepreneurs accepted under the traditional method but rejected based on their EFL score are 8.6 percentage points more likely to have been in arrears by more than 90 days during the 12 months after being screened by the EFL tool, compared to 14.5 percent of entrepreneurs accepted under both methods.

We did not observe that the EFL tool had the ability to screen out higher-risk borrowers for unbanked entrepreneurs approved under the traditional method (column 6). The differences in repayment behavior here are smaller and not statistically different from zero. Moreover, the signs of the estimates do not point consistently in the same direction.

Testing Hypothesis 2: Credit to New Borrowers

Table 3 presents the results of testing hypothesis 2: Entrepreneurs who were rejected under the traditional model but accepted based on their EFL score did not display poorer loan repayment behavior than entrepreneurs who were accepted under the traditional model. Each pair of columns presents regressions of the outcome variables on a dummy variable equal to one if the entrepreneur was rejected under the traditional model and accepted based on the EFL score, and equal to zero if the entrepreneur was accepted under the traditional model. The first column presents the constant coefficient (the average for entrepreneurs accepted under the traditional method), while the second column presents the dummy variable coefficient (the difference between entrepreneurs rejected under the traditional model and accepted based on their EFL score and entrepreneurs accepted under the traditional model).

Table 3 shows evidence against hypothesis 2 (column 2). In fact, entrepreneurs rejected under the traditional model and accepted based on their EFL score exhibited poorer loan repayment behavior than those accepted under the traditional method. These results seem to be driven by banked entrepreneurs and suggest that the traditional screening method—which, for banked entrepreneurs, incorporates valuable information about their past repayment behavior—is a powerful tool to screen out high-risk applicants (column 4).

The differences in the loan repayment behavior patterns of unbanked entrepreneurs are smaller and not statistically different from zero (column 6). Moreover, the size of the coefficients is small compared to the coefficients for banked entrepreneurs. Our results thus suggest that the EFL tool can be used to offer loans to unbanked applicants who are rejected under the traditional method without increasing the risk of the loan portfolio.
V. Conclusions

In this paper we detail the use of a psychometric credit application, developed by EFL, to reduce information asymmetries and to better assess credit risk for small businesses. In the context of a pilot exercise conducted by the fifth-largest bank in Peru, we found that the EFL tool can add value to a traditional credit-scoring method in different ways for banked and unbanked entrepreneurs.

For banked entrepreneurs—that is, those with a credit history—the EFL tool can be used as a secondary screening mechanism to reduce the portfolio risk. But for banked entrepreneurs with negative credit histories who have been rejected using the traditional credit-scoring method, the EFL tool has limited power and can even lead to an increase in the portfolio risk. That is, with respect to portfolio risk, the EFL tool does not successfully replace credit history information, but it does well at complementing this information. For unbanked entrepreneurs—that is, those with no formal credit history—our results suggest that the EFL tool can be used to make additional loans to applicants rejected based on the traditional screening method without increasing portfolio risk.

Our findings clearly show the importance of information in assessing credit risk and making accurate credit decisions. They highlight the power of traditional screening methods, based mainly on applicants’ credit history, to screen out loan applicants with poor loan repayment behavior. Increasing the quality of the information that credit bureaus can access—for example, including data from retailers and utility companies in addition to banks and financial institutions and allowing positive information (payment history on accounts in good standing) in addition to negative information (late payments, number and amount of defaults and arrears, and bankruptcies)—could improve credit-scoring models and increase credit markets’ confidence in their credit scores, even for entrepreneurs who have not previously borrowed from formal financial institutions. In the meantime, EFL offers a practical solution to financial institutions in countries where well-developed credit bureaus are in the process of consolidation.

References


Unitary or Noncooperative Intrahousehold Model? Evidence from Couples in Uganda

Nathan Fiala and Xi He

We present an overview of the evidence regarding the unitary, collective and non-cooperative models of household decision making and discuss how they can affect individual and household welfare. We then discuss the results of an artefactual experiment conducted in Uganda with spouses in order to test whether household members maximize common preferences, or instead are willing to pay a significant cost to hide money from their spouse. We find that both the unitary and non-cooperative models exist in the intra-household decision making process and that a “one-size fits all” model of household decision making is unlikely to be satisfactory. JEL codes: D13, J12, O12

The decision-making process within the family has great implications for both individual outcomes, like health and education, and public policies. However, some of the basic questions about the family decision-making process remain unclear. Theoretical models of household behavior offer different assumptions about what households maximize: the unitary approach (Samuelson 1956; Becker 1965) assumes that households maximize a common set of preferences where all income is pooled and the identity of the income recipient does not affect household decisions, while the collective approach (McElroy and Horney 1981; Manser and Brown 1980) assumes that the household members maximize a weighed sum of individual preferences where the identity of the individual controlling resources affects decisions, and bargaining power depends on control of allocation. Both of these approaches however can lead to a Pareto optimal outcome. Under the noncooperative household production model, the household bargaining process can lead to important inefficiencies (Carter and Katz 1997; Fafchamps 2001).

We first present an overview of the evidence regarding the unitary, collective, and noncooperative models and discuss how they can affect individual and
household welfare. The scope of the evidence of when the unitary model does not hold presents strong evidence that intrahousehold bargaining is an important issue in many contexts.

We then present the results of an artefactual experiment conducted in Uganda with spouses in order to test whether household members maximize common preferences or instead are willing to pay a significant cost to hide money from their spouse. In sum, 731 couples are offered the tradeoff between maximizing household income or paying a penalty to gain greater control over that income. We find that both the unitary and noncooperative models exist in the intrahousehold decision-making process and that a “one-size fits all” model of household decision making is unlikely to be satisfactory.

Our work is very close to that of Zou (2015), who conducts an artefactual experiment in Burkina Faso in which married individuals reveal their relative valuation of spouse’s earnings over their own. Our contribution to the literature lies in that we use a specific experimental design that can directly investigate whether couples share the same preference or not, and we provide direct evidence that both unitary and noncooperative models exist, which gives credit to the popularity of the intrahousehold bargaining model.

The remainder of this article is structured as follows. In section 1 we discuss existing evidence of the unitary model of the household. We present the experimental design in section 2 and the results in section 3. Section 4 concludes.

I. Evidence of the Unitary Model

There is growing evidence that the unitary household model fails to describe intrahousehold decision making. Recent research lends strong support to an intrahousehold model that considers the existence of other individuals in the households that, through a bargaining process, play a complex role in household decision making. This document presents a short outline of the literature that explores when the unitary model is not sufficient to describe household behavior.

There is a broad literature on how household bargaining impacts household decision-making in different fields. One implication of household bargaining is on household expenditure. Hashemi et al. (1996) finds that being a member of Grameen Bank or BRAC (Bangladesh Rural Advancement Committee) increases the involvement in household decisions and in making purchases. Holvoet (2005) presents evidence that group-based lending for women in South India has a positive effect on household decision-making regarding loan use and money management. In Ghana, Doss (1996) shows that using the assets held by women in the households as a measure of bargaining power has implications for household expenditure decisions.

Some studies show the impact of household bargaining on consumption of specific goods. De Brauw et al. (2013) study conditional cash transfers (CCTs) given to women in Brazil and find impacts on the purchase of household durable goods. In the same line and also for Brazil, Polato e Fava and Arends-Kuenning
show how bargaining power in the household influences the decision of buying household production durables goods or entertainment durable goods. For Côte d’Ivoire, Duflo and Udry (2004) show that rain shocks can affect household bargaining power, leading to an impact on expenditure. Based on the idea that in many African countries there is some gender specialization in farming crops, it finds that shocks that increase the output of “female crops” shift the expenditures to all types of food (except staples), but if the shock affects the output of “male crops,” it has no effect on the purchase of food. Rubalcava et al. (2009) study the CCT’s PROGRESA program in Mexico and find that poor rural women that have resources under their control are more likely to spend on investment on children and in small-scale livestock. Based on the same data, Bobonis (2009) includes in the analysis a variation attributable to localized rainfall shocks. He also finds that income changes for women are positively related with expenditure on children’s goods. However, the rainfall shock does not have a significant effect on household expenditure. For United Kingdom, Lundberg et al. (1997) also show an impact on the expenditure on children, particularly on children’s clothing and on women’s, when a child allowance is transferred to wives.

Peters (2011) studies a family planning program in Bangladesh and finds that women treated by the program are less likely to be able to make certain purchases without their husband’s permission. In this case, the program is one that endows men with a new asset because the program is provided in certain random areas, and in Bangladesh married women moved in with their husbands. Other studies that find that the introduction of a family planning and health program decreases the bargaining power of treated women include Orrefice (2007) and Chiappori and Orrefice (2008).

The literature also stresses the impact of household decision-making on labor supply and time allocation. Orrefice (2007) studies the impact of legalization of abortion in some states of the United States and finds that it affects household bargaining, and its effect is a decrease in labor supply of married women and an increase in their husband’s labor supply. Heath and Tan (2014) find the opposite effect in India when bargaining power is given by an inheritance law that improves women’s situation to inherit property. Gray (1998) shows for some states of United States that marital-property laws determine how unilateral-divorce laws affect household bargaining. He finds that wives increase their labor supply and this change reflects changes in home-production hours and in leisure time. However, based on the same data, Stevenson (2008) shows that any unilateral divorce law increases wives labor supply. The same effect is found in Iversen and Rosenbluth (2006) in their study for several countries (most established democracies, some east European countries, and the Philippines). In the study of Holvoet (2005), where credits were given to women, is shown that group-based lending for women also affects time allocation to women. For the United States, Friedberg and Webb (2005) study the effect of spouse’s relative wages as a measure of household bargaining on time used during the weekend. They find that when the wives’ relative wage increases, they enjoy more leisure time and
dedicate less time to household work. In the same line, Bittman et al. (2003) find that in the United States and Australia women decrease the time in household work if their earnings increase.

There is also evidence that household bargaining affects investment and agriculture supply. Yilmazer and Lich (2015), based on data for the United States, find that different risk preferences of wife and husband affect the household portfolio asset allocation, and consequently, the level of risk of investment depends on the risk tolerance of the spouse with more power bargaining. With respect to the impact on agriculture supply, Lim et al. (2007) present evidence for Ethiopia, measuring bargaining power as the value of livestock each spouse would have in the event of a divorce. The study finds that women’s bargaining power influences negatively in cash crop production if the earnings of the cash crop are controlled by the husband and it is based on the wife’s work.

Fertility, education, health, and nutrition of children have also been found to be affected by household bargaining. For fertility, the evidence can be found in Thomas (1990), which finds an impact on children’s survival in Brazil; Rasul (2008) shows that for Malaysia and China the impact depends on the commitment to future actions in the marriage; and Ashraf et al. (2014) for Zambia finds that it depends on moral hazard, specifically in asymmetric information in the use of contraception. Regarding education, health, and nutrition of children, de Brauw et al. (2013) find that CCTs given to women in the “Bolsa Familia” program in Brazil increase the decision-making power of women in contraception use, as well as children’s school attendance and health. Additionally, for Brazil, Thomas (1990) finds that unearned income in hands of women has more effect on household health than income in the hands of the husbands, and Thomas (1993) shows that income in general in the hands of women is positive related to expenditure on household education and health. Park (2007) studies the resource allocation of children’s nutrition and education in Indonesia; he shows that for children’s nutrition the unitary model is rejected while for education the results are mixed. Maluccio and Quisumbing (2003) show a positive correlation between indicators of female bargaining power and expenses on food and education for Bangladesh, Ethiopia, Indonesia, and South Africa.

The impact of household bargaining is also observed in domestic violence. In studies where microcredits given to women were analyzed, the impact was mixed. There are some studies that find a decrease in domestic violence (Hashemi et al. 1996; Kabeer 2001), while others find that microcredits or any means that make women richer increase it (Goetz and Sen Gupta 1996; Rahman 1999; Bloch and Rao 2002; Balasubramanian 2013). Literature also finds an effect on policy preferences, for example, the work of Iversen and Rosenbluth (2006) finds that increase in women’s labor participation has an effect on women’s policy preferences. This is driven by the idea that women will try to improve their exit options from the marriage and their household bargaining, and they will choose the policies that fit best with their preferences.
Some more general studies include Kebedea et al. (2014), who runs an experiment with couples in Ethiopia where information about endowments is varied. He finds that information improves efficiency in only some treatments, meaning information is context dependent. Husbands’ expectations of their wives contributions are higher than their wives’ actual contributions, and wives’ expectations of their husbands’ are lower than their husbands’ actual contributions. Carlsson et al. (2009) find differences in risk preferences when spouses make decisions separately and together. The couples’ risk preferences become more similar the richer the spouses and the higher the relative income contribution of wives. And Ashraf (2009) looks at the effect of observability and communication on financial choices of married individuals in the Philippines. When choices are private, men put money into their personal account. When choices are observable, men commit money to consumption for their own benefit, but when spouses are able to communicate, men put money into their wife’s account.

II. EXPERIMENTAL DESIGN AND DATA

The experiment we describe here is designed to understand how control over income becomes an issue when preferences over money usage differs between spouses. We study the role of this factor using a design where participants are required to allocate money across two investment options. This question is inspired by Mani’s (2011) work in India that found participants were willing to accept a low return option over a high return option in order to have greater personal control of money over their spouse.

The experiment is conducted in Northeast and Central Uganda. After completing a simple survey of respondent’s demographic and income we asked individuals if we can meet with their spouse the next day to ask the same questions. If they said yes, the enumerator read the following:

Thank you very much for letting us meet with you and your partner. Before I go, I have one final question. This question involves real money. I have 2,000 USH to give you for your time. You can choose to invest the money in two businesses. Let’s call them business A and business B. If you invest in business A, the money will be doubled to 4,000 USH. I can pay this money to you now in cash. If you invest in business B, the money will be tripled to 6,000 USH. Another member of our research team will then give this money to your partner tomorrow at their business. Which investment would you prefer to make?

We believe this game is credible for three reasons. First, this data collection was part of a larger panel data collection where individuals had been visited four times previously. Second, the amount offered in the game is relatively large at about half a day’s income for the average person. Finally, we test for whether income of the individual is correlated with the decision in the game and find no relationship.

In this game, the investor faces a tradeoff between generating larger household income and greater control over that income. We believe the interpretation of the game is quite straightforward: if both husband and wife prefer higher household
income, the unitary model is validated, and if either of the spouses prefers
greater control rather than greater income, then the noncooperative household
model is supported. We ask main respondents first, track their spouse’s basic in-
f ormation and answer to the above game later, and then match the spouse’s
choices to investigate whether they share the same set of preference or not.

In sum, 731 married couples participated in the game; 60 percent of the
primary participants are female. For the female subgroup, 8.8 percent of the par-
ticipants are 18–23 years old, while people ranging from 24 to 35 comprise 66.8
percent. In addition, 14.9 percent are between 36 and 41 years of age, and the re-
maining 9.6 percent are from 41 to 50 years of age. For the male main respon-
dents, 12 percent are aged from 18 to 23, 65.2 percent are aged from 24 to 35,
11.1 percent are aged from 36 to 41, and the remaining 11.7 percent aged from
41 to 50. About 82 percent of participants have completed primary school edu-
cation, and 6.17 percent of participants have completed secondary school or
have higher education level. There are 2.82 children in each household on
average.

III. EXPERIMENTAL RESULTS

The results of the games are presented in table 1. There are several features worth
noticing. First of all, about 37.7 percent couples choose to invest in the business
that generates higher profits instead of choosing lower-return business and
keeping the money to themselves. This presents some validation of the unitary
household model. However, 17.9 percent of couples jointly prefer greater
control over income to greater income, indicating that both husband and wife
hope to have greater control and more bargaining power in household decision-
making.

The remaining 44.4 percent of couples have different preferences: one pursues
higher income while the other one pursues higher control over that income, in-
validating the assumption of unitary model. The contradicting preferences over
income and control among couples suggest that couples compete for bargaining

<table>
<thead>
<tr>
<th>Table 1. Game Summaries</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

|                         |
|                         |
|                         |
|                         |

|                         |
|                         |
|                         |
|                         |

|                         |
|                         |
|                         |
|                         |

|                         |
|                         |
|                         |
|                         |

Notes: Trust spouse takes the value of 1 if they sent the money to their spouse. It takes value 0 if
they decided to keep the money themselves at a cost.
power and the decision-making process within the household might be very complicated, which is also the reason why the household bargaining affects both the individual outcomes, like education, health, labor supply, as well as public policies like income transfer, and so forth, as evidenced by the various empirical studies.

Another feature worth noticing is that women and men have significant preferences over higher income and greater control. In sum, 41.6 percent of male respondents prefer higher income to greater control, while only 24.1 percent female respondents choose higher income rather than greater control. Moreover, 52 percent of the female respondents are trusted by their spouses, while 41.1 percent of the male respondents are trusted by their spouses, suggesting that male participants are more likely to choose higher income yet female respondents are more likely to choose higher control over income. This conclusion is similar to some other studies claiming that women are more cautious with money and they prefer to keep money by themselves because they are afraid that their husbands might use money unwisely.

IV. Conclusion

Theoretical models have different assumptions about the intrahousehold decision process. There are many alternative empirical studies rejecting the unitary model and supporting the noncooperative approach, which assumes that spouses have different preferences and depend on bargaining power to allocate household resources. We conduct a game among 731 couples in Uganda to investigate whether couples have different preference over higher income and greater control over that income. Experiment results suggest that there is no “one-size fits all” model. There are households in which couples both prefer higher income and the unitary model fits well, while there are also couples that forgo overall efficiency and pursue higher individual power over money, which lends credit to the cooperative household model. Moreover, women tend to prefer higher control over income than men, and this indicates that women are more likely to feel unsafe to let their husbands dispose of money at will.

There is growing strong evidence in the literature that shows whether the unitary or noncooperative model of household decision-making holds has significant implications for household welfare. However, what determines the type of household decision-making process is still unclear, and future studies can move forward in this direction to explore the factors impacting different intrahousehold decision-making models.

REFERENCES


Opportunity versus Necessity: Understanding the Heterogeneity of Female Micro-Entrepreneurs

Gabriela Calderon, Leonardo Iacovone, and Laura Juarez

Entrepreneurs that voluntarily choose to start a business because they are able to identify a good business opportunity and act on it—opportunity entrepreneurs—might be different along various dimensions from those who are forced to become entrepreneurs because of lack of other alternatives—necessity entrepreneurs. To provide evidence on these differences, this article exploits a unique data set covering a wide array of characteristics, including cognitive skills, noncognitive skills, and managerial practices, for a large sample of female entrepreneurs in Mexico. Descriptive results show that on average opportunity entrepreneurs have better performance and higher skills than necessity entrepreneurs. A discriminant analysis reveals that discrimination is difficult to achieve based on these observables, which suggests the existence of unobservables driving both the decision to become an opportunity entrepreneur and performance. Thus, an instrumental variables estimation is conducted, using state economic growth in the year the business was set up as an instrument for opportunity, to confirm that opportunity entrepreneurs have higher performance and better management practices. JEL codes: J24, L26, O12

1 Introduction

Support programs for micro-businesses have become increasingly common in developing countries in recent years for at least two reasons: first, micro-enterprises employ a substantial fraction of individuals in these economies (about 47 percent in Mexico); and second, despite their prevalence, the majority of these micro-enterprises tend to stay small and have low productivity. However, the impact of such programs—many of which provide business grants, training or a combination of both—has been mixed at best. This raises the question about whether

Gabriela Calderon is a nonresident research fellow at Poverty and Governance, Stanford University; her e-mail address is gabcal@alumni.stanford.edu; Leonardo Iacovone is a senior economist at The World Bank, 1818 H Street NW, Washington, DC; his e-mail address is liacovone@worldbank.org; Laura Juarez (corresponding author) is a researcher at Dirección General de Investigación Económica, Banco de México, Av. 5 de mayo 18, Mexico, D.F.; her e-mail address is ljuarezm@banxico.org.mx. We gratefully acknowledge funding from Instituto Nacional del Emprendedor (INADEM) at the Ministry of Economy in Mexico, USAID, The World Bank, and the Instituto Nacional de las Mujeres (INMUJERES). We thank Paulina López and Sandra Aguilar for outstanding research assistance. The views expressed in this article are solely those of the authors and do not necessarily reflect those of Banco de Mexico or the World Bank.

© The Author 2016. Published by Oxford University Press on behalf of the International Bank for Reconstruction and Development / THE WORLD BANK. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
these impacts depend on characteristics or attitudes of the entrepreneur, and therefore better targeting could significantly improve them. In fact, some recent evidence suggests that, even though the mean effects of business training might be small and not significant, greater returns are concentrated in high-potential entrepreneurs, who are the most likely to adopt better entrepreneurial practices and earn higher profits after training (De Mel et al. 2012; Calderon et al. 2013; Fafchamps et al. 2014).

In this article, we provide some novel evidence on the heterogeneity of female microentrepreneurs in urban Mexico by comparing those who started their business out of opportunity with those who did so out of necessity. Our data come from the baseline survey for the randomized evaluation of a large business training program for female entrepreneurs, funded by the National Institute of Entrepreneur and implemented by the NGO CREA, Mujeres Moviendo México. This data set is unique in that it provides detailed information on business outcomes, access to credit, cognitive and noncognitive skills for a large sample of female micro-entrepreneurs in selected urban areas in Mexico.

For the empirical analysis, we first show mean differences in business performance, characteristics and skills between the two groups of microentrepreneurs. Then we use discriminant analysis to find the combination of observable characteristics that best distinguishes opportunity from necessity entrepreneurs. Finally, we estimate the effect of being an opportunity entrepreneur on the profitability of the business and the quality of management practices, using state GDP growth at the time when the business was opened as an instrument for opportunity.

Our results suggest that on average opportunity entrepreneurs have significantly higher profits, better management practices, and higher cognitive and selected noncognitive skills. Our discriminant analysis shows that management practices, some business characteristics and skills of the entrepreneur can be used to separate both groups, but the distinction based on these observables is only partially successful, which suggests the existence of unobservables determining both the decision to set up a business, based on a perceived opportunity, and firm performance. Finally, our instrumental variables results confirm that opportunity entrepreneurs have higher performance measured as profits and management practices, even after controlling for the observable skills of the entrepreneur.

The policy relevance of these findings is twofold. First, as mentioned before, identifying the observable characteristics that define opportunity entrepreneurs could potentially improve the targeting of business support programs, by concentrating them on the entrepreneurs with the highest growth potential. This by no means implies disregarding low-performing necessity entrepreneurs. On the contrary, the latter might be better served by interventions that help ease their

1. Discriminant analysis is a technique of species classification that has been used before to study microentrepreneurship by de Mel, McKenzie and Woodruff (2010) for Sri Lanka; and by Bruhn (2013) for Mexico.
transition to salaried employment. Second, our focus on female entrepreneurs is also of great relevance given that their businesses seem to be at a particular disadvantage when trying to scale-up. For instance, in Latin America, microfirms led by women have been found to have an even smaller size and lower productivity, compared to those led by men (Bruhn 2009). Thus, the evidence provided in this article can be readily applied to those in more need of targeted support.

2 Data and Descriptive Analysis

Our data come from a baseline survey conducted in 2014 in eight urban areas in Mexico: Aguascalientes, Dolores Hidalgo, Irapuato, León, Mexico City, Querétaro, San Juan del Río, and Toluca. For the sampling design, a number of census tracks (called AGEBs by the Mexican Statistical Institute, INEGI) with relatively high concentration of commercial activity were chosen.2 The questionnaire was applied only to women who are business owners or partners, that is, who take managerial, financial, and marketing decisions, and the sample included firms with no more than five employees or sales that are less than four million Mexican pesos per year (which is the definition of a micro-enterprise).3

Our survey has information on a uniquely broad range of variables: sociodemographic characteristics, current and initial business characteristics, business outcomes (sales, profits, and costs), managerial practices, cognitive and noncognitive skills of the entrepreneur, access to credit, growth expectations and obstacles, among others. The sample consists of 10,275 female micro-entrepreneurs. To our knowledge, our survey is the first one measuring both cognitive and noncognitive skills, in addition to business characteristics and outcomes, for such a large sample of micro-entrepreneurs in Mexico.

We classify the women in our sample as “opportunity” or “necessity” entrepreneurs according to the self-reported reason for opening their business. A respondent is classified as opportunity entrepreneur if she reports opening her business either because (i) she wanted to become independent, (ii) she had money and found a good business opportunity, or (iii) she wanted to practice her profession or develop her career profile. Conversely, necessity entrepreneurs are those who started their business because they could not find a well-paid or suitable job and needed a source of income. After excluding women who started their business out of family tradition or other reasons, we are left with a sample of 8,949 entrepreneurs (21 percent classified as opportunity and 79 percent as necessity).

In table 1, we report the mean differences between opportunity and necessity entrepreneurs in the following characteristics: (i) Business performance measures (weekly profits, sales, and sales per worker); (ii) business practices, as measured

2. These areas with higher concentration of commercial activity were identified relying on the 2009 Economic Census data.
3. Fixed and semi-fixed stands on the street, and dwellings in which the woman was found selling or producing something (or had a sign indicating so) also qualified as enterprises and were included in the survey.
TABLE 1. Mean Difference between Opportunity and Necessity Entrepreneurs (Full Sample)

<table>
<thead>
<tr>
<th>Measures of performance</th>
<th>Opportunity group mean</th>
<th>Necessity group mean</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekly profits (self reported)</td>
<td>1937.635</td>
<td>1349.177</td>
<td>588.4576***</td>
</tr>
<tr>
<td>Weekly sales (self reported)</td>
<td>4507.258</td>
<td>3540.586</td>
<td>966.6726***</td>
</tr>
<tr>
<td>Weekly sales/workers</td>
<td>4395.333</td>
<td>4302.254</td>
<td>93.09963</td>
</tr>
<tr>
<td>Composite Business Practice score (standarized)</td>
<td>.3060325</td>
<td>-.0842768</td>
<td>.3903093***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Business characteristics</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of entrepreneur</td>
<td>41.81658</td>
<td>45.92807</td>
<td>-4.111494***</td>
</tr>
<tr>
<td>Age of business in months</td>
<td>90.59086</td>
<td>103.6411</td>
<td>-13.05026***</td>
</tr>
<tr>
<td>Proportion with one worker or more</td>
<td>.3185</td>
<td>.2278</td>
<td>.0907***</td>
</tr>
<tr>
<td>Costs: monthly salary expenses</td>
<td>1025.357</td>
<td>498.751</td>
<td>526.6064***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cognitive skills</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score of raven test (standarized)</td>
<td>.0827833</td>
<td>-.0211648</td>
<td>.1039482***</td>
</tr>
<tr>
<td>Total score of digit span test (standarized)</td>
<td>.1459293</td>
<td>-.0328984</td>
<td>.1788277***</td>
</tr>
<tr>
<td>Years of schooling</td>
<td>9.789416</td>
<td>8.251668</td>
<td>1.537748***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Noncognitive skills</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion (standarized)</td>
<td>-.0365413</td>
<td>.055396</td>
<td>-.0520809**</td>
</tr>
<tr>
<td>Agreeableness (standarized)</td>
<td>-.0093652</td>
<td>.0075027</td>
<td>-.0168679</td>
</tr>
<tr>
<td>Conscientiousness (standarized)</td>
<td>-.069074</td>
<td>.0121551</td>
<td>-.0812292***</td>
</tr>
<tr>
<td>Neuroticism (standarized)</td>
<td>-.0246642</td>
<td>.0130341</td>
<td>-.0376982</td>
</tr>
<tr>
<td>Intellect/imagination (standarized)</td>
<td>.0074094</td>
<td>-.0046432</td>
<td>.0120526</td>
</tr>
<tr>
<td>Self efficacy (standarized)</td>
<td>-.0957561</td>
<td>.0110521</td>
<td>-.1068082***</td>
</tr>
<tr>
<td>Locus of control (standarized)</td>
<td>.0954217</td>
<td>-.0202922</td>
<td>.1157139***</td>
</tr>
<tr>
<td>Impulsiveness (standarized)</td>
<td>.0480202</td>
<td>-.0170598</td>
<td>.06508**</td>
</tr>
<tr>
<td>Self confidence (standarized)</td>
<td>.1074716</td>
<td>-.0390423</td>
<td>.1465139***</td>
</tr>
<tr>
<td>Attitude toward risk (standarized)</td>
<td>.0975181</td>
<td>-.0298163</td>
<td>.1273345***</td>
</tr>
<tr>
<td>Self satisfaction (standarized)</td>
<td>.1163573</td>
<td>-.0443902</td>
<td>.1607475***</td>
</tr>
<tr>
<td>Optimism (standarized)</td>
<td>.138387</td>
<td>-.0400507</td>
<td>.1784377***</td>
</tr>
<tr>
<td>Attitude toward trust (standarized)</td>
<td>.0137271</td>
<td>-.0011845</td>
<td>.0149116</td>
</tr>
<tr>
<td>Attitude toward growth (standarized)</td>
<td>.0621355</td>
<td>-.0119533</td>
<td>.0740888***</td>
</tr>
</tbody>
</table>

Observations 8949

*p < 0.1, **p < 0.05, ***p < 0.01.

by a standardized Composite Business Practices score (CBP)⁴; (iii) characteristics of the business and the entrepreneur (age, proportion with at least one worker, and monthly salary expenses); (iv) cognitive skills (education, standardized scores of Raven and digit span recall tests); and (v) noncognitive skills (standardized measures of agreeableness, conscientiousness, extraversion, intellect/imagination and neuroticism⁵, impulsiveness, locus of control, willingness to take risks, self-confidence, self-efficacy, optimism, self-satisfaction, and trust).

⁴. The CBP score is an index that measures how well entrepreneurs in our sample manage their business and is constructed considering measures of marketing, keeping stock, record keeping and financial planning, following Fafchamps and Woodruff (2014).

⁵. These personality traits are referred to as “the big five” in the literature.
Regarding business performance, table 1 shows that mean weekly profits and sales are higher for opportunity than for necessity entrepreneurs and the difference is statistically significant. Mean weekly sales per worker are also higher for opportunity entrepreneurs, but the difference is not statistically significant, probably because, as shown in the next panel, those entrepreneurs have a significantly larger number of workers. Opportunity entrepreneurs also have a significantly higher composite business practice score, compared to necessity ones. This descriptive evidence confirms that opportunity entrepreneurs have both better business performance and are manage their businesses significantly better. Opportunity entrepreneurs are four years younger than necessity ones on average, and their businesses are about 13 months younger. About 32 percent of opportunity entrepreneurs have at least one worker or more, whereas only 23 percent of necessity ones do. As a result, opportunity entrepreneurs have about double the salary expenses per month compared to necessity ones. All these mean differences are statistically significant.

The next panel in table 1 shows that, compared to necessity entrepreneurs, opportunity ones also have significantly higher mean cognitive skills. They have significantly higher mean scores in the Raven and digit span recall tests, and 1.5 more years of schooling. The mean difference in schooling between groups corresponds roughly to that between having finished secondary school and proceeded to the first year of high school and not having finished secondary school.

The last panel in table 1 shows that opportunity entrepreneurs have statistically significant higher mean scores for locus of control, impulsiveness, self-confidence, self-satisfaction, willingness to take risks, optimism and their attitude toward business growth, compared to their necessity counterparts, and the opposite is observed for extraversion, conscientiousness, and self-efficacy. However, it is reasonable to think that at least three of those traits in which opportunity entrepreneurs score higher on average—namely, willingness to take risks, optimism, and their attitude toward business growth—could be correlated with better business outcomes.

### 3 Discriminant Analysis

To complement the descriptive analysis presented so far, we use discriminant analysis to explore whether management practices, business and entrepreneur characteristics, and skills can be used to distinguish opportunity from necessity entrepreneurs; and then use the estimates to predict whether a given observation belongs to each group.\(^6\)

Table 2 presents the results for our full sample of both necessity and opportunity entrepreneurs, and then for the opportunity group compared to the top and bottom profit quartiles of the necessity group. We vary the set of characteristics that are used to separate entrepreneurs into different species. For instance, in

---

\(^6\) For a description of this methodology, please refer to de Mel, McKenzie, and Woodruff (2010).
panel A, for the full sample of entrepreneurs, using only measures of business performance would lead us to correctly classify 52 percent of entrepreneurs as opportunity and 63 percent as necessity. Using measures of business characteristics and skills seems to improve the classification of opportunity entrepreneurs, as measured with a larger proportion of correctly classified ones but not necessarily so for necessity entrepreneurs. In the last row of panel A, using all variables combined results in 60 and 63 percent of opportunity and necessity entrepreneurs to be correctly classified, respectively. This is about a 10 percentage point improvement from classifying entrepreneurs as opportunity or necessity over a random classification, that is, one that would be obtained by simply flipping a coin, indicating that there might be other unobserved factors that explain that a women is a necessity or an opportunity entrepreneur.

The last two columns in panel A show that, when including all the explanatory variables together, our model would classify 42 percent of the entrepreneurs as opportunity and 58 percent as necessity. This suggests that some entrepreneurs who report themselves as necessity ones, are in fact more similar to their opportunity counterparts, according to the discriminant analysis.

Panels B and C show the results when including all the opportunity entrepreneurs in our sample and only necessity entrepreneurs in the bottom and top quartiles of self-reported daily profits. The objective is to see whether high and low performing necessity entrepreneurs are in fact classified as more or less similar to opportunity ones. For instance, panel B shows that, when all variables are
combined in the discriminant analysis, about 74 percent of low-performing necessity entrepreneurs are correctly classified as being so. In comparison, in Panel C, about 49 percent of high-performing entrepreneurs are correctly classified as being necessity ones. This suggests that some high-performing necessity entrepreneurs are more similar in the characteristics we use for the discriminant analysis to opportunity ones, as would be expected. The last column shows that, although in panel B about 57 percent of entrepreneurs overall are classified as necessity, which is not very different from the results for the overall sample in panel A, in panel C only 44 percent of high-performing entrepreneurs are classified as necessity ones, confirming that a higher proportion of high-performing necessity entrepreneurs in fact “look like” their opportunity counterparts.

In conclusion, the discriminant analysis suggests that necessity and opportunity entrepreneurs differ in key characteristics and abilities, but some among those starting their business out of necessity actually resemble their more able opportunity counterparts. Therefore classifying entrepreneurs based on observable characteristics seems not to be very accurate. This suggests the existence of unobservable traits that drive both the decision to set up a business to pursue a good opportunity and the entrepreneur’s performance. Accordingly, in order to confirm our descriptive results suggesting that opportunity entrepreneurs have higher performance, we need to adopt an IV approach, which we discuss in the following section.

4 Instrumental Variables Approach

Opportunity entrepreneurs appear to perform better than necessity ones. In order to estimate the partial effect of opening a business out of opportunity, controlling for the relevant observable characteristics of the firm and the entrepreneur, we estimate the following regression:

$$y_i = \alpha + \beta \text{opportunity}_i + \Theta^T X_i + \psi_i,$$

where $y_i$ is alternatively the log of self-reported weekly profits or CBP score observed in 2014. Our key independent variable is $\text{opportunity}$, a dummy indicating that the entrepreneur reported opening her business out of opportunity, defined as before, and $X_i$ is a row vector that includes age of entrepreneur, Raven test score, span test score, years of schooling, locus of control, risk attitude, self satisfaction and optimism which were the observable characteristics of the entrepreneur and the firm that were statistical significant among all of the variables used in our descriptive analysis. Robust standard errors are estimated for this equation.

Even after controlling for all the measures of skills that are available to us, $\text{opportunity}$ might be endogenous in equation (1), because it might be correlated with unobservable characteristics of the entrepreneur, like her social networks.
that contribute to have higher profits and are captured in $\psi_i$. To overcome this endogeneity, we use a two-stage least square (2SLS) estimation in which the first-stage equation is:

$$opportunity_i = \alpha + \beta GDPgrowth_{t0s} + \Gamma^T X_i + \epsilon_i,$$

(2)

where $GDPgrowth_{t0s}$, the instrument, is the GDP growth observed in state $s$ where the entrepreneur $i$ lives, in the year $t_0$ she decides to open her business. Given that our survey measures profits and management performance in 2014, our assumption is that state GDP growth at the time when the business was set up is exogenous to profits (and management performance) various years later and only influences them through the choice of starting a business out of necessity or opportunity. When we use a 2SLS specification we include years of opening fixed effects in both stages, in this way, we compare firms that opened in the same year and have similar characteristics.

Table 3 presents the estimation results. Panels A and B show the effect of the opportunity dummy on weekly profits and the CBP score, respectively. Panel C shows the first stage of the 2SLS presented in column 2. Column 1 shows that when estimating equation (1) by OLS, the effect of opportunity on weekly profits is positive and significant at 1 percent in panel A and so is the effect of this same variable on the CBP score in panel B. In column 2, controlling for the potential endogeneity of this variable yields a much larger positive coefficient for opportunity for both profits and the CBP score, significant at 10 percent for both. The estimate in panel A implies that women who open their business out of opportunity in year $t_0$, induced by higher economic growth in their state in that same year, have 2.6 times higher weekly profits than necessity ones; and that part of this effect can be explained by a 2 times better CBP score, that is, by better management practices.

The partial F statistic shown in panel C has a value less than 10 but larger than 5, which means that the maximum IV bias we are tolerating is less than 20 percent (Staiger and Stock 1997). The fact that the 2SLS estimates in column 2 are larger than the OLS ones in both panels A and B, might indicate that opportunity is measured with error. An alternative explanation is that our instrument is identifying the effect of opportunity for a specific subset of “compliers,” in our case, women responding to become opportunity entrepreneurs when there is a higher state GDP growth in the year they opened their business or those who become necessity entrepreneurs when there is a low state GDP growth, that is, a local average treatment effect (LATE), which might not be representative of the

---

7. Poschke (2013) presents a model in which, in each period, individuals decide whether to work or to run a firm. Selection of agents into entrepreneurship comes from the heterogeneity of the relevant outside option and the heterogeneity in firm’s performance.

8. Poschke (2013) argues that agents that have a low productivity in the labor market have lower performance as entrepreneurs. Therefore, it is possible that this type of people is commonly laid off when there is low GDP growth and always have low entrepreneurial performance.
average population of female entrepreneurs. The third column in table 3 excludes the entrepreneurs who opened their businesses after 2011 from the estimation sample to address the concern that our instrument might directly affect profits observed in 2014, the year our survey was conducted, through a correlation with state GDP growth in that year. The estimate for opportunity in panel A and the first-stage coefficient of our instrument in panel C are very similar in magnitude to our estimates in column 2, which is reassuring, even though both of them lose statistical significance, probably due to the smaller number of observations. The effect of opportunity on the CBP score in column 3 is positive, larger than in column 2, and significant at 5 percent.

Table 4 presents some robustness checks. Specifically, we use state GDP growth observed either the year before \(t_0 - 1\), in column 1) or the year after \(t_0 + 1\), in column 2) as instruments, instead of that observed during the same year \(t_0\). In panel C, the estimate for lagged state GDP growth is close to zero and insignificant, whereas the one for \(t_0 + 1\) is positive, of similar magnitude compared to the one for \(t_0\) in table 3, and significant at 1 percent. In column 2, panels A and B show that the effects on profits and CBP score are positive, but only the latter is significant at 10 percent. Finally, column 3 includes the state GDP growth in \(t_0 - 1\), \(t_0\) and \(t_0 + 1\) together as instruments. In panels A and B,
the effects of opportunity on profits and the CBP score are positive and significant at 10 and 5 percent, respectively, confirming that becoming an opportunity entrepreneur, induced by our instruments, increases business profitability in part due to better management. In panel C, only the state GDP growth in $t_0$ and $t_0 + 1$ have a positive and significant effect on the endogenous variable, but the first-stage partial F statistic for the three instruments is low (3.8).

5 Conclusions

We provide new evidence suggesting that opportunity entrepreneurs have more profitable businesses, manage them better and have higher measures of cognitive and of some noncognitive skills, compared to necessity ones. Using a discriminant analysis, we show that combining these characteristics allows us to separate both types of entrepreneurs, although not perfectly, and identify those that even though they started their business out of necessity are more similar in observables to their opportunity counterparts. Given that the discriminant analysis only increases 10 percentage points of the prediction between necessity and opportunity entrepreneur compared to a random assignment, the characteristics of the firm and of the entrepreneur considered in the analysis cannot fully explain when an entrepreneur is either an opportunity or a necessity one. In this way, we proceed to analyze this characteristic with an instrumental variable approach, which confirms that businesses led by opportunity entrepreneurs are significantly more
profitable than those led by necessity entrepreneurs, in part because of better management practices. These results are relevant for improving the targeting of business support programs. For instance, these programs could obtain larger impacts by concentrating on the entrepreneurs with the highest growth potential, whereas other types of interventions might help low-performing necessity entrepreneurs improve their prospects for salaried employment.

REFERENCES


Highways and Spatial Location within Cities:
Evidence from India

Ejaz Ghani, Arti Grover Goswami, and William R. Kerr

We investigate the impact of the Golden Quadrilateral (GQ) highway project on the spatial organization and efficiency of manufacturing activity. The GQ project upgraded the quality and width of 3,633 miles of roads in India. We use a difference-in-difference estimation strategy to compare non-nodal districts based upon their distance from the highway system. For the organized portion of the manufacturing sector, we find that GQ led to improvements in both urban and rural areas of nonnodal districts located 0–10 km from GQ. These higher entry rates and increases in plant productivity are not present in districts 10–50 km away. The entry effects are stronger in rural areas of districts, but the differences between urban and rural areas are modest relative to the overall effect. For the unorganized sector, we do not find material effects from the GQ upgrades in either setting. These findings suggest that in the time frames that we can consider—the first five to seven years during and after upgrades—the economic effects of major highway projects contribute modestly to the migration of the organized sector out of Indian cities but are unrelated to the increased urbanization of the unorganized sector. JEL codes: L10, L25, L26, L60, L80, L90, L91, L92, M13, O10, R00, R10, R11, R14

Adequate transportation infrastructure is an essential ingredient for economic development and growth. Rapidly expanding countries like India and China face severe constraints on their transportation infrastructure. Business leaders, policy makers, and academics describe infrastructure as a critical hurdle for sustained growth that must be met with public funding, but to date there is a limited understanding of the economic impact of those projects. We study how proximity to a
major new road network affects the organization of manufacturing activity. Additionally, we also compare how GQ influenced the manufacturing operations and entrepreneurship rates in the urban and rural portions of the districts through which it passed.

Our setting is the Golden Quadrilateral (GQ) Project, a large-scale highway construction and improvement project in India, using plant-level data from 1994 to 2009. The GQ project sought to improve the connection of four major cities of India—Delhi, Mumbai, Chennai, and Kolkata. Comprising 3,633 miles of road upgrades and new construction, the GQ network connected many of the major industrial, agricultural and cultural centers of India after its construction began in 2001. We find that districts located within 10 km to the GQ network experienced substantial increases in entry levels and higher organized sector productivity. The study suggests a significant increase in the overall output for the average district located on the GQ network, compared to no response in their nearby peers. We find that although entry effects for the organized sector are present in both urban and rural areas, the effects are much stronger and significant in rural settings. The differences are most substantial where new plants with major output levels are locating. By contrast, the productivity gains for the organized sector are very similar in both urban and rural locations.

Ghani et al. (2012) found that district level infrastructure is partly facilitating the relocation of organized manufacturing to rural locations while the unorganized manufacturing is migrating to urban locations. Our work on GQ suggests that such movement in the organized sector seems to be partially explained by national level highways as well, especially with regard to output of young plants. Since the unorganized sector is driving urbanization of Indian manufacturing, it is important for policy makers to understand the dynamics of this sector in the design of policies to promote urbanization. Our work suggests a very limited impact of the GQ upgrades on unorganized manufacturing outside of the nodal districts. We see traces of evidence of the organized sector findings repeating themselves in the unorganized sector but the results are substantially diminished in economic magnitudes. We confirm that these basic patterns are true in both urban and rural settings; they also hold true regardless of the gender of the business owner in the unorganized sector.

Our project contributes to the literature on the economic impacts of transportation networks and infrastructure investments in developing economies, which is unfortunately quite small relative to its policy importance (e.g., Ghani et al., 2012; Datta, 2011). Beyond India, several recent studies find positive economic effects in nonnodal locations due to transportation infrastructure in China (e.g., Banerjee et al. 2012; Baum-Snow et al. 2012; Roberts et al. 2012), Africa (e.g., Jedwab and Moradi 2015), and the United States (e.g., Fernald 1998; Chandra and Thompson 2000; Lahr et al. 2005; Baum-Snow 2007; Michaels 2008; Duranton and Turner 2011). This study also contributes to a literature seeking to understand the development of the manufacturing sector in India (e.g., Ahluwalia 2000; Besley and Burgess 2004; Kochhar et al. 2006) and especially
... those emphasizing the importance of infrastructure constraints (e.g., Mitra et al. 1998; Gupta et al. 2008; Gupta and Kumar 2010). Finally, our work also links to studies seeking to understand the allocation of activity across regions and the productivity of firms (e.g., Desmet et al. 2015, Hsieh and Klenow 2009).

I. DATA AND ESTIMATION METHODOLOGY

We employ repeated cross-sectional surveys of manufacturing establishments carried out by the government of India. Data for organized sector surveys are sourced from Annual Survey of Industries conducted in 1994–95 onwards to 2009–10. In all cases, the survey was undertaken over two fiscal years (e.g., the 1994 survey was conducted during 1994–1995), but we will only refer to the initial year for simplicity. This time span allows us two surveys before the GQ upgrades (1994 and 2000) began in 2001 and several years post the upgrade. For unorganized manufacturing, we employ plant-level data from the years 1994, 2000, and 2005. Ghani et al. (2013, 2015) provide additional details on the data sources and preparation.

To introduce the spatial impact of GQ on Indian manufacturing across urban and rural regions, we estimate a difference-in-difference estimation, where a pre-post analysis is conducted with explanatory variables being interactions of indicator variables for how far a district is from the GQ highway network with an indicator variable for the post-GQ upgrades (equal to one in 2005 and 2007). Indexing districts with \( i \) and years with \( t \), the specification takes the form:

\[
Y_{i,t} = \sum_{d \in D} \beta_d \cdot GQ \text{ Dist}_{i,d} \cdot \text{PostGQ}_t + \eta_i + \gamma_t + \epsilon_{i,t}. \tag{1}
\]

The set \( D \) contains three distance bands with respect to the GQ network: a nodal district, 0–10 km from the GQ network, and 10–50 km from the GQ network. The excluded category includes districts more than 50 km from the GQ network. The \( \beta_d \) coefficients measure, by distance band, the average change in outcome \( Y_i \) over the post period compared to the reference category. Most outcome variables \( Y_i \) are expressed in logs, with the exception of TFP, which is expressed in unit standard deviations. District fixed effects control for the overall levels of entry rates in each district across the sample; these fixed effects also control for the main effects of distance from the GQ network. In a similar manner, the year fixed effects control for aggregate annual changes in the Indian economy, including the main effects of the post-GQ upgrades period. Thus, the interactions and their \( \beta_d \) coefficients quantify differences in outcomes after the GQ upgrades by spatial band compared to the excluded group that comprises districts located more than 50 km from the GQ network.

Estimations report clustered standard errors, weight observations by log total district population in 2001, and have 312 observations representing the included districts. We winsorize outcome variables at the 1%/99% level to guard against...
outliers. Our district sample is constructed such that employment, output, and establishment counts are continuously observed.

II. IMPACT OF GQ UPGRADES ON ORGANIZED INDIAN MANUFACTURING

Table 1 presents the results from difference-in-difference estimations of equation 1 for organized manufacturing. Column headers provide the outcome variables studied. Column 1 considers the entry of young plants by their log count in the district, Column 2 considers the log employment in these young plants, and Column 3 considers the log output in these young plants. Columns 4–6 repeat the same for urban areas while Columns 7–9 reproduce the results for rural sections of a district.

The top row of Column 1 shows a very substantial increase in young firm counts in nodal cities after the upgrades in comparison to districts more than 50 km from GQ. We are very cautious about interpreting these results much given that the upgrades were built around the connectivity of the nodal cities. The imprecision in these estimates is mostly due to the fact that there are only nine nodal districts. As effects for our other distance categories are being measured for each band relative to districts more than 50 km from the GQ network, the inclusion or exclusion of the nodal districts does not impact our core results regarding nonnodal districts.

The key pattern emphasized in Ghani et al. (2015) is shown in the second and third rows. There is a substantial increase in organized sector entry in districts within 10 km of GQ that is not reflected in districts farther away at 10–50 km. Our sample includes 76 districts within 10 km, and 42 districts in the 10–50 km band. To some degree (substantiated further in Ghani et al. 2015), the upgrades of the GQ network can be taken as exogenous for these districts since neither group is in the nodal district set around which the upgrades were based. The districts within 0–10 km of GQ have a 0.4–0.9 log point increase in entry activity after the GQ upgrade compared to districts more than 50 km away.

The remainder of this table applies the methodology from Columns 1–3 to urban and rural areas of districts independently. Some districts do not contain an urban or rural area. Specifically, of our 312 district, 278 have urban areas and 291 have rural areas. In the estimations that consider levels of activity in urban areas, we will restrict our samples to districts that have manufacturing establishments in urban areas in all the years surveyed. In a similar manner, the estimates for rural areas require that the district have rural areas across the full sample period. This provides a consistent panel for each group.

The results for urban areas in Columns 4–6 are more muted than the total effects document in Columns 1–3. We find economically and statistically significant increases in plant entry rates for Column 4. By contrast, Columns 5 and 6 find positive point estimates for employment and output, respectively, but these coefficients are not precisely estimated like the base results in Columns 2 and 3.
Table 1. Pre-post estimations of the impact of GQ improvements in urban vs. rural areas for organized sector entrants

<table>
<thead>
<tr>
<th></th>
<th>Plants</th>
<th>Employment</th>
<th>Output</th>
<th>Plants</th>
<th>Employment</th>
<th>Output</th>
<th>Plants</th>
<th>Employment</th>
<th>Output</th>
<th>Plants</th>
<th>Employment</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
<td>(8)</td>
<td>(9)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Base spatial horizon measuring effects relative to districts 50+ km from the GQ network</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post GQ upgrades*</td>
<td>0.702</td>
<td>1.167</td>
<td>1.647+</td>
<td>0.388</td>
<td>0.631</td>
<td>0.383</td>
<td>-0.165</td>
<td>0.202</td>
<td>0.723</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nodal district</td>
<td>(0.662)</td>
<td>(0.814)</td>
<td>(0.951)</td>
<td>(0.433)</td>
<td>(0.566)</td>
<td>(0.723)</td>
<td>(0.692)</td>
<td>(0.804)</td>
<td>(1.215)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post GQ upgrades*</td>
<td>0.436++</td>
<td>0.471+++</td>
<td>0.928+++</td>
<td>0.319+</td>
<td>0.221</td>
<td>0.443</td>
<td>0.442+++</td>
<td>0.597+++</td>
<td>1.059+++</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>District 0-10 km from GQ</td>
<td>(0.172)</td>
<td>(0.239)</td>
<td>(0.346)</td>
<td>(0.176)</td>
<td>(0.252)</td>
<td>(0.380)</td>
<td>(0.183)</td>
<td>(0.260)</td>
<td>(0.389)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post GQ upgrades*</td>
<td>-0.012</td>
<td>-0.056</td>
<td>-0.263</td>
<td>-0.009</td>
<td>-0.056</td>
<td>-0.351</td>
<td>-0.015</td>
<td>0.006</td>
<td>-0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>District 10-50 km from GQ</td>
<td>(0.240)</td>
<td>(0.357)</td>
<td>(0.537)</td>
<td>(0.225)</td>
<td>(0.312)</td>
<td>(0.427)</td>
<td>(0.261)</td>
<td>(0.380)</td>
<td>(0.620)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>District and year fixed effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1248</td>
<td>1248</td>
<td>1248</td>
<td>1112</td>
<td>1112</td>
<td>1112</td>
<td>1164</td>
<td>1164</td>
<td>1164</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** Districts are local administrative units that generally form the tier of local government immediately below that of India's subnational states and territories. These are the smallest entities for which data is available with ASI. Nodal districts include Delhi, Mumbai, Kolkata, and Chennai and their contiguous suburbs (Gurgaon, Faridabad, Ghaziabad, and NOIDA for Delhi; Thane for Mumbai). The indicator variable for District 0–10 km from GQ takes a unit value for non-nodal districts that have minimum straight-line distance from the GQ network of less than 10 km; other distance-related indicator variables are defined analogously. Estimations consider the logged values of economic activity for young plants of organized sector manufacturing activity in 312 Indian districts for 1994, 2000, 2005, and 2007 from the Annual Survey of Industries. Young plants are those that are less than four years old. The table estimates effects of GQ upgrades for nearby districts relative to districts more than 50 km from the GQ network. The Post GQ upgrades variable takes unit value for the years 2005 and after, once the GQ upgrades commenced in 2001. Outcome variables are winsorized at their 1% and 99% levels. Estimations report standard errors clustered by district, include district and year fixed effects, and weight observations by log total district population in 2001. * is used to denote the interaction of two variables, while +, ++, and +++ denote statistical significance at the 10%, 5%, and 1% levels, respectively.
The output coefficients are also substantially smaller, between one-third and a half the size of what is evident in Column 3 for the full sample.

When isolating rural areas of districts in Columns 7–9, the outcomes have several key differences. First, the rural entry responses are generally larger than those estimated for urban areas, regardless of the metric. Second, unlike the emphasis on plant counts in urban areas, the rural estimates place more emphasis on employment and outputs of the organized sector. That is, the GQ effort appears to have increased the entry rates of organized sector plants in both urban and rural settings, but the largest increases in terms of employment and outputs are seen in rural settings. In fact, the development of large employment and output bases for organized manufacturing after the GQ upgrades plays the strongest role in the aggregate gains experienced on these dimensions. Ghani et al. (2013) provide a full set of results and those across extended distance bands.

In a similar manner, Table 2 considers labor productivity and TFP estimations, using the Sivadasan (2009) methodology to calculate the Levinsohn-Petrin (L-P) style TFP estimates. Columns 1 and 2 show a core increase of productivity associated with the GQ implementations in districts nearby the highway system. These increases are primarily driven by adjustments in the incumbent establishments of the districts. We do not quantify the labor productivity and TFP changes of new entrants, as much of the impact of new entrants comes from the extensive margin and these plant-level traits are not defined in these cases. In contrast to the differences observed for entry rates across urban and rural areas, the productivity effects appear quite uniformly realized,

### Table 2. Pre-post estimations of manufacturing productivity for organized sector

<table>
<thead>
<tr>
<th></th>
<th>Total GQ organized sector effect, entrants</th>
<th>Columns 1–2 in urban areas</th>
<th>Columns 1–2 in rural areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor TFP</td>
<td>(1) (2)</td>
<td>(3) (4)</td>
<td>(5) (6)</td>
</tr>
<tr>
<td>Post GQ upgrades*</td>
<td>0.084 −0.022</td>
<td>0.037 0.010</td>
<td>0.126 0.011</td>
</tr>
<tr>
<td>Nodal district</td>
<td>(0.135) (0.033)</td>
<td>(0.145) (0.043)</td>
<td>(0.264) (0.167)</td>
</tr>
<tr>
<td>Post GQ upgrades*</td>
<td>0.177 +0.086++</td>
<td>0.165 0.098</td>
<td>0.198 ++0.099+</td>
</tr>
<tr>
<td>District 0–10 km from GQ</td>
<td>(0.093) (0.042)</td>
<td>(0.139) (0.066)</td>
<td>(0.099) (0.056)</td>
</tr>
<tr>
<td>Post GQ upgrades*</td>
<td>0.043 −0.005</td>
<td>−0.000 −0.021</td>
<td>0.115 0.068</td>
</tr>
<tr>
<td>District 10–50 km from GQ</td>
<td>(0.132) (0.074)</td>
<td>(0.186) (0.085)</td>
<td>(0.175) (0.099)</td>
</tr>
<tr>
<td>District and year fixed effects</td>
<td>Yes Yes Yes Yes</td>
<td>Yes Yes Yes Yes</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1248 1244</td>
<td>1108 1100</td>
<td>1160 1160</td>
</tr>
</tbody>
</table>

**Notes:** See Table 1. Labor productivity is calculated as total output per employee, and TFP is calculated as the residual of value added over capital and labor inputs in a standard Cobb-Douglas production function using the L-P Sivadasan technique. See Ghani et al. (2015) for details on the methodology. * is used to denote the interaction of two variables, while +, ++, and +++ denote statistical significance at the 10%, 5%, and 1% levels, respectively.
although the coefficients remain significant for rural regions vis-à-vis urban areas. The labor productivity elasticity is slightly stronger in rural areas, likely reflective of the larger-scale production functions that can be used in these locations.

Ghani et al. (2015) confirm these results for overall organized manufacturing using a long-difference estimation technique. Additionally, there are natural concerns about the endogenous placement of the GQ highway system. For example, policy makers might have known about the latent growth potential of regions and attempted to aid that potential through highway development. Ghani et al. (2015) address these concerns by providing three alternative robustness checks described next.

**Placebo Estimations**

Ghani et al. (2015) compare districts proximate to the GQ network to districts proximate to the North South-East West (NS-EW) highway network (see figure 1 for a map of both GQ and NS-EW). The idea behind this comparison is that

**Figure 1.** GQ and NS-EW Highway route structure
districts that are at some distance from the GQ network may not be a good control group if they have patterns of evolution that do not mirror what districts immediately on the GQ system would have experienced had the GQ upgrades not occurred. The null results observed for districts close to the NS-EW corridor in Ghani et al. (2015) provide a stronger foundation in this regard, especially as its upgrades were planned to start at the same time as those of the GQ network before being delayed. The identification assumption is that unobserved conditions such as regional growth potential along the GQ network were similar to those for the NS-EW system (conditional on covariates).

**Instrument Variable (IV) Estimations**

Continuing with potential identification challenges, Ghani et al. (2015) also consider if the GQ planners were better able to shape the layout of the network to touch upon India’s growing regions vis-à-vis the NS-EW planners. The districts 0–10 km from the GQ network are instrumented with being 0–10 km from a (mostly) straight line between the nodal districts of the GQ network. The identifying assumption in this IV approach is that endogenous placement choices in terms of weaving the highway towards promising districts (or struggling districts) can be overcome by focusing on what the layout would have been if the network was established based upon minimal distances only. Ghani et al. (2015) find that the first-stage relationship of this IV estimation is quite strong. The IV specifications generally confirm the OLS findings, and, in most cases, the null hypothesis that the OLS and IV results are the same cannot be rejected.

**Dynamic Estimations**

Dynamic patterns around these reforms provide additional assurance about the role of the GQ upgrades in these economic outcomes and insight into their timing. By separately estimating effects for each year, it is feasible to observe whether the growth patterns appear to follow the GQ upgrades hypothesized to cause them. As an example, we present results for dynamics estimations of output for young plants. Figure 2 plots the coefficient values for log new output and their 90% confidence bands. Vertical lines in the figure mark when the GQ upgrades began and when they reached the 80% completion mark. Effects are measured relative to 1994, and we see no differences in 1999 or 2000 for non-nodal districts within 10 km of the GQ compared to those 10–50 km apart. Once the GQ upgrades commence, the patterns for output of young plants is pretty dramatic. These patterns confirm that the timing of the GQ upgrades coincides with the timing of growth in output of young plants.

**III. Impact of GQ on Unorganized Indian Manufacturing**

This section summarizes the connection (or lack thereof) between GQ upgrades and the broad development of unorganized manufacturing activity. One effort in
this project was to implement the GQ empirical strategy used in Ghani et al. (2015) in the unorganized sector to compare against the large, rapid effects observed for the organized sector. Our unorganized sector data stops in 2005, compared to 2007 for difference-in-difference estimations of the organized sector. Although we hope in future work to extend this time frame to 2010 for both sectors, even by 2005 the organized sector shows a strong response in terms of young firm activity and productivity (Ghani et al. 2015) and sourcing/inventory management (Datta 2011). It is thus useful to quantify whether the unorganized sector has a similarly rapid response.

Table 3 reports one of the portions of our basic framework for the unorganized sector from Ghani et al. (2013) to show differences and similarities with the organized sector. We highlight the following key lessons here. First, unlike the organized sector, difference-in-difference estimations find that non-nodal districts close to the GQ network behave similarly to those located farther away with respect to aggregate levels of unorganized manufacturing. In Ghani et al. (2013) we show that this comparability of districts 0–10 km from the GQ network with those that are 10–50 km apart holds irrespective of the whether
the highway development close to the district was an upgrade or a new construction (although there are traces of a stronger response for new construction). Likewise, we do not find differences in unorganized activity depending upon the completion date of the GQ upgrade (e.g., considering areas completed prior to 2002) or when examining the gender balance of plant owners. In sum, it appears that the highway improvements had limited aggregate effects for the unorganized sector, especially in comparison to the organized sector. Khanna (2014) provides complementary evidence outside of manufacturing using nighttime lights data.

These results are reasonable given the structure of unorganized manufacturing in India. By definition, unorganized enterprises are small establishments. Most employees in these enterprises are not full time, and many of these establishments are household based. Given the informality of this sector, it is less likely that such establishments and their new location decisions would depend as strongly on materials and products and connectivity to the market through national highways. The small-scale nature of these establishments makes it more likely that they target local product and labor markets. Thus, it is likely that they are more dependent on state or district roads vis-à-vis national highways.

### Table 3. Pre-post estimations of the impact of GQ improvements on unorganized activity

<table>
<thead>
<tr>
<th></th>
<th>Log levels of total activity</th>
<th>Log levels of young firm activity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Plants (1) Employment (2) Output (3)</td>
<td>Plants (4) Employment (5) Output (6)</td>
</tr>
<tr>
<td>Base spatial horizon measuring effects relative to districts 50+ km from the GQ network</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post GQ upgrades*</td>
<td>0.184</td>
<td>0.259</td>
</tr>
<tr>
<td>Nodal district</td>
<td>(0.177)</td>
<td>(0.146)</td>
</tr>
<tr>
<td>Post GQ upgrades*</td>
<td>-0.086</td>
<td>-0.096</td>
</tr>
<tr>
<td>District 0-10 km from GQ</td>
<td>(0.105)</td>
<td>(0.109)</td>
</tr>
<tr>
<td>Post GQ upgrades*</td>
<td>-0.031</td>
<td>0.006</td>
</tr>
<tr>
<td>District 10-50 km from GQ</td>
<td>(0.116)</td>
<td>(0.122)</td>
</tr>
<tr>
<td>District and year fixed effects</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>1089</td>
<td>1089</td>
</tr>
</tbody>
</table>

Notes: Estimations consider the location of unorganized sector manufacturing activity in 312 Indian districts for 1994, 2000, and 2005 from the National Sample Statistics. Young plants are those that are less than four years old. The table estimates effects of GQ upgrades for unorganized sector in nearby districts relative to districts more than 50 km from the GQ network; Ghani et al. (2013) also includes extended spatial rings to measure effects relative to districts 200 km away from the GQ network. The Post GQ upgrades variable takes unit value for the year 2005 after the GQ upgrades commenced in 2001. Outcome variables are winsorized at their 1% and 99% levels, and entry variables are coded at the 1% level where no entry is observed to maintain a consistent sample. Estimations report standard errors clustered by district, include district and year fixed effects, and weight observations by log total district population in 2001. * is used to denote the interaction of two variables. See Ghani et al. (2013) for more details.
IV. Conclusions

Many discussions of India’s future highlight the need and expected growth of urbanization. Another set of discussions, which often overlap extensively, highlight the importance of enhancing India’s infrastructure (e.g., McKinsey 2010, 2012; World Bank 2012). In this paper, we have quantified the degree to which one such infrastructure project—the massive upgrades to the GQ highway network that connects four of India’s major cities—influenced the urbanization of the manufacturing sector in nonnodal locations.

The patterns are quite intriguing. The upgrades are connected to enhancements in organized sector activity in both urban and rural environments. Across the full span of results from Ghani et al. (2013), we conclude that increases in entry rates and plant productivity are fairly balanced between the two settings, with the main exception being that rural areas receive relatively more stimulus in terms of net output growth. By contrast, the unorganized sector of manufacturing is not closely linked to the GQ developments in either location. Thus, the GQ experience suggests that major interdistrict projects are likely to continue and accelerate the spatial adjustments of the organized sector and its general move towards rural locations. On the other hand, the GQ experience does not appear connected to the general movement of the unorganized manufacturing sector into Indian cities.

Our work on India’s GQ highway system provides an important input into policy choices. It builds a framework for estimating the likely quantitative impact of infrastructure development projects and also provides estimates of the relative impacts across districts by distance to the network, thereby offering additional insights into the distributional consequences of large scale infrastructure projects. This methodology would be applicable to similar settings, where, for instance, poor transportation infrastructure severely hinders economic activity. On the whole, the paper speaks to the severe constraints that inadequate infrastructure can have for the development of manufacturing in emerging economies and the potential growth that may follow from alleviating that constraint.

References


Do Deep Trade Agreements Boost Vertical FDI?

*Alberto Osnago, Nadia Rocha, and Michele Ruta*

An increasing number of preferential trade agreements (PTAs) and the rise of offshore production are distinctive features of the modern world economy. Exploiting the WTO database on the content of deep trade agreements, we build on the existing literature to investigate whether deep trade agreements between countries are related to vertical foreign direct investment (FDI). Specifically, we show that deeper trade agreements increase vertical FDI measured with a proxy constructed in Osnago, Rocha and Ruta (2015) with data on foreign firm ownership and subsidiaries’ revenues available in the ORBIS database. JEL codes: F13, F23

How are preferential trade agreements (PTAs) and global value chains (GVCs) related? Two recent features of the world economy brought this question to the forefront of trade research and of the trade policy debate. First, during the recent decades, technological innovation in communication and transportation enabled the unbundling of stages of production processes across time and space leading to an increase in offshoring. Second, starting at the end of the 1990s, more and more countries signed bilateral and regional PTAs.

The answer to this broad question can actually be less straightforward than one could think. Both offshoring and trade agreements present two distinct features. Offshoring can be done in two modes: outside or within the boundaries of the firm. When firms outsource the production of some stages outside their boundaries, that is, when firms engage in foreign outsourcing, they generate arm’s length trade. On the contrary, when firms offshore within their boundaries through (vertical) foreign direct investment (FDI) they generate within-firms trade.

Similarly, PTAs can be classified into two types. Traditional PTAs usually involve reciprocal market access exchanges involving tariff cuts and the reduction of other border measures. On the other hand, modern day PTAs often contain provisions that cover a wide array of non-tariff measures, both at the border and behind-the-border. The literature refers to these new trade agreements as “deep”
to distinguish them from traditional PTAs that focus only on market access commitments—sometimes referred to as “shallow” agreements.

In this paper we briefly illustrate available data on PTAs and the existing literature that studies the relationship between the depth of PTAs and offshoring. We then provide novel results that show how deep PTAs are associated to the mode of offshoring. The empirical evidence that we collect in this paper shows that deeper PTAs are related to more vertical FDI.

**What Are Deep PTAs and What Is Their Content?**

PTAs are usually thought of as reciprocal market access exchanges involving tariff cuts and the reduction of other border measures. With preferential tariffs approaching the zero lower bound, the coverage of PTAs in terms of policy areas has widened over time as documented by WTO (2011).

Modern day trade agreements increasingly contain provisions that cover a wide array of non-tariff measures, both at the border and behind-the-border. For example, several PTAs include provisions regulating technical barriers to trade (TBT), sanitary and phytosanitary (SPS) measures, rules on investment, intellectual property rights (IPR) protection, provisions on anti-corruption, competition policy, labor standards, and so forth.

In this paper we rely on the dataset on the content of preferential trade agreements constructed by the WTO. Following Horn et al (2010), the WTO mapped a total of 52 disciplines across 100 PTAs signed between 1958 and 2011.

The left panel of figure 1 shows that PTAs became deeper over time. Agreements signed between 1987 and 1991 included on average nine provisions whereas agreements signed between 2007 and 2011 included on average 15 provisions. The right panel of figure 1 lists the 20 most common provisions included in the set of agreements mapped by the WTO. As expected, all agreements include reductions in tariffs on manufacturing goods. At the same time, more than 50 percent of agreements include deeper provisions such as anti-dumping and countervailing measures (CVM), rules on competition, movement of capitals and intellectual property rights (TRIPS and IPR). Moreover, TBT, investment disciplines and SPS measures are often included in PTAs. Exploiting the WTO data set, we construct different measures of depth of PTAs based on the content of each agreement. The left panel of figure 1 reports the evolution over time of the average depth of PTAs, measured as the average number of provisions included in PTAs.

**Deep PTAs and Vertical FDI**

The recent wave of PTAs and the surge in offshoring have brought to the forefront of trade research and of the trade policy debate the question of how trade agreements relate to the international organization of production (see, e.g., Lawrence (1996), Baldwin (2011), WTO (2011), and Antras and Staiger...
**Figure 1. Evolution of Depth over Time and most Common Provisions, 1987–2011**

*Source: Authors’ calculations based on the WTO dataset on the content of PTAs.*
The key insight of the theoretical literature is that the “depth” of trade agreements is associated with the international fragmentation of production. Econometric studies are scarce but they suggest that there is a positive relationship between production networks trade and deep integration. Orefice and Rocha (2014) investigate empirically the dual relationship between deep PTAs and trade in parts and components. They find that signing deep trade agreements increases trade in parts and components. At the same time higher levels of trade in production networks increases the likelihood of signing deeper agreements.

In this paper we move the focus on the relationship between deep PTAs and offshoring within the boundaries of the firm, that is, vertical FDI. The key question is whether the depth of trade agreements between two countries is correlated with more vertical FDI.

**Empirical Strategy**

We estimate the following linear equation:

\[
\log(FDI_{ijkt}) = \beta_1 \text{DEPTH}_{ijt} + \beta_2 \text{INSTITUTIONS}_{jt} + \beta_3 \log(Tariff)_{ijkt} + \\
+ \beta_4 \text{BIT}_{ijt} + \gamma_1 X_{jt} + \gamma_2 X_{ij} + \delta_{kt} + \epsilon_{ijkt}
\]

where \( k \) is the parent’s sector, \( t \) is time, \( i \) and \( j \) are country indexes (\( i \) for the “origin” country and \( j \) for the “destination” country). The dependent variable \( FDI_{ijkt} \) is a measure of the intensive margin of vertical FDI. In order to quantify vertical FDI flows we apply an innovative methodology proposed by Alfaro and Charlton (2009), used also by Lanz and Miroudot (2011). Using firm level data obtained from the ORBIS dataset, we determine the ownership relationship among a very large number of firms. We then proxy vertical FDI from country \( i \) to country \( j \) in sector \( k \) at time \( t \) as the aggregate value of the revenues of all subsidiaries owned by firms in country \( i \) producing inputs for sector \( k \) in country \( j \).

The main variable of interest that captures the depth of the agreements is \( \text{DEPTH}_{ijt} \). In the empirical analysis, \( \text{DEPTH}_{ijt} \) can take one of four forms: a dummy equal to one if there is a PTA; the number of provisions included in the PTA; or the log of one of two indexes (\( \log(\text{Top 5}) \) and \( \log(\text{Top 10}) \)) constructed using a principal component analysis as in Orefice and Rocha (2014).

The control variable \( \text{INSTITUTIONS}_{jt} \) is rule of law in the country that receives FDI. \( \text{Tariff}_{ijkt} \) represents the level of tariffs imposed by the origin country \( i \) on product \( k \). This variable helps us to separate the impact of our PTA variable that goes beyond simple tariff liberalization. \( \text{BIT}_{ijt} \) is a dichotomous variable capturing the existence of a bilateral investment treaty between \( i \) and \( j \) at time \( t \).

1. See Osnago et al. (2015) for details on the identification of vertical FDI and the construction of the variable.
$X_{it}$ is a vector of controls for characteristics of the destination country that vary over time. It includes GDP, GDP per capita and destination country remoteness. It also includes the average depth of the agreements signed by the host country with third countries in order to capture the overlapping nature of PTA networks. We also control for country-pair characteristics by including a standard set of gravity variables $X_{ij}$ such as geographical distance, contiguity, common language, and colonial relationship. We finally include sector-country-time fixed effects $\delta_{iklt}$ in order to control for potential omitted variables bias.

**Key Findings**

We show that deeper PTAs are associated to more vertical FDI originating from Germany, Japan, and the United States. The first four columns in table 1 report the coefficients of equation 1 estimated using OLS. Focusing only on the number of provisions for ease of interpretation, column 2 shows that the inclusion of one additional provision in the agreement is associated with an increase in vertical FDI of 1.7 percent.

The relationship between vertical FDI and deep PTAs can go in both directions. Deep PTAs may stimulate the creation of GVCs by facilitating trade of intermediate goods and FDI flows between potential members of a production network. On the other hand, firms involved in intense vertical FDI may lobby for deeper trade agreements in order to secure and increase the profitability of their investments in partner countries.

In order to deal with potential endogeneity issues we also use an instrumental variable approach. We instrument PTA depth between country $i$ and country $j$ with the weighted average depth of all the agreements signed by $i$ and $j$ with third countries. The rationale of this instrument comes from the domino theory of PTAs first introduced by Baldwin and Jaimovich (2010): the higher the level of integration between a country $j$ and its partners, the higher the probability that country $i$ will sign a PTA of similar depth with $j$ to avoid trade diversion effects.

The instrumental variable estimations are reported in columns 5 to 8 of table 1. The coefficients of all our measures of depth remain positive and significant suggesting that deeper integration is an important factor driving the make-or-buy decision of firms. One additional provision increases the value of FDI by almost 5 percent.

2. Rule of law comes from the Worldwide Governance Indicator database, tariffs from TRAINS, BIT from UNCTAD, GDPs from the World Bank World Development Indicators and the gravity variables from CEPII.

3. We restrict our analysis to these countries since they represent the regional hubs of production networks. Our sample is also restricted to the years 2003, 2007, and 2011 for practical reasons.

4. This type of instrumental variable approach has already been used in the literature (see, e.g., Orefice and Rocha 2014).

5. A similar argument has been provided by Chen and Joshi (2010).
Table 1. Vertical FDI and Deep Integration

<table>
<thead>
<tr>
<th></th>
<th>FDI (log of revenues in 1000$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>PTA</td>
<td>0.643***</td>
</tr>
<tr>
<td></td>
<td>(0.220)</td>
</tr>
<tr>
<td>N. of Provisions</td>
<td>0.0169**</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
</tr>
<tr>
<td>log(Top 5)</td>
<td>0.601*</td>
</tr>
<tr>
<td></td>
<td>(0.323)</td>
</tr>
<tr>
<td>log(Top 10)</td>
<td>0.538**</td>
</tr>
<tr>
<td></td>
<td>(0.323)</td>
</tr>
<tr>
<td>Avg depth of PTAs by j</td>
<td>0.035**</td>
</tr>
<tr>
<td>Rule of Law</td>
<td>0.364***</td>
</tr>
<tr>
<td></td>
<td>(0.135)</td>
</tr>
<tr>
<td>Tariff (log)</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>(0.195)</td>
</tr>
<tr>
<td>BIT</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>(0.149)</td>
</tr>
<tr>
<td>Observations</td>
<td>4,816</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.350</td>
</tr>
<tr>
<td>( \delta_{it} ) FE</td>
<td>Yes</td>
</tr>
</tbody>
</table>

\*\*\*p < .01, \*\*p < .05, \*p < .1.

Country-industry-year fixed effects refer to the country of the parent firm. All regressions control for distance, contiguity, colony relationship, common language, a dummy for China, GDP, GDP per capita, average integration and remoteness of the country of the subsidiary. Robust standard errors in parentheses clustered at the 6-digits NAICS.
CONCLUSIONS

A better grasp of the relationship between PTAs and offshoring is important in a world where countries sign more and more trade agreements and firms increasingly seek to engage in international production networks. In this paper, we provide novel evidence that the depth of PTAs is also related to the mode of offshoring. In particular, our findings indicate that signing deeper agreements can increase the flows of vertical FDI between countries.

In a related paper (Osnago et al. 2015), we take a step forward and examine how the content of PTAs, that is, the type of provisions included, is related to firms’ choice between foreign outsourcing and vertical FDI. We find evidence that the positive link between the depth of PTAs and vertical FDI is driven by the provisions that improve the contractibility of inputs provided by foreign suppliers, such as regulatory provisions.

While more work is needed, this line of research is contributing to our understanding of how policymakers can design trade agreements to support firms’ integration into global value chains.

REFERENCES


ICT Use, Competitive Pressures, and Firm Performance in Mexico

Leonardo Iacovone, Mariana Pereira-López, and Marc Schiffbauer

In this paper we present a set of stylized facts regarding the relation between information and communication technologies (ICT) use, firm performance, and competition. Taking advantage of a novel firm-level data set regarding ICT for Mexico, we find that firms facing higher competition appear to be the ones that have more incentives to increase their use of ICT. Accordingly, even though there is indeed a positive relation between ICT use and firm performance, this effect is greater for firms that face higher competition pressures, which is consistent with the theoretical predictions of the trade-induced technical change hypothesis. JEL codes: E23, F14, L25, O33

The question of whether the digital revolution has had effects on productivity has been around for a long time. Firms expect to become more efficient as a result of using more information and communication technologies (ICT) and therefore to obtain positive returns from these investments. Accordingly, ICT has been the most dynamic component of total investment in OECD countries (OECD 2007). However, empirical studies have not found much evidence on this relationship. Moreover, this lack of evidence led to the so-called Solow Paradox during the 1980s in which the computer age was everywhere but in the productivity statistics (Solow 1987).

Recent evidence points to a reversal of this paradox as ICT is considered one of the factors that can potentially increase firm-level productivity (Syverson 2011). According to Brynjolfsson and Hitt (1996) previous nonsignificant results arise from mismeasurement, lags on the effects, as well as the fact that the impact

Leonardo Iacovone is a senior economist at the World Bank; his e-mail address is liacovone@worldbank.org; Mariana Pereira-López is a consultant at the World Bank and Universidad Iberoamericana; her e-mail address is mpereiralopez@worldbank.org; Marc Schiffbauer is a senior economist at the World Bank; his e-mail address is mschiffbauer@worldbank.org. We thank Gerardo Durand, Gerardo López, Simón Sánchez Trinidad, and Gabriel Romero from INEGI for their support with the data use in compliance with the confidentiality requirements set by the Mexican Laws. We also want to thank Martin Guzman from the Ministry of Economy for his help with Mexico’s trade data. Finally, thanks to Remi Jedwab, Bob Rijkers, Pablo Cotler, Irving Rosales, Alejandro Rodríguez, and Gabriela Luna, as well as the participants of the World Bank’s ABCDE Conference 2015 for their valuable comments. All errors are naturally our own. This paper was prepared as background study for the forthcoming World Development Report 2016 “The Internet for Development.”
of ICT on variety, quality, and other intangibles is more likely to be detected at the firm level than using more aggregated data. Using firm-level data for the United States, these authors find that computers have significant effects over output, and that their contribution to marginal product is at least as high as the one from other capital investments. Among other studies that obtain similar results are Stiroh (2002), Oliner et al. (2008), and Jorgenson et al. (2008).

Even though some agreement has been reached regarding the positive effects of ICT on productivity, there is still little evidence on the channels or the factors that could enhance these effects. For example, as Jorgenson et al. (2008) argue, in order to successfully leverage ICT investments, firms must undergo complementary investments such as changes in business organization, human resources, workplace practices, training, and so forth, which Brynjolfsson et al. (2002) refer to as intangible assets. In order to explain the differences in terms of productivity growth after 1995 between the United States and Europe, that had previously been widely studied, Bloom et al. (2012b) dig deeper in the analysis of this statement and conclude that these differences arise from the human resources practices of US firms.

On the other hand, there is a different strand of the literature that analyzes the effects of increasing competition over productivity. As Aghion et al. (2001) argues, the idea behind these studies is that firms decide to innovate and adopt new technologies in order to temporarily escape competition from neck-and-neck rivals. Oliner et al. (2008) analyze those industries that were under the most intense pressure as a result of the 2001 recession and find that they experienced the strongest productivity growth, but ICT does not appear to facilitate the adjustment to competitive pressures. However, recent studies that take advantage of the surge of Chinese competition as an external shock, find evidence in favor of trade-induced technical change, innovation, and thus productivity increases. For example, Bloom et al. (2015) find that this shock led to higher R&D, patenting, innovation, and productivity in the case of European countries. Iacovone et al. (2013) evaluate the effects of Chinese competition for Mexico both on the extensive (firm exit) and intensive margins, obtaining heterogeneous results depending on the size of the firm as small firms experience losses and in some cases are forced to exit the market, bigger firms exhibit the opposite effects leading to productivity growth.

In this paper we take advantage of a very rich and novel data set regarding ICT use in Mexico in order to shed some light on the relation between ICT use and firm’s productivity in the case of a developing country, which has been seldom studied. Furthermore, we build upon the ICT-productivity literature and extend it by taking into consideration the strand of literature that analyzes competition as a factor that speeds up the creative destruction process. Thus, we analyze whether facing higher competition pressures can enhance the relation between ICT use and productivity by providing the incentives for firms to either

---

1. See, e.g., Van Ark et al. (2003), Haltiwanger et al. (2003), and Van Ark et al. (2008).
pursue the organizational changes that the adoption of these new technologies require or simply to use them in a more efficient way.

I. Data

Our data are from Mexico’s National Survey on Information Technologies 2009 and 2013, which is a survey designed by the National Science and Technology Council (CONACYT) and conducted by the National Institute of Statistics and Geography (INEGI). This survey includes detailed firm-level data regarding ICT and some other firms characteristics and it is under constant revision in order to improve the quality of the information obtained.²

Using ENTIC 2009 and 2013, we were able to match 1,017 firms that appeared in both waves of the survey and using the correspondences between North American Industry Classification System (NAICS) codes, which is the sectoral classification used in ENTIC and the Harmonized System (HS1996) codes, we merged this information with trade data from the World Integrated Trade Statistics (WITS), World Bank. Therefore, we ended up considering only the manufacturing sector and our final sample consists of a balanced panel of 715 firms. Due to the representative random sampling design of the survey, as it is not meant to follow the same firms over time, big firms are more likely to appear in both waves of the survey. Thus, 87% of the firms in our sample are big firms and 40% are older than 27 years in 2012.

In order to measure ICT use, we rely on physical measures that were previously used by other authors such as Beaudry et al. (2006), Bloom et al. (2012b), and Bloom et al. (2015). As these authors argue, this kind of measures has the advantage of being recorded consistently in all firms and that they do not depend on price indexes. Furthermore, as Bloom et al. (2012b) argue, they are highly correlated with other measures such as IT capital stock per worker. In this sense, we analyze computers-per-worker, the share of labor with internet, and the share of labor with computer which should behave similarly to computers-per-worker and allows us to test the robustness of our main measure.

As our performance indicator considering the information available in ENTIC, we use sales-per-worker, which even though is an imperfect measure of productivity, has been widely used in previous studies.³ For example, Haltiwanger et al. (1999) and more recently Bloom et al. (2012a) use this indicator as their productivity measure. Table 1 shows descriptive statistics for firms in our sample in terms of performance, ICT use, and other firms’ characteristics.

² Even though the first wave of ENTIC was conducted in 2004, it was included as a short module of the Innovation Survey and most of the relevant information for this analysis is not compatible with what we observe in later waves. Additionally, due to the random sampling design of this survey, only 297 of the firms that were surveyed in 2004 can be found in the 2013 wave.

³ We are unable to construct TFP measures because the survey does not include information on physical non-ICT related capital stocks.
<table>
<thead>
<tr>
<th>Variable</th>
<th>2008</th>
<th></th>
<th></th>
<th></th>
<th>2012</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>sd</td>
<td>p10</td>
<td>p90</td>
<td>Mean</td>
<td>sd</td>
<td>p10</td>
<td>p90</td>
</tr>
<tr>
<td>Computers-per-worker</td>
<td>0.24</td>
<td>0.22</td>
<td>0.06</td>
<td>0.50</td>
<td>0.28</td>
<td>0.23</td>
<td>0.07</td>
<td>0.57</td>
</tr>
<tr>
<td>Share of labor with Internet</td>
<td>0.20</td>
<td>0.19</td>
<td>0.03</td>
<td>0.40</td>
<td>0.20</td>
<td>0.19</td>
<td>0.03</td>
<td>0.40</td>
</tr>
<tr>
<td>Share of labor with computer</td>
<td>0.21</td>
<td>0.19</td>
<td>0.05</td>
<td>0.43</td>
<td>0.24</td>
<td>0.22</td>
<td>0.06</td>
<td>0.50</td>
</tr>
<tr>
<td>Sales-per-worker (thousand pesos)</td>
<td>997.36</td>
<td>1,638.24</td>
<td>123.01</td>
<td>2,192.49</td>
<td>1,005.39</td>
<td>1,407.87</td>
<td>131.20</td>
<td>2,494.11</td>
</tr>
<tr>
<td>Share of white-collar workers</td>
<td>0.22</td>
<td>0.17</td>
<td>0.05</td>
<td>0.42</td>
<td>0.25</td>
<td>0.23</td>
<td>0.06</td>
<td>0.59</td>
</tr>
<tr>
<td>Investment fixed assets/worker (thousand pesos)</td>
<td>46.10</td>
<td>111.03</td>
<td>0.00</td>
<td>130.11</td>
<td>41.34</td>
<td>91.19</td>
<td>0.05</td>
<td>113.05</td>
</tr>
</tbody>
</table>

*The sample includes 715 firms.

Source: Authors’ calculations with data from ENTIC 2009 and 2013, INEGI.
Finally, in order to analyze whether firms face competition, following Iacovone et al. (2013), we use the natural shock generated by the increasing competition from China in the Mexican market. In this sense, we measure competition as the change in the share of China in total Mexican imports for each HS1996 subheading between 2000 and 2008.⁴

II. RESULTS

Fact 1. Firms facing higher competition have increased more their use of ICT.

First of all, we analyzed whether firms that faced more competition, measured as a value above the median in our competition variable, experienced higher increases in terms of ICT use. Figure 1 shows that for the period of 2008–2012 this is indeed the case. Firms producing goods that experienced higher import competition from China invested more in ICT; the difference is statistically significant at the 10% level for computers-per-worker and the share of labor with internet.⁵

A possible mechanism behind the relation between ICT use and competition is trade-induced technical change. Under this hypothesis, competition generates incentives for innovation and the adoption of new technologies, speeding up the creative destruction process (Bloom et al. 2015). In this sense, considering all the

Figure 1. Mean Change in ICT Use 2008–2012

Source: Authors’ calculations with data from ENTEC 2009 and 2013, INEGI. Differences in means significant at the: * 10% level, ** 5% level, *** 1% level.

Note: High Chinese competition is defined as a value above the median in the change of the share of China in total Mexican imports between 2000 and 2008.

⁴. Considering that according to data from Mexico’s Ministry of Economy, the U.S. market accounts for 80% of Mexican exports, we also analyzed the share of China in U.S. imports. Results are similar. This analysis is shown in Iacovone et al. (2015).

⁵. In Iacovone et al. (2015) this is tested using regression analysis and the results indicate that this relation is significant for the three ICT use variables considered, even after controlling for firm characteristics.
side investments in terms of changes in processes and organization that entail increasing ICT use, firms that face higher competition pressures have more incentives to pursue these investments.

**Fact 2. Firms that become more intensive in their use of ICT exhibit higher increases in productivity.**

In order to test whether the predictions from previous studies regarding the relation between ICT and firm performance hold for developing countries, we analyze if firms that showed higher increases in their use of ICT, also increased their productivity.

As figure 2 shows, firms that exhibited changes in their ICT use above the median showed much higher increases in sales-per-worker. Furthermore, as shown in the second panel of the figure, this is the case even when we analyze the residuals of a regression of the change in sales over firm size, age, investment in fixed assets per worker and the share of white-collar workers. That is, even controlling for firm characteristics, firms that increased their use of ICT became more productive.

It is important to note that one of the empirical problems associated with the analysis of this relation is that the decision of investment in ICT cannot be taken as independent from performance (Draca et al. 2007). Relying on the same data Iacovone et al. (2015) adopt an instrumental variable (IV) approach and find that the positive relationship between ICT use and productivity still holds.

**Fact 3. Firms that face higher competition and make more intensive use of ICT have a better performance.**

Testing the mechanism of the trade-induced technical change hypothesis, we analyzed the differences in terms of productivity according to the level of Chinese competition. We used the dummy variable that takes a value of one if our competition variable (change in the share of China in Mexico’s imports) is above the median. As depicted in figure 3, conditional on firm characteristics, only firms that face competitive pressures from China, exhibit improvements in their performance as a result of a more intensive use of ICT (column 2 in the figure). For firms that face low competition there does not appear to be a relation at all.

Once again, the results support the idea that competition from China generates the incentives for firms not only to increase ICT but to use it more efficiently and pursue the complementary investments that these technologies require, meaning training and organizational changes.

Due to the construction of our sample, by definition we are analyzing the effects of competition only in terms of the intensive margin, as the exit of firms (extensive margin) cannot be tested due to data limitations. Considering the results of Iacovone et al. (2013) this should be less of a problem for our purpose as smaller firms are the ones that tend to exit the market under competitive pressures and most of the firms included in our analysis are big firms.
Further evidence on this relation is provided by Iacovone et al. (2015). In that study we also conduct robustness tests and deal with the endogeneity problem of ICT that we already mentioned.

III. Conclusions and Policy Implications

Empirical studies have experienced difficulties in finding evidence about the positive effects of ICT on productivity. Furthermore, the wide variation in terms of

---

**Figure 2. Mean Change in Sales/Worker 2008–2012**

(a) Mean change in sales/worker  
(b) Mean change in sales/worker controlling for firm characteristics.  
*Source:* Authors’ calculations with data from ENIC 2009 and 2013, INEGI. Differences in means significant at the: * 10% level, ** 5% level, *** 1% level.  
*Note:* In panel (a) the means of the differences in ln(sales/worker) are shown for each ICT-use group. In panel (b) we take the residuals of a regression of the difference of ln(sales/worker) over size dummies, age dummies, ln(investment in fixed assets/worker) and the share of white-collar workers.
Figure 3. Change in Sales/Worker vs. ICT Use

Source: Authors’ calculations with data from ENIC 2009 and 2013, INEGI.

Note: For conditional ln(sales/worker) we use the residuals of a regression of ln(sales/worker) over size dummies, age dummies, ln(investment in fixed assets/worker) and the share of white-collar workers.
the elasticities for developed countries, mainly the United States and Europe, creates an expectation of heterogeneity in the effects by country, industry, and type of firm (Draca et al. 2007).

Therefore, it is important to extend this analysis to the case of a developing country such as Mexico, which has the additional characteristic of being one of the countries directly affected by the sudden increase in Chinese competition in recent years. This fact led us not only to analyze whether previous results hold but also to test the relation between ICT and competition.

Our results indicate that ICT alone is not going to improve firm performance unless appropriate incentives are provided through higher competition to use ICT more effectively by making complementary investments in reorganization and skills. That is, in order to cope with the pressure of Chinese competition, firms appear to either make more effective use of ICT or to be willing to make the organizational changes necessary to improve the returns of ICT use.

Thus, the implications for policy are straightforward and suggest going back to the basics and removing existing barriers to private sector competition. This does not mean just enforcing antitrust laws but designing policies aimed at opening markets, building the infrastructure to make them more connected, as well as eliminating red tapes, among other strategies.

References


Do Deep Trade Agreements Boost Vertical FDI?

Alberto Osnago, Nadia Rocha, and Michele Ruta

An increasing number of preferential trade agreements (PTAs) and the rise of offshore production are distinctive features of the modern world economy. Exploiting the WTO database on the content of deep trade agreements, we build on the existing literature to investigate whether deep trade agreements between countries are related to vertical foreign direct investment (FDI). Specifically, we show that deeper trade agreements increase vertical FDI measured with a proxy constructed in Osnago, Rocha and Ruta (2015) with data on foreign firm ownership and subsidiaries’ revenues available in the ORBIS database. JEL codes: F13, F23

How are preferential trade agreements (PTAs) and global value chains (GVCs) related? Two recent features of the world economy brought this question to the forefront of trade research and of the trade policy debate. First, during the recent decades, technological innovation in communication and transportation enabled the unbundling of stages of production processes across time and space leading to an increase in offshoring. Second, starting at the end of the 1990s, more and more countries signed bilateral and regional PTAs.

The answer to this broad question can actually be less straightforward than one could think. Both offshoring and trade agreements present two distinct features. Offshoring can be done in two modes: outside or within the boundaries of the firm. When firms outsource the production of some stages outside their boundaries, that is, when firms engage in foreign outsourcing, they generate arm’s length trade. On the contrary, when firms offshore within their boundaries through (vertical) foreign direct investment (FDI) they generate within-firms trade.

Similarly, PTAs can be classified into two types. Traditional PTAs usually involve reciprocal market access exchanges involving tariff cuts and the reduction of other border measures. On the other hand, modern day PTAs often contain provisions that cover a wide array of non-tariff measures, both at the border and behind-the-border. The literature refers to these new trade agreements as “deep”

Alberto Osnago (corresponding author) is a consultant at the World Bank. His e-mail address is aosnago@worldbank.org. Nadia Rocha is a senior economist at the World Bank, and Michele Ruta is a lead economist at the World Bank Group.
to distinguish them from traditional PTAs that focus only on market access commitments—sometimes referred to as “shallow” agreements.

In this paper we briefly illustrate available data on PTAs and the existing literature that studies the relationship between the depth of PTAs and offshoring. We then provide novel results that show how deep PTAs are associated to the mode of offshoring. The empirical evidence that we collect in this paper shows that deeper PTAs are related to more vertical FDI.

**What Are Deep PTAs and What Is Their Content?**

PTAs are usually thought of as reciprocal market access exchanges involving tariff cuts and the reduction of other border measures. With preferential tariffs approaching the zero lower bound, the coverage of PTAs in terms of policy areas has widened over time as documented by WTO (2011).

Modern day trade agreements increasingly contain provisions that cover a wide array of non-tariff measures, both at the border and behind-the-border. For example, several PTAs include provisions regulating technical barriers to trade (TBT), sanitary and phytosanitary (SPS) measures, rules on investment, intellectual property rights (IPR) protection, provisions on anti-corruption, competition policy, labor standards, and so forth.

In this paper we rely on the dataset on the content of preferential trade agreements constructed by the WTO. Following Horn et al (2010), the WTO mapped a total of 52 disciplines across 100 PTAs signed between 1958 and 2011.

The left panel of figure 1 shows that PTAs became deeper over time. Agreements signed between 1987 and 1991 included on average nine provisions whereas agreements signed between 2007 and 2011 included on average 15 provisions. The right panel of figure 1 lists the 20 most common provisions included in the set of agreements mapped by the WTO. As expected, all agreements include reductions in tariffs on manufacturing goods. At the same time, more than 50 percent of agreements include deeper provisions such as anti-dumping and countervailing measures (CVM), rules on competition, movement of capitals and intellectual property rights (TRIPS and IPR). Moreover, TBT, investment disciplines and SPS measures are often included in PTAs. Exploiting the WTO data set, we construct different measures of depth of PTAs based on the content of each agreement. The left panel of figure 1 reports the evolution over time of the average depth of PTAs, measured as the average number of provisions included in PTAs.

**Deep PTAs and Vertical FDI**

The recent wave of PTAs and the surge in offshoring have brought to the forefront of trade research and of the trade policy debate the question of how trade agreements relate to the international organization of production (see, e.g., Lawrence (1996), Baldwin (2011), WTO (2011), and Antras and Staiger
**Figure 1.** Evolution of Depth over Time and most Common Provisions, 1987–2011

Source: Authors' calculations based on the WTO dataset on the content of PTAs.
The key insight of the theoretical literature is that the “depth” of trade agreements is associated with the international fragmentation of production.

Econometric studies are scarce but they suggest that there is a positive relationship between production networks trade and deep integration. Orefice and Rocha (2014) investigate empirically the dual relationship between deep PTAs and trade in parts and components. They find that signing deep trade agreements increases trade in parts and components. At the same time higher levels of trade in production networks increases the likelihood of signing deeper agreements.

In this paper we move the focus on the relationship between deep PTAs and offshoring within the boundaries of the firm, that is, vertical FDI. The key question is whether the depth of trade agreements between two countries is correlated with more vertical FDI.

**Empirical Strategy**

We estimate the following linear equation:

\[
\log(FDI_{ijkt}) = \beta_1 DEPTH_{ijt} + \beta_2 INSTITUTIONS_{jt} + \beta_3 \log(Tariff)_{ijkt} + \beta_4 BIT_{ijt} + \gamma_1 X_{jt} + \gamma_2 X_{ij} + \delta_{ikt} + \epsilon_{ijkt}
\]

where \(k\) is the parent’s sector, \(t\) is time, \(i\) and \(j\) are country indexes (\(i\) for the “origin” country and \(j\) for the “destination” country). The dependent variable \(FDI_{ijkt}\) is a measure of the intensive margin of vertical FDI. In order to quantify vertical FDI flows we apply an innovative methodology proposed by Alfaro and Charlton (2009), used also by Lanz and Miroudot (2011). Using firm level data obtained from the ORBIS dataset, we determine the ownership relationship among a very large number of firms. We then proxy vertical FDI from country \(i\) to country \(j\) in sector \(k\) at time \(t\) as the aggregate value of the revenues of all subsidiaries owned by firms in country \(i\) producing inputs for sector \(k\) in country \(j\).

The main variable of interest that captures the depth of the agreements is \(DEPTH_{ijt}\). In the empirical analysis, \(DEPTH_{ijt}\) can take one of four forms: a dummy equal to one if there is a PTA; the number of provisions included in the PTA; or the log of one of two indexes (log(Top 5) and log(Top 10)) constructed using a principal component analysis as in Orefice and Rocha (2014).

The control variable \(INSTITUTIONS_{jt}\) is rule of law in the country that receives FDI. \(Tariff_{ijkt}\) represents the level of tariffs imposed by the origin country \(i\) on product \(k\). This variable helps us to separate the impact of our PTA variable that goes beyond simple tariff liberalization. \(BIT_{ijt}\) is a dichotomous variable capturing the existence of a bilateral investment treaty between \(i\) and \(j\) at time \(t\).

1. See Osnago et al. (2015) for details on the identification of vertical FDI and the construction of the variable.
$X_{\mu}$ is a vector of controls for characteristics of the destination country that vary over time. It includes GDP, GDP per capita and destination country remoteness. It also includes the average depth of the agreements signed by the host country with third countries in order to capture the overlapping nature of PTA networks. We also control for country-pair characteristics by including a standard set of gravity variables $X_{ij}$ such as geographical distance, contiguity, common language, and colonial relationship. We finally include sector-country-time fixed effects $\delta_{ikt}$ in order to control for potential omitted variables bias.

**Key Findings**

We show that deeper PTAs are associated to more vertical FDI originating from Germany, Japan, and the United States. The first four columns in table 1 report the coefficients of equation 1 estimated using OLS. Focusing only on the number of provisions for ease of interpretation, column 2 shows that the inclusion of one additional provision in the agreement is associated with an increase in vertical FDI of 1.7 percent.

The relationship between vertical FDI and deep PTAs can go in both directions. Deep PTAs may stimulate the creation of GVCs by facilitating trade of intermediate goods and FDI flows between potential members of a production network. On the other hand, firms involved in intense vertical FDI may lobby for deeper trade agreements in order to secure and increase the profitability of their investments in partner countries.

In order to deal with potential endogeneity issues we also use an instrumental variable approach. We instrument PTA depth between country $i$ and country $j$ with the weighted average depth of all the agreements signed by $i$ and $j$ with third countries. The rationale of this instrument comes from the domino theory of PTAs first introduced by Baldwin and Jaimovich (2010): the higher the level of integration between a country $j$ and its partners, the higher the probability that country $i$ will sign a PTA of similar depth with $j$ to avoid trade diversion effects.

The instrumental variable estimations are reported in columns 5 to 8 of table 1. The coefficients of all our measures of depth remain positive and significant suggesting that deeper integration is an important factor driving the make-or-buy decision of firms. One additional provision increases the value of FDI by almost 5 percent.

---

2. Rule of law comes from the Worldwide Governance Indicator database, tariffs from TRAINS, BIT from UNCTAD, GDPs from the World Bank World Development Indicators and the gravity variables from CEPII.

3. We restrict our analysis to these countries since they represent the regional hubs of production networks. Our sample is also restricted to the years 2003, 2007, and 2011 for practical reasons.

4. This type of instrumental variable approach has already been used in the literature (see, e.g., Orefice and Rocha 2014).

5. A similar argument has been provided by Chen and Joshi (2010).
**Table 1. Vertical FDI and Deep Integration**

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>IV</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
<td>(8)</td>
</tr>
<tr>
<td>FDI (log of revenues in 1000$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTA</td>
<td>0.643*** (0.220)</td>
<td></td>
<td></td>
<td></td>
<td>0.910*** (0.269)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N. of Provisions</td>
<td>0.0169** (0.007)</td>
<td>0.601* (0.323)</td>
<td></td>
<td></td>
<td>0.0492*** (0.009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log(Top 5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.190*** (0.440)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log(Top 10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.538** (0.234)</td>
<td></td>
<td></td>
<td>1.459*** (0.282)</td>
</tr>
<tr>
<td>Avg depth of PTAs by j</td>
<td>0.037** (0.015)</td>
<td>0.035** (0.016)</td>
<td>0.042*** (0.016)</td>
<td>0.041*** (0.016)</td>
<td>0.040*** (0.013)</td>
<td>0.028** (0.014)</td>
<td>0.043*** (0.014)</td>
<td>0.044*** (0.014)</td>
</tr>
<tr>
<td>Rule of Law</td>
<td>0.364*** (0.135)</td>
<td>0.344** (0.133)</td>
<td>0.342** (0.134)</td>
<td>0.351*** (0.135)</td>
<td>0.389*** (0.124)</td>
<td>0.409*** (0.128)</td>
<td>0.418*** (0.127)</td>
<td>0.419*** (0.127)</td>
</tr>
<tr>
<td>Tariff (log)</td>
<td>0.042 (0.195)</td>
<td>-0.198 (0.179)</td>
<td>-0.123 (0.181)</td>
<td>-0.096 (0.181)</td>
<td>0.107 (0.177)</td>
<td>-0.256* (0.152)</td>
<td>-0.001 (0.159)</td>
<td>0.027 (0.159)</td>
</tr>
<tr>
<td>BIT</td>
<td>0.075 (0.149)</td>
<td>0.026 (0.151)</td>
<td>0.012 (0.151)</td>
<td>0.028 (0.153)</td>
<td>0.121 (0.138)</td>
<td>0.167 (0.143)</td>
<td>0.155 (0.145)</td>
<td>0.155 (0.144)</td>
</tr>
<tr>
<td>Observations</td>
<td>4,816</td>
<td>4,692</td>
<td>4,692</td>
<td>4,692</td>
<td>4,816</td>
<td>4,692</td>
<td>4,692</td>
<td>4,692</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.350</td>
<td>0.346</td>
<td>0.345</td>
<td>0.346</td>
<td>0.348</td>
<td>0.340</td>
<td>0.336</td>
<td>0.340</td>
</tr>
<tr>
<td>δ_t FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

***p < .01, **p < .05, *p < .1.

Country-industry-year fixed effects refer to the country of the parent firm. All regressions control for distance, contiguity, colony relationship, common language, a dummy for China, GDP, GDP per capita, average integration and remoteness of the country of the subsidiary. Robust standard errors in parentheses clustered at the 6-digits NAICS.
CONCLUSIONS

A better grasp of the relationship between PTAs and offshoring is important in a world where countries sign more and more trade agreements and firms increasingly seek to engage in international production networks. In this paper, we provide novel evidence that the depth of PTAs is also related to the mode of offshoring. In particular, our findings indicate that signing deeper agreements can increase the flows of vertical FDI between countries.

In a related paper (Osnago et al. 2015), we take a step forward and examine how the content of PTAs, that is, the type of provisions included, is related to firms’ choice between foreign outsourcing and vertical FDI. We find evidence that the positive link between the depth of PTAs and vertical FDI is driven by the provisions that improve the contractibility of inputs provided by foreign suppliers, such as regulatory provisions.

While more work is needed, this line of research is contributing to our understanding of how policymakers can design trade agreements to support firms’ integration into global value chains.

REFERENCES


Exporters, Engineers, and Blue-collar Workers

Irene Brambilla, Daniel Lederman, and Guido Porto

This article investigates differences in the composition of employment between exporting and nonexporting firms. In particular, it asks whether exporting firms hire more engineers relative to blue-collar workers than nonexporting firms. In a stylized partial-equilibrium model, firms produce goods of varying quality and exporters tend to produce higher quality goods, which are intensive in engineers relative to blue-collar workers. Firms are heterogeneous and more productive firms become exporters and have a higher demand for engineers. The article provides causal evidence in support of these theories using the Chilean Encuesta Nacional Industrial Anual (ENIA), an annual census of manufacturing firms. The results from an instrumental variable estimator suggest that Chilean exporters indeed utilize a higher share of engineers over blue-collar workers. JEL Codes: F13, F14

There is widespread evidence of both a wage and an employment premium in exporting vis-à-vis nonexporting firms (Bernard and Jensen 1999; Bernard et al. 2007). This article investigates differences in the composition of employment of exporting firms. In particular, it asks whether exporting firms hire more engineers relative to blue-collar workers than nonexporting firms. This can happen because firms produce goods of varying quality and exporters tend to produce higher quality goods. As in the literature, more productive firms become exporters (Melitz 2003). In our framework, goods are produced with tasks using assignment rules (Acemoglu and Zilibotti 2001; Costinot and Vogel 2010; Acemoglu and Autor 2011). Production involves different tasks, such as managing, output, and design, and maintenance services. Exporters and nonexporters...
fill managerial tasks with skilled workers and fill maintenance services tasks with unskilled workers. However, output and design can be performed by engineers or by blue-collar workers. Engineers can deliver higher quality goods than blue-collar workers because they are more efficient in input supervision, assembly, and general product attributes that determine quality (Bernard and Jensen 1997; Yeaple 2005; Acemoglu and Zilibotti, 2001; Verhoogen 2008; Brambilla, Lederman, and Porto 2012; Kugler and Verhoogen 2012; Bustos 2014; Bastos, Silva and Verhoogen 2014; Caron, Fally and Markusen, 2014). For given factor prices, more productive firms can afford to enter export markets, choose to produce on average higher quality products, and hire on average a higher share of engineers relative to blue-collar workers than less productive, nonexporting firms.

To test these predictions, the empirical analysis exploits detailed information of firms demand for different tasks contained in the Chilean Encuesta Nacional Industrial Anual (ENIA). The results from an instrumental variable estimator show that Chilean exporters indeed utilize a higher share of engineers over blue-collar workers than nonexporting firms. The results provide causal evidence of the skilled tasks demanded by exporters relative to nonexporters in Chile.

The rest of the article is organized as follows. Section 1 sketches a simple model of exports, quality, and the demand for engineers and blue-collar workers. Section 2 discusses the main empirical results. Section 3 concludes.

I. A Model of Exports, Engineers, and Blue-collar Workers

We combine elements from Verhoogen (2008) with elements from Acemoglu and Autor (2011) and Costinot and Vogel (2010) in a simple partial equilibrium model of exports and employment composition. From Verhoogen (2008), we adopt the theoretical framework underlying the idea that exporting requires quality upgrades that are inherently intensive in skilled labor. This framework can also be found in Kugler and Verhoogen (2012), Brambilla, Lederman and Porto (2012), Brambilla and Porto (2016), and Bastos, Silva and Verhoogen (2014). From Acemoglu and Autor (2011) and Costinot and Vogel (2010), we adopt an assignment model of skills to tasks. In this setting, quality for exports is produced with a collection of tasks that can be performed by workers with different skills. Exporting firms assign higher skilled workers such as engineers to fundamentally similar tasks to achieve higher quality. As a consequence, exporting firms hire a higher share of engineers over blue-collar workers than nonexporters.

We adopt a multinomial logit demand framework. The aggregate demand function of product \( j \) with quality \( \theta \) and price \( p \) is

\[
x_j(p_j, \theta) = \exp(\alpha \theta_j - p_j),
\]

where \( \alpha \) captures quality valuation which, as in Verhoogen (2008), is higher for exporting firms.
Firms produce differentiated products in monopolistically-competitive markets. Firms can choose the level of vertical differentiation of their products \( \theta \). This depends, ceteris paribus, on the quality valuation parameter \( \alpha \). This gives firms the option to provide more vertical differentiation to export markets with higher \( \alpha \). Exporting, in turn, incurs a fixed cost \( F_x \).

As in Verhoogen (2008), production of output of quality \( \theta \) requires activities to produce physical units and activities to produce quality itself. We assume that the production of quantity (physical units of output) is separated from the production of quality. Firms produce quantity and quality with tasks (Acemoglu and Autor 2011). To simplify the exposition, we assume that there are three different types of tasks: maintenance services (cleaning, machine and building maintenance), production and design (production activities, accounting, packaging, marketing, engineering, logistics, input control, supervision), and managerial activities (firm direction). We refer to these tasks as task 1, 2, and 3, respectively. In turn, tasks are produced by either skilled or unskilled workers which earn wages \( w_H \) and \( w_L \), respectively.

Regarding quantity production, following Acemoglu and Autor (2011), tasks provide services that generate output. Letting \( y_x(i) \) be the task services, output \( x \) is:

\[
x = \exp \left( \sum_{i=1}^{3} \alpha_x(i) \ln y_x(i) \right),
\]

where \( y_x(i) \), with \( i=1,2,3 \), is the service of tasks \( i \), and \( \alpha_x(i) \) are the Cobb-Douglas parameters that capture the intensity of task \( i \). The production function has constant return to scale, \( \alpha_x(1) + \alpha_x(2) + \alpha_x(3) = 1 \). The production function of task services \( y_x(i) \) is Ricardian. The productivity of unskilled and skilled labor in task \( i \) is \( a_{Lx}(i) \) and \( a_{Hx}(i) \), respectively. We rank tasks in increasing order of skill intensity so that task 1 is the most unskilled intensive task, while task 3 is the most skilled intensive task. To determine the assignment of skills to tasks, we assume that wages are exogenous to the firm. The unit cost of using unskilled labor in task \( i \) is thus \( w_L / a_{Lx}(i) \), and the unit cost of using skilled labor is \( w_H / a_{Hx}(i) \). These assumptions imply that wages of unskilled and skilled workers are the same across tasks, but the workers’ productivity is not.

Without loss of generality, we consider a scenario where wages and productivity are such that

\[
\frac{a_{Hx}(1)}{a_{Lx}(1)} < \frac{a_{Hx}(2)}{a_{Lx}(2)} < \frac{w_H}{w_L} < \frac{a_{Hx}(3)}{a_{Lx}(3)}.
\]

This assumption implies that, in output production activities, firms allocate unskilled workers to tasks 1 and 2 and skilled workers to task 3. Given constant factor prices and labor productivity, these allocations deliver a cost function of
output which features constant returns to scale. The marginal cost function is

$$c = c(w_L, w_H),$$  \(\text{(4)}\)

which depends on factor prices (and technology), but it is independent of the quantity (and quality) produced.\(^1\)

Turning to the technology to produce quality, we assume that the production of quality requires the same collection of services produced by tasks 1 to 3 as in output production. The production function of quality is

$$\theta = \lambda \exp \left( \sum_{i=1}^{3} \alpha_{\theta}(i) \ln y_{\theta}(i) \right)^{\beta},$$  \(\text{(5)}\)

where \(\lambda\) is firm productivity in quality production. Firms differ in \(\lambda\) and these differences allow firms to select themselves into exporters and nonexporters and, within exporters, to define their export intensity (the share of exports in sales). To achieve an interior solution, we assume that \(\beta < 1\) so that the production function exhibits decreasing returns to scale and the marginal cost of quality production is increasing.\(^2\)

The production function of task services \(y_{\theta}\) is also Ricardian, with productivity \(a_{L,\theta}(i)\) and \(a_{H,\theta}(i)\), \(i=1,2,3\). Again, without loss of generality, consider a scenario where wages and productivity are such that

$$\frac{a_{H,\theta}(1)}{a_{L,\theta}(1)} < \frac{w_H}{w_L} < \frac{a_{H,\theta}(2)}{a_{L,\theta}(2)} < \frac{a_{H,\theta}(3)}{a_{L,\theta}(3)}.$$  \(\text{(6)}\)

This assumption implies that, in quality production activities, firms allocate unskilled workers to task 1 but skilled workers to tasks 2 and 3.

Our interpretation of these assumptions is as follows. Production of output and quality require similar tasks. In basic tasks such as cleaning, janitorial services, maintenance operations and so on, firms hire similar types of unskilled workers both in output and quality production. In more sophisticated tasks such as managerial services or firm direction, firms hire similar types of skilled workers in all firm activities. The firm’s sales director has similar skills as the design director. The difference is in task 2, operations and designs. In the production of

1. This result, which is a consequence of the assumption that the production of quantity and quality are independent, simplifies the solution of the model, but it is not strictly necessary for our conclusions. Extensions where the marginal cost of quantity production depends on quality can be found in Verhoogen (2008), Brambilla, Lederman, and Porto (2012), Kugler and Verhoogen (2012), and Bastos, Silva, and Verhoogen (2014). See, e.g., the review in Brambilla and Porto (2016).

2. This assumption guarantees that the choice of quality is bounded. A similar assumption is in Verhoogen (2008).
physical output, firms hire blue-collar workers; in the production of quality, firms hire engineers.

A key difference between output and quality production is the structure of the cost function. While output production is subject to constant marginal cost $c$, quality production features increasing marginal costs. The total cost function is:

$$ F = F(\theta, \lambda, w_L, w_H). $$ (7)

Given the production function for $\theta$, equation (6), $F'_{\theta} > 0$ and $F''_{\theta} > 0$. That is, the marginal cost of producing quality is positive and increasing in the level of quality. We also know that $F_{\lambda} < 0$ and that $\partial F_{\theta}/\partial \lambda < 0$ (the marginal cost of producing quality decreases with productivity). Note also that, given optimal quality, the production of quality is independent of the sales of physical units of the good. In this sense, $F$ is a fixed costs (in terms of $x$).

Having determined the technology of producing physical units and quality, we can now study firm choices of quality and price to maximize profits (net of fixed costs), $\pi = (p - c)x(p, \theta) - F(\theta)$. The first order conditions are

$$ p = 1 + c, $$ (8)

$$ \alpha x(p, \theta) = F'(\theta). $$ (9)

In this setting, firms charge the same price for goods of any quality. This price is a markup over the marginal cost $c$ (given by (4)). Since it is costly to produce quality, vertical differentiation occurs because firms can sell higher quantities of higher-quality products. The optimal choice of quality $\theta$ is determined by the equality of the marginal cost ($F'(\theta)$) and the marginal benefit (higher sales measured by $\alpha x(p, \theta)$) of quality provision. Since $\partial F_{\theta}/\partial \lambda < 0$, optimal quality increases with the productivity of the firm.

To derive the demand for different types of workers, consider a firm with productivity $\lambda$ that chooses a good of quality $\theta$. For the lowest skill tasks, task 1, only unskilled workers are utilized, both for output production and for quality production. Total unskilled utilization is $L(1) = \alpha_x(1) (p - 1)x/w_L + \alpha_{\theta}(1) F(\theta)/w_L$. For the highest skill tasks, task 3, only skilled workers are utilized. Total skilled utilization is $H(3) = \alpha_x(3) (p - 1)x/w_H + \alpha_{\theta}(3) F(\theta)/w_H$. For intermediate tasks, task 2, unskilled workers are utilized in output production and skilled workers in quality production. This can refer to an engineer, who monitors production, and a blue-collar worker, that operates machines. Unskilled utilization is

3. The result that firms charge the same price for goods of different qualities is probably unrealistic, but it is a simplification that allows us to easily illustrate the implications of productivity and of exporting on the demands of skilled tasks. See Brambilla and Porto (2016) for alternative modeling strategies.

4. For an interior solution, we assume that the marginal cost increases in $\theta$ at a sufficiently high rate (concretely, the second order condition for profit maximization is $F''(\theta) > \alpha^2 x(p, \theta)$).

5. For a detailed derivation of factor demands, see Brambilla, Lederman, and Porto (2015).
We want to establish whether there are differences in $\rho$ between firms, in particular between exporters and nonexporters. Firms differ in $\lambda$, quality productivity, and more productive firms choose to sell higher quality goods. A firm can separately choose a quality to sell domestically and a quality to sell abroad. As in the literature, export markets have a higher quality valuation (a higher $\alpha$ in the demand function (1) and demand higher quality. In the spirit of Verhoogen (2008), Figure 1 illustrates the model. The right panel of the figure plots the positive relationship between productivity $\lambda$ and quality $\theta$. The curve AA depicts the profile of domestic quality chosen by firms of varying $\lambda$. The curve BB, which depicts the profile in foreign markets, shifts up because of higher quality valuation abroad. There is a fixed cost of exporting and a (lower) fixed cost of entering the domestic market. This gives rise to two productivity cutoffs. Let $\lambda_{min}$ be the cutoff productivity to enter domestic markets and $\lambda_{exp}$ be the cutoff productivity of exporters. The solid curve shows the profile of average quality as a function of

\[ L(2) = \alpha_x(2)(p - 1)x/w_L, \]\n
while skilled utilization is

\[ H(2) = \alpha_\theta(2)F(\theta)/w_H. \]

To streamline the exposition and assess the empirical implications, we focus on the ratio of engineers to blue-collar workers:

\[ \rho(\theta) = \frac{H(2)}{L(2)} = \frac{\alpha_\theta(2)F(\theta)/w_H}{\alpha_x(2)(p - 1)x/w_L} \quad (10) \]
productivity. Average quality jumps at the exporting cutoff and increases steadily with $\lambda$. On average, exporters produce higher quality products than nonexporters. In addition, within exporters, firms with higher productivity $\lambda$ will ship a larger share of their sales abroad and will produce even higher quality output.

Consider now a more productive firm (i.e., a firm with a higher $\lambda$) that chooses to produce a higher quality $\theta$. It is easy to show that $\rho$ is increasing in $\theta$ at a decreasing rate.\(^6\) On the left panel of figure 1, we plot the relationship between the ratio of engineers to blue-collar workers and quality. The curve AA corresponds to the ratio utilized in the production for the domestic market. The curve BB is the ratio utilized in the production for foreign markets. BB lies above AA because, for a given quality, foreign markets with higher quality valuation also demand higher quantities. The solid curve depicts the average $\rho$ for different firms. As it can be seen, there is a jump in $\rho$ at the quality exporting cutoff. Exporters produce on average higher quality products and, consequently, hire on average a higher share and a higher ratio of engineers to blue-collar workers.

II. Evidence

We use two sources of data, firm-level data, and customs records. The firm-level data come from the Encuesta Nacional Industrial Anual (ENIA), an annual industrial census run by Chile’s Instituto Nacional de Estadística that interviews all manufacturing plants with 10 workers or more. It is a panel. The customs data provide administrative records on firms exports by destination. We manually matched both databases for the period 2001–2005. As a result, we built a 5-year panel database of Chilean manufacturing firms.

The data have several modules. The main module contains information on industry affiliation, ownership type, sales, exports, input use, imports of materials, workers and wages. Industry affiliation is defined at the 4-digit ISIC Revision 3 level, which totals 113 industries.

We are mostly interested in the employment information. The data on workers are presented at detailed categories, which allows us to explore the demand for different tasks. From the detailed employment records, we define the following tasks: management (directors), administrative services (accountants, lawyers), engineers (specialized skilled production workers), blue-collar activities (nonspecialized unskilled production workers), and general maintenance services (unskilled nonproduction workers).

Table 1 presents summary statistics for the key variables in our model. We present the unconditional averages as well as averages for exporting firms and nonexporting firms. In terms of employment, exporters are larger than nonexporters, as expected. On average, 39 percent of workers in Chilean firms are skilled workers and 61 percent are unskilled workers. Exporters utilize a higher

\(^6\) Note that since the production functions are Cobb-Douglas, the parameters $\alpha_x$ and $\alpha_\theta$ are fixed and, in particular, do not depend on the $x$ or $\theta$.\n
share of skilled workers (41 percent) than nonexporters (39 percent). Within the skilled and unskilled categories, we are especially interested in engineering and blue-collar employment. The share of engineering employment is roughly 18.1 percent at both exporters and nonexporters, while the share of blue-collar employment is 52.0 and 54.6 at exporters and nonexporters, respectively. The ratio of engineers to blue-collar workers is 4.7 at exporters and 2.1 at nonexporters. Finally, the average exporter ships around 32 percent of its sales abroad. Among all firms, exports accounts for only 5 percent of total firm sales.

To formally study the relationship between exporting and the composition of employment, we work with the following regression model

\[
y_{ijt} = \gamma E_{ijt} + x'_{ijt}\beta + \phi_i + \phi_{jt} + \epsilon_{ijt},
\]

where \( i \) is a firm, \( t \) is time, and \( j \) is an industry. Outcomes are denoted by \( y_{ijt} \) and export intensity (the ratio of exports to total sales) is \( E_{ijt} \). We add a vector \( x \), which includes firm level variables such as log total employment and initial conditions (sales and exporting status) interacted with year dummies to account for firm-specific trends. The regression includes firm fixed effects, \( \phi_i \) and industry year effects, \( \phi_{jt} \).

To estimate causal effects, we instrument \( E_{ijt} \) following a strategy similar to Park et al. (2010), Brambilla, Lederman, and Porto (2012), Bastos, Silva, and

| Table 1. Summary Statistics National Annual Industrial Survey Chile 2001–2005 |
|---------------------------------|------------------|------------------|
|                                 | All firms        | Exporters        | Nonexporters     |
| A) Skilled and unskilled labor  |                  |                  |
| Log skilled employment         | 2.37             | 2.47             | 2.36             |
| Log unskilled employment       | 2.88             | 2.88             | 2.87             |
| Share skilled employment       | 38.7             | 40.6             | 38.5             |
| Share unskilled employment     | 61.3             | 59.5             | 61.5             |
| B) Tasks                       |                  |                  |
| Log managerial employment      | 0.60             | 0.79             | 0.58             |
| Log engineering employment     | 1.22             | 1.36             | 1.21             |
| Log services employment        | 1.22             | 1.34             | 1.21             |
| Log blue-collar employment     | 2.71             | 2.72             | 2.71             |
| Log maintenance employment     | 0.46             | 0.48             | 0.46             |
| Share managerial employment    | 7.2              | 8.7              | 7.0              |
| Share engineering employment   | 18.8             | 18.2             | 18.8             |
| Share services employment      | 12.7             | 13.8             | 12.7             |
| Share blue-collar employment   | 54.4             | 52.0             | 54.6             |
| Share maintenance employment   | 6.9              | 7.3              | 6.9              |
| Ratio engineers/blue-collar   | 2.5              | 4.7              | 2.1              |
| C) Exports                     |                  |                  |
| Exports/sales                  | 0.05             | 0.32             | 0.00             |

Verhoogen (2014), and Brambilla and Porto (2016), among others. Intuitively, exogenous export opportunities for a firm are likely to arise when foreign demand expands and this will happen when exchange-rate variations make Chilean exports relatively cheaper. Given these exogenous changes, a firm will be more likely to take advantage of these export opportunities if it is exposed to those markets. A natural measure of destination exposure in this case is the share of a firm’s exports to that destination in total firm sales. As in Brambilla, Lederman, and Porto (2012), we also interact this variable with initial firm sales (i.e., log sales in 2001) to include any firm advantages in profiting from export opportunities based on firm size. To assess the power of these instruments, we can look at the first stage results. These are reported in panel A of table 2 for four specifications. In column 1, the specification includes only firm fixed-effects and year-effects. The instruments work well and have high explanatory power. In column 2, we add log employment to control for size. The instruments work well and have high explanatory power. In column 2, we add log employment to control for size. This means we compare

<table>
<thead>
<tr>
<th>Table 2. Exports, Engineers, and Blue-collar Workers IV Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>A) First stage Results</td>
</tr>
<tr>
<td>Average real exchange rate</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>Average real exchange rate * initial sales</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Average real exchange rate</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>R²</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>0.4129         0.4147                                 0.4144               0.4146</td>
</tr>
<tr>
<td>F-statistic</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>888.09          882.96                                  888.42               884.31</td>
</tr>
<tr>
<td>Prob &gt; F</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>0.0000          0.0000                                 0.0000               0.0000</td>
</tr>
<tr>
<td>B) Tasks</td>
</tr>
<tr>
<td>Share managers</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>-0.01**         -0.01                                  -0.01                -0.009</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Share engineers</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>0.09***         0.10***                                0.10***              0.10***</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Share services</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>-0.009          0.009                                 -0.009               -0.009</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Share blue-collar</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>-0.07**         -0.08**                                -0.08**              -0.08**</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Share maintenance</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>-0.01**         -0.01*                                 -0.01*               -0.01*</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Ratio engineers to blue-collar</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>14.42**         13.54**                                13.67**              13.58**</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Firm and year FE</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>Yes             Yes                                   Yes                  Yes</td>
</tr>
<tr>
<td>Log employment</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>No              Yes                                   Yes                  Yes</td>
</tr>
<tr>
<td>Industry-specific trends</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>No              No                                    No                   Yes</td>
</tr>
<tr>
<td>Firm-specific trends</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(1)              (2)                               (3)                (4)</td>
</tr>
<tr>
<td>No              No                                    No                   Yes</td>
</tr>
</tbody>
</table>

Notes: IV-FE regressions of employment shares on export intensity (exports/sales). The instruments are the weighted average the real exchange rate of a firm export partners and the weighted average of the real gdp of a firm export destinations. Column (1): firm fixed-effects and year fixed-effects; column (2): adds log total employment (firm size); column (3): adds controls for industry-specific trends (i.e., interactions between year dummies and industry dummies); column (4): adds initial conditions to control for firm-specific trends. Data are from the Encuesta Nacional Industrial Anual (National Annual Industrial Survey), Chile 2001–2005.
firms of equal size, with different export intensity. To account for industry trends, such as industry-specific growth processes, we add in column 3 interactions between year dummies and industry dummies. In column 4, we also add initial conditions to account for firm-specific trends (Brambilla, Lederman, and Porto 2012). The results are very robust and the magnitudes of the coefficients are also stable across specifications.

The causal impacts of export intensity on employment are reported in panel B of table 2. We find that the share of engineering employment is higher among exporters. This holds for all our specifications, that is, conditional on firm- and year-fixed effects only (column 1), conditional on size and fixed effects (column 2), and also conditional on industry- and firm-specific trends (columns 3 and 4). Evaluated at the average share of export sales, conditional on exporting (32 percent in our data), an exporter hires 3.2 percentage points more engineers than a nonexporter. This is compensated with lower shares of blue-collar employment. The shares of all other types of employment are not statistically different, except for the share of maintenance workers, which is only marginally smaller among exporting firms. These results can also be seen in terms of the ratio of engineers to blue-collar works. In all four specifications, the ratio increases with exports. These findings are consistent with our theory.

III. CONCLUSIONS

Chilean exporters demand more engineers relative to blue-collar workers. Foreign consumers value product quality and quality is intensive in engineering tasks. A simple partial equilibrium model formalized these mechanisms and the evidence from a panel of Chilean firms supports the model predictions. The findings have implications for empirical research and policy design. The notion that trade, and exports in particular, affects the wage premium and thus wage-inequality needs to be carefully assessed. Exporting from developing countries may raise the demand for specific sets of skills, thus creating potential inequality even within skilled labor categories. In Chile, our results show that export opportunities boost the demand for technical skills such as engineering skills. These conclusions should contribute to our understanding of the skilled tasks needed for exports, the role of potential education policies consistent with a successful long-run export performance, and the design of social policies to reduce wage inequality and help the losers from trade.

REFERENCES


The Causal Impacts of Child Labor Law in Brazil: Some Preliminary Findings

Caio Piza and André Portela Souza

This paper investigates the causal impact of the change in law of December 1998 that increased the minimum legal age of entry into the labor force from 14 to 16. We used a difference-in-differences (DD) approach to estimate the impact of this law change on labor force participation rates as a whole, as well as for the formal and informal sectors separately. Our results showed that the ban reduced participation rates for boys by 4 percentage points and that this effect was mostly driven by the informal sector. We found no effect on girls. JEL codes: J08, J22, J23

Literature on child labor has grown considerably over the last 15 years, and this is not just because of increasing data availability. Child labor has fallen over the years, but the worldwide figures are still alarming. According to the International Labour Organization (ILO), 264 million children aged 5 to 17 participated in the labor market in 2012. Of this, 168 million were child labor, that is, either under minimum legal age and/or working in hazardous activities.

Due to the negative externalities associated with children’s participation in the labor force, it is argued that the public sector could intervene by changing the circumstances that cause parents to send their children to work (see Basu and Van 1998). In fact, many countries have adopted bans or other mechanisms to break down the ‘inter-generational child labor trap’ (see Emerson and Souza [2003] and Edmonds [2008] for a survey).

Basu and Van (1998), for instance, argue that a parent’s decision to send a child into the labor force might be seen as a rational choice in a poor household facing many constraints. Given a few assumptions, the labor market may have multiple stable equilibria, one characterised by children participating in the labor force and depressed adult wages, and another in which children do not participate in the labor force and adult wages are higher. Because these two equilibria are stable, the authors argue that if children participate in the labor force, a

Caio Piza is an economist at The World Bank Development Research Group; his e-mail address is caiopiza@worldbank.org; André Portela Souza is a professor of economics at Sao Paulo School of Economics (EESP-FGV); his e-mail address is andre.portela.souza@fgv.br.
A ban could be put in place by the government to shift the economy from this equilibrium to one without child labor.\(^1\)

Although many policies have been put in place to fight child labor, too little is known about the causal impact of such interventions, particularly in developing countries where empirical evidence is almost nonexistent.\(^2\) This paper helps fill this gap by delving into the consequences of two recent federal legislations in Brazil aimed at children of a certain age range.\(^3\)

In December 1998, Brazil increased the minimum employment age from 14 to 16. This policy change gave rise to a natural experiment that we will investigate in this paper.\(^4\)

We used a difference-in-differences (DD) approach to estimate the impact of the law passed in December 1998. The results suggest that the change in law reduced labor force participation rates of 14 year-old boys as a whole, but not of girls, who seem to have shifted to the informal sector. The results for boys are interesting as they indicate that a law can be a powerful instrument that affects individuals, including those working informally.\(^5\)

The remaining part of this paper is organized as follows: The first section briefly discusses the change in law and provides the rationale as to how these two laws might affect children’s time allocation. The second section presents the data, while the third presents the identification strategy. The results are discussed in section 4, and the conclusion highlights the main findings of the paper and outlines some policy recommendations.

### I. The Change in Law of December 1998

The Brazilian Constitution of 1988 set the minimum legal age of entry into the labor market at 14. In 1990, a federal rule named “The Statute of Children and Adolescents”\(^6\)
established children and youth rights. Complementary to the Constitution of 1988, the statute is considered the legal framework for children and youth in the labor market. From 1988 to November 1998, the minimum legal working age in Brazil was 14, and individuals below age 17 were prohibited from working in hazardous conditions.

As a consequence of comprehensive modifications approved for the pension system on December 15, 1998, Constitutional Amendment No. 20 increased the minimum legal age for entry into the labor market from 14 to 16. This change in law did not affect children who turned 14 before the law was passed; but those who turned 14 after the change were banned from the formal labor force.

The relatively large informal sector in Brazil can cast doubts on the effectiveness of the law. However, the effect of this intervention on participation rates of the treatment group depends on its enforceability and also on the size of the problem it is trying to fix. If some of the children participating in the formal sector simply shifted to the informal sector after the ban, the effect of the law on children’s participation rates would be negligible or even positive. But, if some employers decided to no longer employ children below age 16 to avoid legal consequences (such as paying fines), the law would probably reduce participation rates in the informal sector as well.

II. Empirical Strategy

To estimate the effect of the law of 1998, children who turned age 14 after the law change of December 1998 (treatment) were compared with those who turned age 14 before the change in law (comparison). As differences in labor force participation can arise from differences in ages or cohorts, we applied two DD estimates as control for these potential confounding effects. One estimate uses as counterfactual the difference in participation rates between those who turned 14 after December 1997 and those who turned 14 before December 1997. This is the different-cohorts-and-same-ages DD. In the second estimate, the counterfactual is the difference in participation rates between those who turned 13 after December 1997 and those who turned 13 on or before December 1997. This is the different-ages-and-same-cohorts DD. Table 1 shows how the “treatment” and “comparison” groups are defined for these two estimates.

Since we controlled for age (in weeks) linearly, the first difference in the DD exercise can be interpreted as a parametric regression discontinuity design. The identification strategy for the DD depends on two assumptions: (1) any difference in labor force participation between the treatment and control groups exists in level but not in difference. That means the groups would show a common trend in the absence of the changes in law; and (2) all unobserved

7. ILO considers an individual 17 years or younger as a child. In this paper, the terms children, teenagers, and youth are used interchangeably.
8. The law was passed on December 15, 1998, and came into effect the following day.
characteristics that can be correlated with the eligibility status of the individual or other covariates are additive and time-invariant. In our first regression model, we assumed that any differences in participation rates between treated and comparison groups that are due to age effects, is constant over time. Whereas in the second regression model, we assumed a cohort effect that is fixed over time (see Meyer 1995).

The estimation of the impact of the law of 1998 on the outcomes of interest is conducted through the linear probability regression model:

$$Y_{it} = \beta_0 + X_{it}'\beta_1 + \beta_2 T_i + \beta_3 D_{99} + \delta T_i D_{99} + u_{it},$$

where $Y_{it}$ is the outcome variable of individual $i$ in time $t$, $X_{it}$ is the vector of observed characteristics of individual $i$ in time $t$. The vector includes dummies gender and ethnicity, the years of schooling of the household head, the age of the household head, the gender of the household head, and dummy variables for regions and the metropolitan region. In our first regression, $T_i$ is a dummy variable that equals 1 if individual $i$ turned aged 14 in the first semester of 1998 or 1999, and 0 if individual $i$ turned 14 in the second semester of 1997 or 1998, $D_{99}$ is a year dummy that takes value 1 in 1999 (after the law was passed) and zero for 1998 (before the law was passed), and $u_{it}$ denotes an idiosyncratic error term. For this regression, we also include as control the individuals age (in weeks). In our second regression, $T_i$ takes the value of 1 if individual $i$ in cohort $c$ turned age 14 in the first semester of 1999 and 0 if (s)he belongs to the cohort that turned 14 in the second half of 1998. Table 1 shows the groups under comparison in each case. Because the data provide exact dates of birth, we run the regressions using a 26 week-bandwidth (6-month) as can be inferred from table 1.

In both regression models, the parameter of interest, $\delta$, provides the intent-to-treat (ITT) effect. In other words, the impact of the ban on the cohort hindered from participating in the formal labor force at age 14.

### Table 1. Definition of the Eligible and Comparison Groups for the DD Estimates

<table>
<thead>
<tr>
<th>Different cohorts and same ages</th>
<th>PNAD 1998</th>
<th>PNAD 1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment group</td>
<td>Turned 14 between January and</td>
<td>Turned 14 between January and</td>
</tr>
<tr>
<td></td>
<td>June 1998</td>
<td>June 1999</td>
</tr>
<tr>
<td>Comparison group</td>
<td>Turned 14 between July and</td>
<td>Turned 14 between July and</td>
</tr>
<tr>
<td></td>
<td>December 1997</td>
<td>December 1998</td>
</tr>
<tr>
<td>Different ages and same cohorts</td>
<td>Turned 13 between January and</td>
<td>Turned 14 between January and</td>
</tr>
<tr>
<td></td>
<td>June 1998</td>
<td>June 1999</td>
</tr>
<tr>
<td>Treatment group</td>
<td>Turned 13 between July and</td>
<td>Turned 14 between July and</td>
</tr>
<tr>
<td></td>
<td>December 1997</td>
<td>December 1998</td>
</tr>
<tr>
<td>Comparison group</td>
<td>Turned 13 between July and</td>
<td>Turned 14 between July and</td>
</tr>
<tr>
<td></td>
<td>December 1997</td>
<td>December 1998</td>
</tr>
</tbody>
</table>
III. Data and Results

The sample used in this paper is drawn from the Brazilian household surveys (Pesquisa Nacional por Amostra de Domicílios – PNAD) of 1998 and 1999. The PNAD is an annual household survey that covers around 100,000 households and about 320,000 individuals. It is a major source of microdata in Brazil, and is a nationally representative survey that contains detailed information on each household’s socio-economic characteristics, demographic data, household income, and labor force status. The Brazilian Bureau of Statistics (IBGE) provides the exact reference date for when data collection is carried out and we used that information to identify the treatment and comparison groups. Our results are exclusive for urban areas.

We used three measures of child labor: labor force participation rates as a whole without distinguishing between sectors, participation in the formal labor force, and participation rates in the informal labor force.

Table 2 shows ITT estimates on child labor incidence and reports the mean for the comparison group before the change in law. As can be seen, participation rates among boys are much higher compared to girls, and predominantly informal. We could therefore expect a higher impact of the change in law on boys. The first block of table 2 compares different cohorts keeping the age differential constant, whereas the second block compares the same cohort before and after the change in law. The estimates are very similar and point to a reduction in child labor of 4 percentage points for boys. Compared to a different cohort, the fall in participation rates was of 36 percent over the comparison group. Instead, if we used participation rates of the comparison group one year before, the drop in participation rates approached 80 percent. In both cases, the effect is almost fully driven by a reduction in participation rates in the informal sector. We found no impact on girls.

IV. Conclusion

This paper contributes to the literature on the impact of child labor laws by delving into the short-term impacts of the Brazilian Constitutional Amendment of December 1998, which increased the minimum legal age of entry into the labor force from 14 to 16. Unlike other estimates available in the literature, we

---

9. According to IBGE, the reference date is the last week of September. There can be some variability in the exact day, but it is usually September 26 or 27.

10. The dummy variable for labour force participation rates takes the value of 1 if the individual is either (1) employed or looking for a job or was an active worker in the week of reference but was prevented from working due to external causes in the week of reference or (2) worked in the last 12 months, and zero otherwise. A formal worker is someone working with a work permit (carteira assinada) issued by the Brazilian Ministry of Labour and zero otherwise, whereas an informal worker is someone who works without such permission. This definition does not include domestic servants. In Brazil, domestic servants are covered by a separate legislation.
<table>
<thead>
<tr>
<th></th>
<th>Labor force participation rates</th>
<th>Participation rates—formal labor force</th>
<th>Participation rates—informal labor force</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Boys</td>
<td>Girls</td>
</tr>
<tr>
<td>Eligible* (D_{99}) (DD)</td>
<td>-0.019** -0.040*** 0.0035</td>
<td>-0.0048 -0.0074 -0.0016</td>
<td>-0.014* -0.032*** 0.0051</td>
</tr>
<tr>
<td></td>
<td>(2.40)</td>
<td>(2.89)</td>
<td>(0.45)</td>
</tr>
<tr>
<td>Eligible</td>
<td>-0.017*** -0.024*** -0.011**</td>
<td>-0.0091*** -0.016*** -0.0032</td>
<td>-0.0076* -0.0089 -0.0076*</td>
</tr>
<tr>
<td></td>
<td>(3.46)</td>
<td>(2.87)</td>
<td>(2.26)</td>
</tr>
<tr>
<td>(D_{99}) (1998 =0, 1999 =1)</td>
<td>0.040*** 0.068*** 0.012**</td>
<td>0.0078*** 0.011** 0.0046</td>
<td>0.032*** 0.057*** 0.0077</td>
</tr>
<tr>
<td></td>
<td>(6.53)</td>
<td>(6.25)</td>
<td>(2.05)</td>
</tr>
</tbody>
</table>

| Notes: Robust T-statistics in parentheses. *, **, *** Statistically significant at 10 percent 5 percent, and 1 percent, respectively. The controls include age (in weeks), dummy variables for gender (male), ethnicity (white), head’s years of schooling, age, and gender (=1 if male), dummy for states and metropolitan area. |
present estimates for both formal and informal sectors separately. DD estimates show that the ban reduced labor force participation rate for boys only. The impact is almost fully explained by the fall in participation rates in the informal sector. We found no impact on girls.

These preliminary findings suggest that the law was binding and general measures are more effective when the incidence is somewhat pervasive, as in the case of boys. One may conjecture the reasons why there is a strong effect among informal workers. It is possible that the child labor ban reduced verification costs by local authorities and increased chances of firms getting caught for breaking the law. Before the change in law, a child aged 14 could work formally. The cost of verification of a worker hired formally or informally might be greater than the simpler verification of whether a child is working at all. The firms seem to have responded to the change in the law by reducing the number of working children. Policymakers thus should have a broader perspective when they pass such laws, and take into account their heterogeneous effects on time allocation of children.

References


The Effect of Publicly Provided Health Insurance on Education Outcomes in Mexico

Carlo Alcaraz, Daniel Chiquiar, María José Orraca, and Alejandrina Salcedo

In this paper we study the causal effect of a large expansion of publicly provided health insurance on school enrollment rates and on children’s academic performance using the case of Mexico. Access to free health insurance could improve education outcomes directly by making household members healthier or indirectly by raising the amount of resources available for education expenses. Using a panel of municipalities from 2007 to 2010, we find that the expansion of the Mexican public health insurance program, Seguro Popular, had a large positive, statistically significant effect on school enrollment rates and on standardized test scores. JEL codes: I13, I15, I25, I38.

Access to free health insurance could have a positive effect on education, which in turn is a relevant determinant of growth and economic development. In this paper we exploit the case of a large expansion of a publicly provided health insurance program in Mexico and find it had a large positive effect on education outcomes. It is possible to consider different mechanisms through which publicly provided health insurance could improve education (school attendance or performance). The most direct channel is through better health of both parents and children. (See Miguel (2005) for a review on evidence from various countries or Glewwe et al. (2001) for the Philippines, Alderman et al. (2001) for the case of Pakistan, Miguel and Kremer (2004) for Kenya, Todd and Winters (2011) for Mexico, and Lavy et al. (2012) for Israel.)

Carlo Alcaraz is an economist at Banco de México; Daniel Chiquiar is Director General of Economic Research at Banco de México; María José Orraca is a Ph.D. student in economics at the University of Pennsylvania, currently on leave from Banco de México; Alejandrina Salcedo (corresponding author) is Director of Economic Measurement at Banco de México; her e-mail address is asalcedo@banxico.org.mx. The authors would like to thank Gustavo Rivera and Cristina Gutiérrez for very useful comments and help with the Seguro Popular data, as well as Sergio Cárdenas for sharing his knowledge about ENLACE. We would also like to thank Marco Gonzalez-Navarro, an anonymous reader, participants at the NEUDC 2012 conference, the LACEA Meeting 2013, the Annual Bank Conference on Development Economics 2015, and our colleagues from Banco de México for useful comments. The views and conclusions presented in this paper are exclusively of the authors and do not necessarily reflect those of Banco de México. A previous version of this paper circulated under the title “The Effect of Publicly Provided Health Insurance on Academic Performance in Mexico.”

Advance Access Publication April 21, 2016

© The Author 2016. Published by Oxford University Press on behalf of the International Bank for Reconstruction and Development / THE WORLD BANK. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Another mechanism through which access to free-of-charge publicly provided health insurance could improve educational outcomes is by increasing the amount of resources available for education expenses. Health insurance could reduce both out-of-pocket and catastrophic health expenditures. An increase in disposable income may facilitate access to food, transportation, books, pencils, and notebooks, which could in turn increase the return to time spent in school.\(^1\) (See Brown and Park (2002), Morán et al. (2003), and Lochner and Monge-Naranjo (2011) for evidence in this direction.) Moreover, a reduction on uncertainty regarding future income could also improve education outcomes when access to health insurance is provided (see Flug et al. (1998) and Gruber and Yelowitz (1999)).

Hanson and Woodruff (2003) argue that the main costs of education in poor countries are generally the foregone earnings of the child rather than the direct costs. Therefore, it is costly to keep children in school when they could instead be involved in productive activities. Access to free health insurance may alleviate budget constraints and therefore avoid the need to take children out of school and send them to work in order to reduce expenses and increase income. Edmonds and Schady (2012), Alcaraz et al. (2012) and Yang (2008) provide evidence in this direction.

In 2002, the Mexican government introduced Seguro Popular (SP), a health insurance program free of charge for households not covered by social security institutions, estimated to be around half of the country’s population. We use the introduction of this program to study the effect on enrollment rates of children attending primary and secondary school, and on standardized test scores for primary school children. We estimate a fixed effects model on a panel of municipalities where we control for the per capita amount of public transfers through the poverty reduction program Oportunidades, which has both a health and an education component. Although unfortunately we do not have pre-program data for education outcomes in the years before the introduction of SP, we perform a placebo test to show that there is no evidence that the evolution of education outcomes in municipalities that enrolled to SP earlier is statistically different from that in municipalities that received SP later on. Our findings suggest that SP coverage had a positive and statistically significant effect on children’s scores in standardized tests and on school enrollment of children to late primary and secondary school. Although the effect on younger children’s enrollment (early primary) is also positive, it is not statistically significant in the main specification.

Our reduced form estimates do not allow us to identify the relative importance of the different mechanisms through which health insurance improves schooling outcomes. However, given the strong evidence showing that the introduction of SP has led to savings in health expenditures that has been documented in previous papers (Gakidou et al. 2006; Knaul et al. 2006; Barros 2008; King et al. 1.

---

1. Budget constraints actually interfere with households’ education choices, particularly in higher grades. For example, according to a survey conducted by the Mexican Ministry of Education, 50 percent of Mexicans report lack of money to pay for material, transport or tuition as one of the main causes for not starting high school or dropping out from it (SEP 2012).
2009; Galárraga et al. 2010), it is likely that households are allocating part of those resources to education expenditures or that these additional resources are allowing parents to send children to school instead of work, thus increasing enrollment and helping children achieve higher test scores. We interpret the differential effect on older versus younger children in enrollment rate estimates as an indication that households could indeed have been budget constrained and that SP has relaxed these constraints.

I. EXPANSION OF THE PUBLIC HEALTH INSURANCE SYSTEM

The public health system in Mexico is divided in two sectors. One of them provides health services to the population covered by social security institutions managed by the government (mainly IMSS and ISSSTE). Access to these institutions is obtained through a formally registered employer and once workers are covered, their families can be beneficiaries as well. Households not covered by social security institutions, which comprise around half of the population, receive health services from the state governments. There was an important gap in terms of resources allocated to health care between these two sectors, and SP was introduced in 2002 as a means to close such gap and to improve health services in the latter sector (Barros 2008). The main eligibility criterion to enroll in SP is precisely not to be covered by a social security institution and health services are provided basically free of charge. (For a detailed description regarding SP, see Levy (2008), Barros (2008), and Bosch and Campos-Vazquez (2014).) Given the large segment of the population that SP intends to covers, the program was implemented gradually across the country and full coverage was expected to be achieved by 2013 (Levy 2008). As of 2012, SP had enrolled 52.7 million beneficiaries (Sistema de Protección Social en Salud 2012). Some authors like Barros (2008) and Bosch and Campos-Vazquez (2014) have argued that political factors could have influenced the rollout schedule of the program.

The evidence on the impact of SP on health is somewhat mixed (see Barros 2008; Knox 2008; King et al. 2009; Sosa-Rubí et al. 2009a, b). Evidence seems more compelling with respect to the income channel: SP has been found to reduce out-of-pocket health expenditure (Barros 2008; Galárraga et al. 2010), and catastrophic health expenditures (Gakidou et al. 2006; Knaul et al. 2006; Barros 2008; King et al. 2009; Galárraga et al. 2010).

2. IMSS stands for Instituto Mexicano del Seguro Social and ISSSTE stands for Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado. These are the two main social security institutions in Mexico.

3. Although in principle there is a fee schedule for registering to SP that depends on household income, Levy (2008) argues that it is unlikely that SP fees can be enforced and estimates that in 2006 fees represented only 0.8% of the program’s budget.

4. An important branch of the SP literature has focused on whether this program could induce informality (workers with access to social security institutions moving to jobs without such access), as suggested by Levy (2008). No strong evidence has been found in this direction (Barros 2008; Campos-Vázquez and Knox 2010; Azuara and Heckman 2010).
II. Empirical Model, Data, and Descriptive Statistics

In order to identify the effect of SP on education outcomes we estimate a fixed-effects model with school enrollment rates and academic performance at the municipality level as dependent variables, and SP coverage, along with an additional control, as independent variables. The regressions take the following form:

\[
\text{education outcome}_{it} = \beta^p SP_{it} + \gamma Op_{it} + \pi_i + \tau_{st} \cdot \text{year}_t + u_{it},
\]

where \( SP_{it} \) is SP coverage in municipality \( i \) at time \( t \); \( \pi_i \) are municipality fixed effects; \( state_s \) and \( year_t \) are state and year dummies, respectively, so that by considering the interaction in the regression, we include state-specific time fixed effects to control for possible changes over time that could affect all municipalities in the same state in a similar fashion; \( Op_{it} \) is per capita public expenditure in Oportunidades transfers in state \( i \) at time \( t \); and \( u_{it} \) is the error term. Municipality fixed effects allow us to control for observable and unobservable characteristics at the municipality level that do not change over time and that could simultaneously affect education and SP enrollment levels. This addresses possible endogeneity problems related to constant-over-time unobservables. We control for Oportunidades transfers per capita because as is well known (see Levy and Rodríguez 2005), it is a poverty reduction conditional cash transfer program that has important education and health components, and Oportunidades beneficiaries were particularly targeted to be enrolled to SP. Additionally, because resources for education and health expenditures are provided federally to state governments, the state-specific fixed effects we include should control for the availability of such resources.

In order to estimate equation (1), we construct a panel of municipalities with annual data from 2007 to 2010. We restricted the analysis to this time period due to data availability. Regarding enrollment rates, we take the number of children attending grade three in preschool to grade three in primary school and divide it by the number of children aged five to nine in the municipality, and we refer to this indicator as the enrollment rate to early primary school. We will refer to the number of children attending grades five in primary school to two in secondary school divided by the number of children aged 10 to 14 in the municipality as the enrollment rate to late primary and secondary school. Data on the number of children enrolled in each grade were obtained from the Ministry of Education webpage. Estimations on yearly population size at the municipality

5. For example, the recent economic crisis of 2009 could have affected children’s academic performance, and this could have taken place differentially across states depending on the way the crisis affected the region.

6. Our sample consists of 2,419 municipalities (98% of the total) over the span of four years. We do not have information for every municipality each year. In particular, data on ENLACE test scores for the whole state of Michoacan in 2007 is unavailable.
level were obtained from the Ministry of Health webpage, which are based on projections by the National Population Council (CONAPO).7

The academic performance variable we focus on is the publicly available score in the yearly national standardized test on academic achievement ENLACE (Evaluación Nacional de Logro Académico en Centros Escolares), which consists mainly of math and Spanish language questions, but may also include questions on other subjects, all related to topics in the official curricula. We use the mean score at the municipality level.8 Our fixed effects model should account for differences across time in the exam, such as possible changes in the difficulty of the questions or the implementation process.

Our indicator of yearly SP coverage by municipality was constructed as the number of persons enrolled in SP relative to the total population in the municipality.9 Data on the number of persons registered in SP at the end of each month come from the SP National Registry and was provided to us by the Ministry of Health. Again, estimations on yearly population size at the municipality level were obtained from the Ministry of Health webpage. A possible concern with this measure of SP coverage is that it may be correlated with some characteristics of the municipality such as the percentage of the population with access to social security institutions, which could in turn be linked to the municipality’s wealth, health status, or to the quality of its education system. However, municipality fixed effects address this concern under the assumption that the proportion of the population eligible for SP (without access to social security) is constant in time. This is reasonable considering the nature of the variable and the four years time span of this study.10

Data on the total amount of money spent on Oportunidades at the municipality level were downloaded from INEGI’s National States and Municipalities Information System (SIMBAD). Using total population data from the Ministry of Health as before and the consumer price index, we constructed per capita Oportunidades transfers in constant pesos.

Figures 1a, 1b, and 1c show kernel density estimates for the education outcomes. It can be noted that there is important variation between municipalities.11

---

7. A possible concern regarding the construction of our database is that CONAPO’s projections for 2006–2050 could be underestimating the size of the Mexican population (García 2011). This could be exacerbated when considering population by age group. It is also important to bear in mind that although in theory children in these age ranges should be attending the mentioned grades, there are some who lag behind or who did not enroll to school at the recommended age. Moreover, children from one municipality may go to school in a different municipality. Consequently, the enrollment rate variable may take values greater than one.

8. We took into account only those scores that according to the Ministry of Education satisfied a certain standard of quality, or were “representative” at the school level. For robustness we also performed exercises including all schools, and we find very similar results.

9. Given that school starts in August, the SP coverage variable used for the estimations of its effect on school enrollment rate corresponds to the coverage achieved up to that month of each year. In turn, we take SP coverage by April of each year for the estimation of its effect on standardized tests scores, since this exam takes place on that month.

10. As mentioned before, there is no evidence that SP has itself caused a shift in the proportion of households with and without access to social security.
throughout the period. Figure 1d shows that there is also important variation in SP coverage between municipalities in each of the four years that constitute our database. Also, in the early years of our sample a relevant proportion of municipalities had very low SP coverage (note the spike in 2007). Indeed, the distribution has shifted to the right over time, suggesting that coverage has increased across municipalities, which is important for our identification strategy.
Table 1 presents the main results of the estimations of equation (1) for all three dependent variables.\textsuperscript{11} Once we control for Oportunidades, we do not find a statistically significant effect of SP on early primary school enrollment (column 2). The coefficients on SP coverage indicate a positive and statistically significant effect on the enrollment rate to late primary and secondary school (column 4). Attending secondary school is generally more expensive, not only in terms of direct educational expenses, but also because of the higher opportunity cost of older children in terms of their foregone earnings. It could be that the decrease in budget constraints is allowing households to pay for the additional expenditures that the assistance of children to secondary school entails. More importantly, SP may be allowing households to send children to school instead of sending them to work. The comparison of these results with those for younger children suggests that this channel may indeed be more important for older children. Therefore, the evidence is consistent with the hypothesis that one of the channels through which SP coverage affects education is through an indirect income effect. An additional explanation that may account for the lack of significance in the results for younger children is the little variation in accumulated schooling before the age of 10, given that primary school attendance has been mandatory for many years in Mexico (Hanson and Woodruff 2003). This implies that there is a lower margin for increasing enrollment rates to early primary school.

Our results imply that going from no coverage to having all the population registered to SP would translate in an increase of 2.5 percentage points in the enrollment rate to late primary and secondary school (column 4 of table 1). However, SP is targeted to the population not covered by social security, and therefore not every person in the municipality is expected to be enrolled. A more reasonable range would be to go from no coverage to an enrollment rate of 52 percent, which was the mean SP coverage across municipalities in 2010. Considering an increase of this magnitude, our coefficient implies an increase of 1.3 percentage points in enrollment to late primary and secondary school. This is an increase of considerable magnitude, as it represents 1.6% of the mean enrollment rate to these grades in the relevant period.

With respect to the estimations on test scores of primary school children, the coefficient in column 6 indicates a positive and statistically significant effect of SP coverage on test scores. An increase in SP coverage of 52 percentage points would imply an increase of 5.9 points in the average test score, according to the reported coefficient. This corresponds to 15.6 percent of a standard deviation of the average score across municipalities.\textsuperscript{12} According to the World Economic

\textsuperscript{11} In all cases standard errors are clustered at the municipality level.

\textsuperscript{12} As another reference point, the gap between the score of the state in the 90th percentile of test scores across states and that in the 10th percentile is equal to 43.39 points. An increase in SP coverage of 52 percentage points would be associated to an increase in test scores equal to 13.6 percent of this gap.
Table 1. Fixed Effects Estimation Results of the Effect of SP on Education Outcomes

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Enrollment rate to early primary school</th>
<th>Enrollment rate to late primary and secondary school</th>
<th>ENLACE test score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(5)</td>
</tr>
<tr>
<td>SP coverage</td>
<td>0.036*</td>
<td>0.020</td>
<td>0.035**</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.015)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Oportunidades</td>
<td>0.004</td>
<td>0.004**</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.002)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.924***</td>
<td>0.926***</td>
<td>0.789***</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.005)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.284</td>
<td>0.349</td>
<td>0.081</td>
</tr>
<tr>
<td>Observations</td>
<td>9,491</td>
<td>9,342</td>
<td>9,491</td>
</tr>
<tr>
<td>Number of municipalities</td>
<td>2,409</td>
<td>2,408</td>
<td>2,409</td>
</tr>
<tr>
<td>Municipality fixed effects</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>State-specific time fixed effects</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Notes: Seguro Popular (SP) coverage is defined as the proportion of the population in the municipality enrolled in SP. ENLACE test scores correspond to the average score of the schools at the municipality. Enrollment rate to early primary school corresponds to the number of students registered to grades three in preschool to three in primary school at the municipality level divided by the population in the municipality aged five to nine years. Enrollment rate to late primary and secondary school corresponds to the number of students registered to grades five in primary school to two in secondary school at the municipality level divided by the population in the municipality aged 10 to 14 years.

***p < .01, **p < .05, *p < .1. Clustered standard errors in parenthesis.
Forum (2009), a policy that increases test scores in 10 percent of the standard deviation may be considered successful. We therefore believe that the effects we find are sizable, although in interpreting the results, one should remember that it may have taken a municipality around eight years to go from no coverage to 52 percent. On the other hand, it is also important to bear in mind that we are estimating the effects of a program that did not target academic performance directly.\(^\text{13}\)

We can provide suggestive evidence of the absence of pre-existing trends following a simple test (as was implemented by de Janvry et al. (2014)). The test basically consists of restricting the sample to those municipalities that, even though the program was already in place at the national level, enrolled to SP at a later stage, late enough for us to have information about the evolution of its education outcomes before SP was implemented. We then compare education outcomes of early adopters, among those in the restricted sample, with the outcomes of municipalities that enrolled to SP even later on. Using the restricted sample, we find evidence indicating that the trends in schooling outcomes are not correlated with the timing of enrollment to SP. The details are presented in the appendix.

IV. Final Remarks

The evidence presented in this paper suggests that the introduction of Seguro Popular, a program implemented in Mexico that intended to provide public health insurance essentially free of charge to population not covered by social security, had a positive and statistically significant effect on education outcomes, namely on enrollment rates to late primary and secondary school and on academic performance of children in primary school. Although the main mechanisms behind the effects are not identified in the paper, an income effect seems to be playing an important role. We find remarkable that SP has improved academic performance of Mexican children when its main goal is to provide better health services.

V. Appendix

In order to provide evidence on the absence of pre-existing trends in the education outcomes, we perform the following test. We restrict the sample to municipalities that had SP coverage below 10 percent by 2008, and estimate the following regression for the municipalities in this sub-sample:

\[
\Delta \text{education outcome } 2007 - 2008_i = \alpha + \delta \text{ SP in } 2009_i + \gamma \text{ Op } 2009_i + u_i \tag{2}
\]

where \(\Delta \text{education outcome } 2007 - 2008_i\) is the change between 2007 and 2008 in the enrollment rate to early primary school, enrollment rate to late primary and secondary school, or ENLACE test score at municipality \(i\); \(\text{SP in } 2009_i\) is a

\(^{13}\) As additional robustness checks we conducted exercises to consider other forms of time fixed effects. Results support those previously described.
Table A1. Relationship between the Date of Introduction of SP and pre-program Education Outcomes

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Change in enrollment rate to early primary school between 2007–2008</th>
<th>Change in enrollment rate to late primary and secondary school between 2007–2008</th>
<th>Change in ENLACE test scores between 2007–2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Received SP in 2009</td>
<td>-0.014 (0.010)</td>
<td>-0.002 (0.010)</td>
<td>0.554 (3.458)</td>
</tr>
<tr>
<td>Control for Oportunidades</td>
<td>×</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Municipalities with coverage below 10% in 2009</td>
<td>140</td>
<td>140</td>
<td>106</td>
</tr>
<tr>
<td>Municipalities with coverage over 20% in 2009</td>
<td>43</td>
<td>43</td>
<td>90</td>
</tr>
</tbody>
</table>

Note: The sample is restricted to municipalities that by 2008 had SP coverage below 10 percent. Observations that in 2009 had between 10 and 20 percent coverage are dropped. Received SP in 2009 is a dummy variable indicating whether SP coverage at the municipality was above 20 by 2009. ENLACE test scores correspond to the average score of the schools at the municipality. Enrollment rate to early primary school corresponds to the number of students registered from grades three in preschool to three in primary school at the municipality level divided by the population in the municipality aged five to nine years. Enrollment rate to late primary and secondary school corresponds to the number of students registered from grades five in primary school to secondary school at the municipality level divided by the population in the municipality aged 10 to 14 years.

***p < 0.01, **p < 0.05, *p < 0.1. Robust standard errors in parenthesis.
dummy variable that takes value 1 when SP coverage in 2009 in municipality $i$ was above 20 percent (early adopter) and zero if it still was below 10% (late adopter)\textsuperscript{14}; and $Op_{2009_i}$ is per capita public expenditure in Oportunidades transfers in 2009. Lack of significance of the $\delta$ coefficient would indicate that the date in which SP was introduced in a municipality is not correlated with the evolution of education outcomes, supporting the assumption of parallel trends in education outcomes across municipalities before the program was implemented. The first row of table A\textsuperscript{1} provides evidence supporting the assumption that pre-program trends in education outcomes were not correlated with the time of expansion of SP. Indeed, the coefficient of $SP_{2009}$ is not statistically significant for any of the three education outcomes we study.

\textbf{References}


\textsuperscript{14} Note that we have dropped those municipalities that had a coverage between 10 and 20 percent in 2009 with the purpose of not biasing our test towards not finding a significant result due to the fact that late adopters (control group) could be similar to those that registered a small increase in SP coverage, but large enough to be considered as early adopters (treatment group).


The Importance of Legitimacy

Robert Akerlof

Within organizations, there are typically limits to leaders’ legitimacy. This article explores how organizations are structured in the face of such constraints. The concept of legitimacy is formalized in the context of a single-agent moral hazard model. The principal can give the agent monetary incentives; in addition, he can give the agent an order. The agent finds it costly to disobey orders provided they are legitimate. We find that it may be optimal for the principal to take costly actions to bolster legitimacy. We argue that many organizational phenomena can be understood as attempts to bolster legitimacy. Examples include: rejection of overqualified workers, bureaucracy, merger decisions, and above-market-clearing wages. JEL codes: D02, L20, Z10

It is widely recognized outside of economics that rules are easier to enforce when they are considered legitimate. There are three reasons that are commonly cited. First, when agents see rules as legitimate, they feel a sense of duty to obey them. Second, agents are motivated to punish violators of rules they consider legitimate (e.g., by reporting them to authorities). Finally, attempts to enforce illegitimate rules tend to provoke anger and resistance. This suggests that legitimacy is important and the need for it might serve as a constraint on the way in which organizations are structured. This paper builds a simple model to explore the implications of such constraints. It highlights a number of organizational features that may be explained by them, including: rejection of overqualified workers, bureaucracy, and above-market-clearing wages.

We formalize the concept of legitimacy by introducing it into a single-agent moral-hazard model. A principal can use variable wages to incentivize an agent to exert effort. He also has a second tool: he can give the agent an order. The agent finds it costly to disobey legitimate orders but it is costless to disobey illegitimate orders. Orders are only legitimate if they are not too tough: specifically, if the order \( \theta \) is less than or equal to \( L \), where \( L \) parameterizes the extent of the principal’s legitimacy. We refer to the condition \( \theta \leq L \) as the “authority maintenance constraint.”

Robert Akerlof is an assistant professor at the University of Warwick. His e-mail address is r.akerlof@warwick.ac.uk. I am grateful to George Akerlof, Kaushik Basu, Bob Gibbons, Oliver Hart, Rachel Kranton, and seminar participants at MIT and the World Bank ABCDE conference for helpful comments.


© The Author 2016. Published by Oxford University Press on behalf of the International Bank for Reconstruction and Development / THE WORLD BANK. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Initially, we take legitimacy $L$ as exogenous. When legitimacy is high, it is optimal to use orders to incentivize the agent. When legitimacy is low, it is better to use monetary—that is, high-powered—incentives. We then examine the case where the principal can bolster legitimacy at a cost. The principal may find that cost worth incurring since this allows him to give the agent tougher orders. Specifically, we find it is optimal to bolster authority when legitimacy is in an intermediate range.

We will argue that a number of phenomena can be understood as attempts by organizations to bolster authority. To give just one example, firms’ hiring decisions are often influenced by legitimacy considerations since workers may be more or less respectful of authority. Bewley (1999) suggests that firms tend to reject “overqualified workers” for this reason. In interviews he conducts with personnel managers, they point to disobedience as the key reason for rejecting overqualified workers. One tells him: “Overqualification is a problem, just as is underqualification. You cannot fulfill the needs of an overqualified person. They will be unhappy and will be a problem.”

While legitimacy has received relatively little attention in economics, several recent papers stress its importance and seek microfoundations (in particular, see Bénabou and Tirole (2011), Akerlof (2016), and Basu (2015)). Chen (2013), looking at a dataset of World War I deserters, also provides empirical support that legitimacy matters for rule enforcement. Legitimacy is highly stressed outside of economics: by legal scholars, political scientists, and sociologists, among others (see, e.g., Blau (1964), Kornhauser (1984), Ostrom (1990), Tyler (1990), and Fagan and Meares (2008)). For instance, Blau argues that in the absence of legitimacy, rules will be disobeyed because “coercive use of power engenders resistance.” Ostrom (1990) suggests that “the legitimacy of rules...will reduce the costs of monitoring, and [its] absence will increase [the] costs.” Fagan and Meares point out that the illegitimacy of the police in inner cities is a key reason why it is difficult to root out gang activity. This article contributes to the literature on legitimacy by suggesting that it serves as a constraint on the way in which organizations can be structured.

The article proceeds as follows. Section 1 presents the model. Section 2 discusses a range of organizational phenomena related to the model. Section 3 concludes. Proofs of formal results are contained in the appendix.

3. Ostrom (1990), 204.
4. While there is relatively little work in economics on legitimacy per se, a number of papers suggest reasons why leaders would take costly actions to increase authority over followers. For instance, costly actions sometimes help a leader persuade followers to obey (see Hermalin (1998), Majumdar and Mukand (2004), and Van den Steen (2009)). Bolton et al. (2013) argue that resoluteness can help a leader coordinate followers. Efficiency wage models also suggest that it may be worthwhile to pay above-market-wages in order to put in place effective punishments for disobedience (see, e.g., Shapiro and Stiglitz (1984)).
I. A Simple Model

Statement of the Problem

Consider a setting with a risk-neutral principal and a risk-neutral agent. The principal’s payoff is given by $\pi = a_1 - w$, where $a_1$ is an effort choice of the agent and $w$ is the wage paid to the agent. The principal only observes an imperfect measure of $a_1$: $q \in \{b, l\}$. The probability $q = b$ depends upon $a_1$ and upon a second effort choice of the agent, $a_2$. Specifically, $\Pr(q = b) = a_1 + \lambda a_2$. One can think of $a_2$ as effort to make $q$ erroneously look good. $\lambda \geq 0$ parameterizes the quality of $q$ as a measure of $a_1$, where a higher value of $\lambda$ corresponds to a worse measure.

One tool available to the principal to incentivize the agent is the wage, which can be made a function of $q$: $w(q)$. The principal also has a second tool. He can give an order to the agent $(\theta)$ regarding the level of $a_1$ he would like.

The agent can participate or not. If he does not participate, he has an outside option that yields 0. If he does participate, the agent’s payoff is given by: $U = w - \frac{1}{2}(a_1^2 + a_2^2) - \frac{1}{2} a_1 \neq \theta \cdot D(\theta)$. $U$ is increasing in the wage $(w)$ and decreasing in effort $(a_1$ and $a_2)$. The agent also suffers a loss $D(\theta)$ if he disobeys the principal’s order.

We assume it is only costly to disobey orders that are “legitimate.” The principal’s order is only legitimate if $\theta \leq L$ (that is, if it is not too severe). $L$ parameterizes the extent of the principal’s legitimacy. Specifically, $D(\theta) = \left\{ \begin{array}{ll} \infty, & \theta \leq L \\ 0, & \theta > L \end{array} \right.$.

We will refer to the condition $\theta \leq L$ as the authority maintenance constraint—or AM constraint. If the AM constraint is satisfied, we will say the principal “maintains authority”; if it is violated, we will say the principal “eschews the use of authority.”

Solution to the Principal’s Problem

First-best. It serves as a useful benchmark to consider what would happen if effort were contractible (i.e., the first-best case). It is easy to show that the first-best levels of effort are: $a_1^{FB} = 1$ and $a_2^{FB} = 0$. $a_1^{FB} > 0$ since effort of type 1 is productive; on the other hand, $a_2^{FB} = 0$ since effort of type 2 has no productive value (it simply distorts the measure $q$).

Second-best. Now, let us turn to the principal’s actual problem (i.e., the second-best case). The principal’s problem is to choose a wage $w(q)$ and order $\theta$ so as to maximize his expected payoff $E(\pi)$. He maximizes subject to a participation constraint (PC) and an incentive compatibility constraint (IC). The participation constraint can be written as follows:

$$E(U) = [w(l) + (a_1 + \lambda a_2)(w(b) - w(l))] - \frac{1}{2} a_1^2 - \frac{1}{2} a_2^2 - 1_{a_1 \neq \theta} \cdot D(\theta) \geq 0 \text{ (PC)}$$

The incentive compatibility constraint the principal faces depends upon whether
he maintains authority (satisfies the AM constraint) or not. If he satisfies AM, he faces the following IC constraint:

\[ a_1 = \theta, a_2 = \lambda(w(h) - w(l)). \quad (\text{IC–AM}) \]

The IC constraint if he violates AM is:

\[ a_1 = w(h) - w(l), a_2 = \lambda(w(h) - w(l)). \quad (\text{IC–no AM}) \]

Notice that, if the principal eschews the use of authority (violates AM), he can obtain productive effort \( a_1 \) by paying a variable wage \( (w(h) > w(l)) \)—or put, differently, by using “high-powered incentives.” The problem with variable wages, though, is that they incentivize the agent to exert unproductive effort \( a_2 \) as well as productive effort. On the other hand, if the principal maintains authority (satisfies AM), he can use orders \( (\theta) \) rather than a variable wage to incentivize the agent. Orders have the advantage that they do not incentivize unproductive effort \( a_2 \).

The principal therefore faces a tradeoff. He can meet the AM constraint, in which case he faces a better IC constraint; or, he can violate AM and face a worse IC constraint. Clearly, when AM is a more restrictive constraint—legitimacy \( (L) \) is lower—the principal’s inclination to violate AM will be greater. This is the intuition that lies behind Proposition 1 (stated below). It shows that authority is maintained when \( L \) is high and eschewed when \( L \) is low.

**Proposition 1.** The solution to the principal’s problem depends upon whether his legitimacy \( (L) \) is low, medium, or high.

1. **High legitimacy** \( (L \geq 1) \): The principal orders the first-best \( (\theta = 1) \); incentives are low-powered \( (w(h) - w(l) = 0) \); and effort is first-best \( (a_1 = 1, a_2 = 0) \).
2. **Medium legitimacy** \( (1 > L \geq 1 - \frac{\lambda}{\sqrt{1+\lambda^2}}) \): The principal gives the maximum legitimate order \( (\theta = L) \); incentives are low-powered \( (w(h) - w(l) = 0) \); and effort is \( a_1 = L, a_2 = 0 \).
3. **Low legitimacy** \( (L < 1 - \frac{\lambda}{\sqrt{1+\lambda^2}}) \): The principal eschews the use of authority; incentives are high-powered \( (w(h) - w(l) = \frac{\lambda}{1+\lambda^2}) \); and effort is \( a_1 = \frac{1}{1+\lambda^2}, a_2 = \frac{\lambda}{1+\lambda^2} \).

If the principal has sufficient legitimacy \( (L \geq 1) \), he can achieve the first-best outcome, using authority rather than variable wages to incentivize the agent. Notice that, when \( L > 1 \), the AM constraint is actually non-binding (the principal has more authority than he needs). If the principal has an intermediate level of legitimacy, it is worth using authority rather than variable wages to incentivize the agent, but it is no longer possible to achieve the first-best. The AM constraint is
binding in this case: or, put differently, the principal sets the maximum legitimate order \((\theta = L)\). When legitimacy is sufficiently low \((L < 1 - \frac{\lambda}{\sqrt{1+\lambda^2}})\), it is no longer worth maintaining authority (the AM constraint is violated). The principal uses high-powered incentives rather than authority to obtain effort. Consequently, the agent exerts both productive effort \((a_1)\) and unproductive effort \((a_2)\).

**Bolstering Authority**

Now, suppose the principal can bolster his authority over the agent at a cost. Specifically, assume his legitimacy is given by \(L = L_0 + b\), where \(b\) denotes the principal’s effort to bolster his authority. The cost to the principal of bolstering is \(b^2\). Proposition 2, stated below, characterizes the solution to the principal’s problem in this case.

**Proposition 2.** It is optimal for the principal to bolster authority if and only if he has an intermediate level of legitimacy:

1. **High legitimacy** \((L_0 \geq 1)\): the principal does not bolster authority \((b=0)\), gives the first-best order \((\theta = 1)\), and incentives are low-powered \((w(b) - w(l) = 0)\).

2. **Medium legitimacy** \((1 - \sqrt{\frac{2\lambda^2}{1+\lambda^2}} \leq L_0 < 1)\): the principal bolsters authority \((b = \frac{1-L_0}{2})\), gives the maximum legitimate order \((\theta = L = 1+L_0)\), and incentives are low-powered \((w(b) - w(l) = 0)\).

3. **Low legitimacy** \((L_0 < 1 - \sqrt{\frac{2\lambda^2}{1+\lambda^2}})\): the principal does not bolster authority \((b=0)\), eschews the use of authority, and incentives are high-powered \((w(b) - w(l) = \frac{1}{1+\lambda^2})\).

We find that it is optimal for the principal to bolster his authority—choose \(b>0\)—when he has an intermediate amount of legitimacy. When he has high legitimacy \((L_0 \geq 1)\), it is not worth bolstering because he can achieve the first-best without doing so. When he has low legitimacy \((L_0 < 1 - \sqrt{\frac{2\lambda^2}{1+\lambda^2}})\), it is not optimal to bolster because he eschews the use of authority. He uses high-powered incentives rather than authority to obtain effort. But, when the principal has an intermediate amount of legitimacy, it is optimal to bolster because authority is worth maintaining and, by bolstering, the principal is able to give a tougher order to the agent. Put another way, bolstering is valuable because it loosens the AM constraint.

**II. Applications**

We will now discuss a range of real-world settings that correspond to the model. In particular, we will argue that a number of phenomena can be understood as attempts by organizations to bolster authority.
Who Receives Orders?

Firms’ hiring decisions may be influenced by legitimacy considerations. Imagine, for example, a firm is deciding between hiring worker A and worker B. Worker A is more skilled/productive; but worker B is more respectful of authority (the firm has legitimacy $L_B > L_A$ if it hires B). Hiring worker B is therefore a way in which the firm can bolster its authority. The model suggests that it may be optimal to hire B even though worker A is more skilled.

We mentioned in the introduction, for example, firms’ disinclination to hire “overqualified workers.” This disinclination has been documented by Bewley (1999) who finds that personnel managers are as wary of overqualification as they are of underqualification. While one could imagine other reasons for such wariness, such as adverse selection, Bewley’s interviews suggest that the main concern is, in fact, lack of respect for authority.

Who Gives Orders?

Just as respect for authority may vary with the choice of worker, it may also vary with the choice of manager. Hence, legitimacy considerations come into play in the selection of managers—and, relatedly, the extent to which authority is delegated within firms. Imagine, for instance, a firm has a choice between having manager A give orders to workers and delegating authority to manager B. Suppose it is costly to delegate to B—for instance, because B’s preferences are not aligned with the firm’s—but manager B has more legitimacy ($L_B > L_A$). The model suggests authority might be delegated despite the cost since doing so bolsters the firm’s authority. Alternatively, there might be benefits associated with delegation: for instance, manager B might have better information than manager A. When there are benefits to delegation, it can nonetheless be optimal for A to retain authority if A has greater legitimacy ($L_A > L_B$).

The sociologist Alvin Gouldner’s study of the General Gypsum Company, entitled Patterns of Industrial Bureaucracy, provides an example. Gouldner was initially puzzled by the degree of bureaucracy within the firm; but he came to understand that it served a purpose. The General Gypsum Company had a serious problem at their Oscar Center Plant where there were “few rules...and fewer still that were strictly enforced.” To deal with this problem, they initially tried installing a new plant manager, Vincent Peele; but, while Peele was well motivated, he lacked legitimacy, and consequently, his attempts to make reforms failed. When he tried to enforce a rule against absenteeism, for instance, workers were angered and retaliated by increasing their absences. Supervisors at the plant concluded the “rule just could not be enforced.” The company eventually worked out a better arrangement. They realized that if, rather than delegating to Peele, the central office set the rules, they would be perceived as more legitimate.

5. Gouldner (1954), 51.
6. Ibid., 142.
Thus, while the resulting organizational form was costly—highly centralized and bureaucratic—it helped the firm bolster its authority over workers.

**Multiple Agents**

Suppose a principal wants to incentivize multiple agents rather than just one. What it takes to be seen as legitimate by one agent (A) might be different from what it takes to be seen as legitimate by another (B). For instance, legitimacy over A might be given by $L_A = L_0 + b$ while legitimacy over B might be given by $L_B = L_0 - b$. In such a setting, it may be optimal to eschew authority over one agent so as to bolster authority over the other.

Such considerations are highly relevant for mergers. When firms have different cultures, merging them may pose a problem of authority maintenance. It may be better for such firms to have a market relationship instead (akin, in the model, to eschewing authority over one agent and using high-powered incentives).

Buono et al. (1985) describe a case where these issues are at play. They examine a 1981 merger of two mutual savings banks. While the banks were in many ways similar, different cultures prevailed. In particular, the leadership style in Bank A was somewhat democratic and participative while the leadership style in Bank B was more authoritarian. Both banks took pride in their cultures and, in each bank, the style of leadership prior to the merger was viewed favorably by employees. Following the merger, the CEO of Bank B played the key role in managing the merged entity (the other CEO became COO). According to Buono et al., the merger was not a success: primarily because there was anger in Bank A post-merger and resistance to the Bank B CEO’s authoritarian leadership style.

**Wage-Setting**

Legitimacy considerations are also relevant for wage-setting: since the level of pay within a firm—as well as pay dispersion—may impact upon legitimacy. Suppose, for instance, a firm’s legitimacy is increasing in the expected wage it pays its workers ($L = L_0 + f(E(w))$). It can be optimal to pay an above-market wage—that is, a wage that makes the participation constraint nonbinding—since doing so bolsters authority. The model therefore suggests a reason for efficiency wages. Observe that the rationale for paying efficiency wages is distinct from other models, such as Shapiro and Stiglitz (1984).

**III. Conclusion**

This article has argued that limited legitimacy of authority plays a significant role in determining organizational behavior and organizational structure. We formalized the concept of legitimacy in a single-agent moral hazard model. The model explains numerous organizational phenomena such as: rejection of over-qualified workers, bureaucratic organization, merger decisions, and above-market-clearing wages.
The article suggests a number of avenues for future research. We briefly mention three. First, there is a burgeoning literature on “persistent performance differences” (PPDs) across firms. To what extent might authority/legitimacy differences across firms explain PPDs—and, more widely, productivity differences across regions and countries? Heterogeneity in management practices are an important source of PPDs (see Bloom et al. (2013)). It has often been thought that heterogeneity in management practices is due to heterogeneity in managerial skill; but the model suggests they might also be accounted for by heterogeneity in legitimacy.

Second, the article considers an environment where decreases in legitimacy reduce welfare. But, in some contexts, low legitimacy might be beneficial. For instance, it might prevent abuse of authority or allow for better information aggregation.7

Finally, the model in this article is reduced form—since its goal is simply to show that legitimacy considerations matter—but there is a need for micro-foundations of legitimacy and for a deeper understanding of how orders come to be seen as more/less legitimate. Several papers mentioned in the introduction make a start towards this goal.

IV. Appendix

Proof of Proposition 1. Let us consider the maximum profits the principal can obtain if he violates AM and the maximum profits he can obtain if he meets AM and compare.

If the principal violates AM, he maximizes $E(\pi)$ subject to (PC) and (IC – no AM). It is easily shown that it is optimal to choose: $w(h) = \frac{1}{2(1+\lambda^2)}$ and $w(l) = -\frac{1}{2(1+\lambda^2)}$ (which implies $w(h) - w(l) = \frac{1}{1+\lambda^2}$). The resulting payoff to the principal is $E(\pi) = \frac{1}{2(1+\lambda^2)}$.

If the principal meets AM, he maximizes $\pi$ subject to (PC) and (IC – AM). It is easily shown that it is optimal to choose: $\theta = \min(1, L)$ and $w(h) = w(l) = \frac{1}{2} (\min(1, L))^2$. The resulting payoff to the principal is $E(\pi) = \min(1, L) - \frac{1}{2} (\min(1, L))^2$.

Observe that if $L \geq 1$, the principal’s profits from meeting AM are $\frac{1}{2} \geq \frac{1}{2(1+\lambda^2)}$, so it is optimal to maintain authority in this region. Now suppose $L < 1$. His profits are higher from meeting AM than not if and only if $L - \frac{1}{2} L^2 \geq \frac{1}{2(1+\lambda^2)}$. It is easily shown this constraint is equivalent to: $L \geq 1 - \frac{\lambda}{\sqrt{1+\lambda^2}}$. This completes the proof.

Proof of Proposition 2. Again, let us consider the maximum profits the principal can obtain if he violates AM and the maximum profits he can obtain if he meets AM and compare.

Suppose AM is violated. From the proof of Proposition 1 we know that the maximum payoff that can be obtained for a given value of $b$ is $\frac{1}{2(1+\lambda^2)} - b^2$. Clearly, it is optimal to set $b=0$, which results in a payoff of $\frac{1}{2(1+\lambda^2)}$.

7. See Landier et al. (2009) for one approach to this topic.
Now suppose AM is not violated. From the proof of Proposition 1 we know that the maximum payoff that can be obtained for a given value of $b$ is $\theta = \min(1, L)$ and $w(b) = w(l) = \frac{1}{2} \left( \min(1, L) \right)^2$, where $L = L_0 + b$. It is easily shown that it is optimal to set $b = 0$ when $L_0 \geq 0$ and set $b = \frac{1-L_0}{2}$ when $L_0 < 1$. The resulting payoff to the principal is $\frac{1}{2}$ when $L_0 \geq 1$ and $(\frac{1+L_0}{2})^2 - (\frac{1}{2} \left( 1-L_0 \right))^2$ when $L_0 < 1$.

Profits are higher from meeting AM than violating AM when $L_0 / C_{21} > 1$, since $\frac{1}{2} \left( 1+L_0 \right)^2 > (\frac{1}{2} \left( 1-L_0 \right))^2$. Now suppose $L_0 < 1$. The principal’s profits are higher from meeting AM than not if and only if $(\frac{1+L_0}{2})^2 - (\frac{1}{2} \left( 1-L_0 \right))^2 \geq \frac{1}{2(1+L_0^2)}$. It is easily shown this constraint is equivalent to: $L_0 \geq 1 - \sqrt{\frac{2A^2}{1+A^2}}$. This completes the proof.

**References**


Gouldner, A. W. 1954. “Patterns of Industrial Bureaucracy.”


Are Politically Connected Firms More Likely to Evade Taxes? Evidence from Tunisia

Bob Rijkers, Hassen Arouri, and Leila Baghdadi

Are politically connected firms more likely to evade taxes? Using tax, social security, and customs records from Tunisia in which firms owned by former president Ben Ali and his family are identified, this paper demonstrates that connected firms were more likely to evade taxes. Ceteris paribus, connected firms are 4.6% more likely not to submit a tax declaration in spite of registering workers and/or customs transactions and 8.4% more likely to report anomalously low sales when submitting a tax declaration. JEL Codes: H26, D22, E26

Are politically connected firms more likely to evade taxation? The jury is out, in spite of the prevalence of political connections (Faccio 2006) and their association with anticompetitive practices (Claessens et al. 2008; Faccio et al. 2006; Mitten and Johnson 2003; Rijkers et al. 2014b). The question is especially relevant for developing countries, since lackluster revenue mobilization (Slemrod and Yitzhaki 2002) and problematic state-business relationships are both symptoms and causes of underdevelopment.

Taking advantage of unique administrative and tax data from Tunisia in which firms with ownership connections to former president Ben Ali confiscated...
in the aftermath of the Jasmin Revolution are identified, this paper answers this question by assessing whether connected firms were more likely to either (i) not declare or (ii) underreport their earnings to the tax authorities.

Tunisia, a small Northern African country at the forefront of the Arab Spring, provides an interesting context to examine this issue. The Ben Ali family had extensive business interests, owning, fully or in part, at least 662 firms, which together accounted for approximately 5% of private sector output. With a revenue to GDP ratio of 21%, its revenue collection performance is not atypical of a country at its stage of development. Findings for Tunisia are thus likely to be relevant for other middle-income and developing countries.

Moreover, Tunisia has a wealth of administrative data, but information sharing between (and within) different government agencies was extremely limited during the Ben Ali era. This in turn facilitates testing for tax evasion through an examination of discrepancies between different administrative databases. More specifically, triangulating social security and customs data with tax data enables us to assess whether firms that reported hiring workers and/or customs transactions submitted a tax declaration and, if so, whether the earnings reported in these declarations are likely to be credible. This simple but powerful strategy of testing for tax evasion can easily be replicated in other countries.

Previewing our main findings, tax evasion is rife in Tunisia; 9% of firms do not submit a tax declaration in spite of reporting hiring workers to the social security administration and/or participating in international trade. Amongst those that submit declarations, 15% report anomalously low sales. Connected firms are among the worst evaders, being ceteris paribus 5% more likely not to declare taxes and 8% more likely to report anomalously low levels of output when submitting tax declarations.

These results contribute to the literature on political connections by highlighting tax evasion as a mechanism by which connected firms reap rents. Estimates of the premium on being connected that do not consider such evasion, which is the norm in the literature, are likely to be downward biased. The findings furthermore add to the literature on the determinants of tax compliance. Our findings also resonate with a companion study in which we demonstrate that connected firms were more likely to evade tariffs by underreporting prices (Rijkers et al. 2015).

I. Hypotheses

Why might politically connected firms be more likely to evade taxes? Models of tax evasion predict it to be a function of rewards, the risk of being caught, and penalties conditional on being caught (see Alm 1999, Andreoni et al. 1998, Slemrod 2007, and Slemrod and Yitzhaki 2002, for reviews of the literature). These factors may vary between connected and nonconnected firms. The Ben Ali family was notorious both for being above the law and for abusing it for personal purposes (Government of Tunisia 2011). Officials might have been more willing to tolerate evasion by connected firms because of career concerns, to ingratiate
themselves with the ruling elite, or because of fear of reprisals. Alternatively, politically connected firms may have been less risk averse because their owners were wealthier or better capable of coping with the consequences of being caught. However, being associated with corrupt behavior could constitute a more severe reputational risk for connected entrepreneurs with political exposure than for nonconnected entrepreneurs. Connected firms seem to have had better access to information and thus might have been better able to anticipate controls and to judge where and when evasion was more likely to remain undetected.

While distinguishing between these different channels is beyond the scope of this note, they form the basis for the null hypothesis that connected firms are more likely to evade taxes. Moreover, one may expect potential differences between connected and nonconnected firms to vary with the strength of connections.

II. Data

Identifying political connections: In the aftermath of the Jasmin revolution, the Tunisian government ordained by presidential decree (Decret-Loi 2011–13) the confiscation of the assets of 114 individuals belonging to the Ben Ali clan, including Ben Ali himself, his relatives, and his in-laws. The list almost exclusively comprises family members, which helps assuage endogeneity concerns, since only a handful of entrepreneurs managed to select into the family through marriage. The confiscation process is still ongoing. Amongst the assets confiscated thus far are boats, yachts, houses, bank accounts, and 662 firms, which we use to identify firms as politically connected. Of these, there were 261 firms for which listed individuals had 100% ownership. These are referred to as “Fully Ben Ali owned.” Another 226 firms were co-owned by the Ben Ali clan and nonlisted individuals. Such firms are referred to as “Partially Ben Ali owned.” For another 175 firms, the exact ownership share of the Ben Ali family could not be determined on the basis of the information provided to us. Such firms are referred to as having “Unknown Ben Ali Ownership shares.” This variation in the extent of Ben Ali ownership is used as a proxy for the strength of political connections.

Administrative data: To assess whether connected firms are more likely to evade taxes, we examine the prevalence of discrepancies between different administrative databases, notably tax records, social security records, and firm-level administrative databases.

1. During interviews Tunisian officials mentioned that speaking up against the Ben Ali family was a precarious proposition. In other countries, officials may have stronger incentives not to tolerate corrupt behavior by politically connected individuals since there is more media scrutiny of such individuals and thus a higher propensity of complicit behavior being detected (Slemrod 2007).

2. www.infojort.com

3. One limitation is that we only identify these firms as being connected in 2011; thus, there might have been some firms that were connected in the period 2006–2009 but that exited before 2011 that we do not identify. We also do not identify firms that benefitted from indirect connections with the president and his family. As such, our results are likely to underestimate the prevalence of political connections.
customs transactions data. This comparison is useful, since interaction and data sharing both between and within different government agencies was minimal—and entrepreneurs were aware of this. Only after the Revolution were we able to assemble this dataset with the help of the Tunisian authorities, the Tunisian Institute National de la Statistique, and the World Bank. In addition, social security and customs data provide useful objective indicators of economic activity and the quality of social security data is high.

The database we assemble covers all registered private sector firms in Tunisia, including the self-employed. We exclude firms in the regime forfaitaire, for whom controls are limited (typically these are very small firms with sales up to 100,000 TND), and focus exclusively on private firms. Firms in the agricultural and hydrocarbon sectors are also excluded since they face different reporting requirements. We focus on the period 2006–2009 since for this period the richest set of firm characteristics is available. Because of these sample restrictions, we have a total of 432 connected firms, of which 192 are fully Ben Ali owned, 147 are partially Ben Ali owned, and 93 have unknown degrees of Ben Ali ownership. Descriptive statistics for the estimation sample are presented in appendix table A1.

III. Nonreporting: Missing Tax Declarations

To start with, we test whether connected firms are less likely to submit tax declarations conditional on being active, that is, reporting (i) hiring workers to the social security administration and/or (ii) having customs declarations on importing and/or exporting. Second, we examine the prevalence of reporting anomalously low levels of sales, defined as declarations in which sales reported to the tax authorities are lower than at least one of the following (i) the wage bill reported to the social security administration (ii) total exports or (iii) total imports recorded in customs transactions data. Entering and exiting firms are excluded from the analysis, since discrepancies in the declarations of such firms may be due to starting up and filing of bankruptcy.

4. Social security data were obtained from the Caisse Nationale de la Sécurité Sociale (CNSS), customs transactions records were obtained from La Direction Générale des Douanes and tax records were obtained from La Direction Générale des Impôts (D.G.I.).

5. Comparing employment rates calculated on the basis of administrative data with employment rates based on labor force surveys suggests informality in the form of nonregistered employment is limited, typically amounting to less than 10% of employment except for in the construction sector (Rijkers et al. 2014a). Social security records are thus a relatively reliable indicator of activity. Similarly, export data also provide a useful benchmark, since firms have limited incentives to underreport exports as these are typically not taxed. Import data are more problematic, since firms have strong incentives to under or misreport imports to evade taxes.

6. Information on import transactions is not available in 2010, while information on foreign ownership shares and tax regime is not available prior to 2006. However, the main results presented in this paper also obtain in 2010 as well as during the period 2001–2006; these results are not presented here in order to conserve space but are available upon request.
Results of OLS regressions are presented in Table 1 in which the dependent variable is an indicator of whether a firm has a missing tax declaration despite being active, that is, recording having hired and/or paid workers to the social security administration and/or recorded customs transactions. The first thing to note is that tax evasion is widespread; 9.0% of nonconnected firms did not submit a tax declaration despite being economically active (see the bottom row).

One might wonder why firms would report hiring workers and/or customs transactions but

### Table 1. Missing Tax Declarations

<table>
<thead>
<tr>
<th>Dependent variable: Tax declaration is missing</th>
<th>Sample: firms which submitted social security and/or customs declarations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ben Ali firm</strong></td>
<td><strong>0.019</strong>  <strong>0.046</strong>*</td>
</tr>
<tr>
<td>Partially Ben Ali owned</td>
<td></td>
</tr>
<tr>
<td>Unknown Ben Ali %</td>
<td></td>
</tr>
<tr>
<td>Fully Ben Ali owned</td>
<td></td>
</tr>
<tr>
<td>Mean dep var non-Ben Ali firms</td>
<td><strong>0.090</strong>  <strong>0.090</strong>  <strong>0.090</strong></td>
</tr>
<tr>
<td>Year (4)</td>
<td>Yes</td>
</tr>
<tr>
<td>Firm controls (6)</td>
<td>Yes</td>
</tr>
<tr>
<td>Governorate (24)</td>
<td>Yes</td>
</tr>
<tr>
<td>Industry (580)</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td><strong>241,839</strong>  <strong>241,839</strong>  <strong>241,839</strong></td>
</tr>
<tr>
<td># Firms</td>
<td><strong>80,199</strong>  <strong>80,199</strong>  <strong>80,199</strong></td>
</tr>
<tr>
<td>Adjusted R2</td>
<td><strong>0.001</strong>  <strong>0.180</strong>  <strong>0.180</strong></td>
</tr>
</tbody>
</table>

*, **, and *** denote results are statistically significant at the 10%, 5%, and 1% level respectively. Standard errors are clustered by firm. The sample comprises all registered private sector incumbent firms that are not in the regime forfaitaire and that recorded hiring works or paying wages to the social security administration and/or recorded customs transactions, with the exception of entrants and exiting firms and firms in the agricultural and hydrocarbon sectors. Firm controls include; dummies for foreign ownership and being in the offshore sector, log age, log(L+1), log(Exports+1), and Log(Imports+1), where missing values are assigned zeros. Sample period: 2006–2009. Estimated using Ordinary Least Squares.

Results of OLS regressions are presented in Table 1 in which the dependent variable is an indicator of whether a firm has a missing tax declaration despite being active, that is, recording having hired and/or paid workers to the social security administration and/or recorded customs transactions. The first thing to note is that tax evasion is widespread; 9.0% of nonconnected firms did not submit a tax declaration despite being economically active (see the bottom row). One might wonder why firms would report hiring workers and/or customs transactions but

7. We present OLS results for ease of interpretation and computational feasibility (we are including a large number of fixed effects).

8. Of course firms have the option of not reporting to any administrative authority, in which case we consider them “inactive.” Of firms contained in the Tunisian National Business Register, the Repertoire National des Entreprises, 23% of firms are inactive, while 14% of connected firms are inactive. The high prevalence of inactive firms is due to a combination of temporary exits, which are strongly concentrated amongst the self-employed (i.e., one-person enterprises), evasion, and the presence of zombie firms; firms that are no longer economically active but have not legally ceased to exit. Since Ben Ali firms tend to be larger, it is perhaps not surprising that they are less likely to be inactive (in addition, identification of connections occurs in 2011, while our sample starts in 2006).
not submit a tax declaration. A possible explanation is that being registered with the social security administration entitles employees to various benefits, including health insurance, while engaging in international trade without going through customs is arguably difficult (though smuggling does happen).

Second, such evasion is even more widespread among active connected firms. The estimates presented in column 1, which control for being connected and year dummies only, suggest that connected firms are 1.9% more likely than nonconnected firms not to submit a declaration, but the differential with nonconnected firms is not statistically significant. Once their industry, location, and superior size and trade performance are controlled for, as is done in column 2, the evasion premium on being connected rises to 4.6% percentage points and becomes significant at the 1% level.

To assess whether the evasion propensity varies with the strength of political connections, columns 3 replicates the second specification 2, but now separates Ben Ali firms into those that are (i) known to be owned only in part by the Ben Ali family (ii) firms which are fully Ben Ali owned and (iii) a residual group for which the exact Ben Ali ownership share is not known. While coefficient estimates are positive for all three groups, they are only significant for the group of fully Ben Ali owned firms, for whom the connectedness premium is highest; ceteris paribus they are 6.4% more likely not to submit any tax declarations than nonconnected firms.

IV. UNDERREPORTING

We now turn to examining the prevalence of underreporting amongst those firms that submitted tax declarations and reported hiring workers and/or customs transactions, by examining the incidence of anomalously low levels of sales defined here as sales that are lower than either the wage bill, total imports, or total exports. 9

Table 2 presents results using specifications that mirror those presented in table 1 but using as dependent variable having submitted an anomalous declaration. Such anomalous declarations are prevalent; no less than 15.3% of nonconnected firms submit a tax declaration in which they claim their sales were lower than their wagebill, total exports, or total imports. Amongst Ben Ali firms submitting tax declarations, this percentage is an additional 6.6% higher (column 1), and the differential with nonconnected firms is statistically significant. Controlling for other firm characteristics, such as industry, location, age, size, exports, and imports, connected firms are 8.4% more likely to submit an anomalous declaration (column 2). Differentiating by type of connection, as is done in column 3, demonstrates that

9. There are a number of firms for which reported sales are 0. We exclude these from the analysis of underreporting since we are concerned these might reflect temporary cessation of economic activity and/or data entry errors. Our results, which are omitted to conserve space but available upon request, become even stronger when we include them.
fully Ben Ali owned firms are most likely to submit anomalous declarations – in
despite being least likely to submit tax declarations in the first place.

Table A2 in the appendix presents robustness checks that demonstrate that the results also obtain when focusing on profit making firms only (columns 1 and 2). Moreover, they are robust to using more stringent definitions of anomalous declarations, which allow for lags between the production and sales processes, notably (ii) defining a declaration to be anomalous if cumulative output over a two year period is lower than either the wagebill, exports, or imports in the base year (columns 3 and 4) and (iii) average output between 2006 and 2009 is lower than either the average wagebill, average imports, or average exports over the same period (columns 5 and 6). Using either of these alternative measures, connected firms are consistently significantly more likely to report anomalously low levels of sales.
V. Conclusion

By examining discrepancies between administrative data from Tunisia that were curated without coordination between different government agencies, this paper unveils widespread tax evasion; 9% of firms do not submit a tax declaration in spite of registering workers, and 15% report earnings which are anomalously low compared to declarations made to other government agencies. Firms owned by the Ben Ali family were among the worst perpetrators, being ceteris paribus both 5% less likely to submit tax declarations and 8% more likely to underreport earnings to the tax authorities. Fully Ben Ali owned firms have the highest evasion propensity.

Overall, our findings not only attest to tax evasion being a mechanism by which connected entrepreneurs reap rents (see also the companion paper, Rijkers et al. 2015, which demonstrates that connected firms were more likely to evade tariffs) but also suggest that estimates of the impact of political connections on firm performance that do not consider such evasion are likely biased downwards. While the methods used in this paper are simple, they are powerful and can easily be replicated in other countries where information sharing across institutions is imperfect.

Appendix Tables

Table A1. Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>Connected firms</th>
<th>Nonconnected firms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Std.Dev.</td>
<td>Obs</td>
</tr>
<tr>
<td>Missing tax declaration</td>
<td>0.109 0.312</td>
<td>813</td>
</tr>
<tr>
<td>Anomalous declaration*</td>
<td>0.219 0.414</td>
<td>718</td>
</tr>
<tr>
<td>Foreign</td>
<td>0.034 0.182</td>
<td>813</td>
</tr>
<tr>
<td>Offshore</td>
<td>0.100 0.300</td>
<td>813</td>
</tr>
<tr>
<td>Log age</td>
<td>2.130 0.736</td>
<td>813</td>
</tr>
<tr>
<td>Log (L+1)</td>
<td>2.268 1.789</td>
<td>813</td>
</tr>
<tr>
<td>Log (Exports+1)</td>
<td>2.577 5.087</td>
<td>813</td>
</tr>
<tr>
<td>Log (Imports+1)</td>
<td>6.322 6.796</td>
<td>813</td>
</tr>
</tbody>
</table>

Note: The sample comprises all registered private sector incumbent firms reporting positive turnover that are not in the regime forfaitaire and that recorded hiring workers and/or paying wages to the social security administration and/or recorded customs transactions, with the exception firms in the agricultural and hydrocarbon sectors; entrants and exiting firms are not included. Sample period: 2006–2009.

* A tax declaration is considered anomalous if sales reported to the tax authorities are strictly lower than the wage bill reported to the social security administration and/or imports and/or exports recorded by customs. Firms that report their sales to be 0 are not included in our analysis of anomalous declarations.
## Table A2. Alternative Definitions of Anomalously Low Output

Underreporting—Alternative definitions

<table>
<thead>
<tr>
<th>Dependent variable: Reported sales are anomalously low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only firms reporting positive profits</td>
</tr>
<tr>
<td>Definition of anomalously low output</td>
</tr>
<tr>
<td>Sales &lt; Wagebill and/or Sales &lt; Exports and/or Sales &lt; Imports</td>
</tr>
<tr>
<td>Sample restriction</td>
</tr>
<tr>
<td>Only firms reporting positive output</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>Ben Ali firm</td>
</tr>
<tr>
<td>(0.025)</td>
</tr>
<tr>
<td>Partially Ben Ali owned</td>
</tr>
<tr>
<td>(0.040)</td>
</tr>
<tr>
<td>Unknown Ben Ali %</td>
</tr>
<tr>
<td>(0.046)</td>
</tr>
<tr>
<td>Fully Ben Ali owned</td>
</tr>
<tr>
<td>(0.043)</td>
</tr>
<tr>
<td>Mean dep var non-Ben Ali firms</td>
</tr>
<tr>
<td>Year (4)</td>
</tr>
<tr>
<td>Firm controls</td>
</tr>
<tr>
<td>Governorate (24)</td>
</tr>
<tr>
<td>Industry (580)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td># firms</td>
</tr>
<tr>
<td>Adjusted R2</td>
</tr>
</tbody>
</table>

*, **, and *** denote results are statistically significant at the 10%, 5%, and 1% level respectively. Standard errors are clustered by firm. Subject to the specific sample restrictions listed in the table heading above, the sample comprises all registered private sector incumbent firms reporting positive turnover that are not in the regime forfaitaire and that recorded hiring workers and/or paying wages to the social security administration and/or recorded customs transactions, with the exception firms in the agricultural and hydrocarbon sectors; entrants and exiting firms are not included. Firm controls include: dummies for foreign ownership and being in the offshore sector, log age, log(L + 1), log(Exports + 1), and Log(Imports + 1), where missing values are assigned zeros. Sample period: 2006–2009. Estimated using Ordinary Least Squares.
REFERENCES


Courting Economic Development

James R. Brown, J. Anthony Cookson, and Rawley Z. Heimer

We show that court enforcement uncertainty hinders economic development using sharp variation in judiciaries across Native American reservations in the United States. Congressional legislation passed in 1953 assigned state courts the authority to resolve civil disputes on a subset of reservations, while tribal courts retained authority on unaffected reservations. Although affected and unaffected reservations had similar economic conditions when the law passed, reservations under state courts experienced significantly greater long-run growth. When we examine the distribution of incomes across reservations, the average difference in development is due to the lower incomes of the most impoverished reservations with tribal courts. We show that the relative underdevelopment of reservations with tribal courts is driven by reservations with the most uncertainty in court enforcement. JEL codes: K10, N22, O10

What causes economic growth? Despite centuries of inquiry and decades of empirical research, economists continue to wrestle with this fundamental question. In particular, while it is clear in a general sense that “institutions” are important for development (e.g., North 1990), identifying the specific mechanisms through which institutions encourage growth is a significant challenge (Sala-i-Martin et al. 2004).

One potentially important institutional determinant of economic performance is the legal system, as countries differ sharply in the extent to which the courts enforce contracts, resolve property disputes, and protect individuals from expropriation by the government. For example, a prominent literature argues that countries with legal origins in civil (rather than common) law have higher levels of administrative procedures in their courts—what Djankov et al. (2003) call “legal formalism.” These countries have “heavier regulation, less secure property rights, more corrupt and less efficient governments, and even less political freedom than do the common law countries” (Glaeser and Shleifer 2002, 1194).
Yet scholars remain skeptical that these studies convincingly show legal institutions or court systems cause economic growth. The primary criticism is that statistical measures of the quality of legal institutions simply proxy for other factors—such as culture, political differences, or historical events—which also have persistent effects on economic development (Glaeser and Shleifer 2002). For example, even La Porta et al. (2008) note that “connecting legal origins to aggregate economic growth . . . has proved difficult.”

This paper addresses this criticism by studying how legal institutions affect economic development across Native American reservations in the United States. The reservation setting avoids many of the confounding explanations from the cross-national setting because—in comparison to countries—reservations are similar geographically, have similar opportunities to trade with the rest of the United States, and have a similar history, particularly with respect to their relation to the US government. In addition, the Native American experience is informative because court systems vary sharply across reservations, in part due to the (external) actions of the US Congress. One such action—Public Law 280 (PL280), passed by Congress in 1953—transferred jurisdiction over civil and criminal proceedings on a subset of reservations to state courts (US Congress 1953). On reservations not affected by PL280, these proceedings are handled by tribal courts. As prior work shows, state court jurisdiction is associated with significantly higher reservation income.

We depart from previous work on PL280 by examining heterogeneity in development outcomes among reservations with tribal courts. We find that many reservations with autonomous tribal courts are just as well off as reservations under state court jurisdiction. Yet the income distribution for tribal court reservations has a much longer lower tail, which means that these reservations are more likely to experience extreme poverty. In addition, we show that this lower tail of outcomes can explain the entire average difference in development outcomes across reservations with state and tribal courts.

Next, we study what contributes to the increased heterogeneity in economic development under autonomous tribal courts. In particular, we evaluate how uncertainty over legal enforcement influences the distribution of incomes across reservations. In our tests, we use the tribal court’s civil caseload to proxy for enforcement uncertainty. This measure is theoretically founded because, when

---

1. Public Law 280 stipulated that state courts have concurrent jurisdiction with tribal courts on affected reservations. On unaffected reservations, tribal courts have exclusive jurisdiction over civil and criminal proceedings because state courts have no authority. Thus, the distinctive feature of a non-PL280 reservation is the absence of state courts, not the presence of tribal courts. In practice, PL280 transferred court activity almost entirely to state courts (see table 2b). For this reason and because the exposition is simpler, we will sometimes refer to PL280 reservations as “state court reservations” and non-PL280 reservations as “tribal court reservations.”

2. For example, Brown et al. (2016) find that income per capita is 7–14% greater on reservations with state courts, after differencing out regional differences in economic development. Also see the evidence in Anderson and Parker (2016).
legal enforcement is well understood (low enforcement uncertainty), disputing
parties are more likely to settle their differences without using the court, resulting
in lower caseload.\textsuperscript{3}

Consistent with a mechanism whereby legal enforcement uncertainty hinders
economic growth, we find that reservations with the busiest (most active) tribal
courts have significantly lower per capita income, and that this effect can explain
the entire mean income difference between state-court and tribal-court reserva-
tions. In fact, using a Komologorov-Smirnov test, we find no difference between
the entire distribution of tribal court reservations with low enforcement uncer-
tainty and state court reservations. In this way, our study points to an important
mechanism for understanding the legal foundations of development and for im-
proving tribal courts from within: Reducing enforcement uncertainty in the
courts can facilitate development and may go a long way to alleviating the persist-
tent underdevelopment of reservations in the United States.

I. Native American Institutions and Development

Native American tribes are nations with limited sovereignty. Their governmental
rule is subordinate to the US federal government, but with few exceptions, not
subject to state government decisions. A set of three Supreme Court decisions
between 1823 and 1832—called the Marshall Trilogy after Chief Justice John
Marshall—established the guiding legal principle that tribes’ relationship to the
federal government “resembles that of a ward to its guardian (\textit{Cherokee Nation
v. Georgia} 1832).” Under this legal framework, a long history of federal legislative
interventions has generated a common set of experiences among the many Native
American tribes. Although the federal government’s attitude toward tribes varied
between assimilation and self-determination during this period, most federal legis-
lation tended to make reservations more similar, and this is true even for pro-
sovereignty measures. For example, although the Indian Reorganization Act of
1934 was a measure to enhance tribal sovereignty and self determination, the law
created a template constitutional form that most reservations followed.

In contrast to most interventions in reservation policy, one of the most
 crucial pieces of legislative intervention on reservations created stark differenc-
es in legal jurisdiction across tribes. Passed by Congress in 1953, Public Law
280 (PL 280) transferred authority to state courts over civil and criminal pro-
cedings on most reservations in Alaska, California, Minnesota, Nebraska,
Oregon, and Wisconsin. State courts would later have jurisdiction in Florida,
Iowa, and Washington, while New York gained jurisdiction three years prior to
PL280. In passing PL280, Congress sought to transfer the burden of federal

\textsuperscript{3} This proxy is a natural outgrowth of the law and economics literature on transaction costs and
property rights. In fact, the notion that cases will be less likely to be brought to the court when
enforcement is clear is an old idea that dates back to at least the original exposition of “The Coase
Theorem” in \textit{Coase (1960)}. In our context, enforcement uncertainty is greater when there are high
transaction costs and improperly specified property rights.
court jurisdiction to the states, particularly in regions with greater perceived lawlessness. The transfer of authority to state courts for criminal jurisdiction was the primary consideration behind the law, but “[C]ivil jurisdiction was an afterthought in a measure aimed primarily at bringing law and order to the reservations, added because . . . it was convenient and cheap” (Goldberg-Ambrose 1997, 50). Table 1 summarizes how PL280 affected the court systems of the largest reservations.

### Table 1. Civil Jurisdiction by State Courts and Tribal Courts

<table>
<thead>
<tr>
<th>State</th>
<th>Civil jurisdiction of contracts</th>
<th>Exemptions or retrocessions</th>
<th># Reservations (tribal courts, state courts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>State courts, mandatory state</td>
<td></td>
<td>(0, 0)</td>
</tr>
<tr>
<td>Arizona</td>
<td>Tribal courts, except for pollution</td>
<td></td>
<td>(9, 0)</td>
</tr>
<tr>
<td>California</td>
<td>State Courts, mandatory state</td>
<td></td>
<td>(0, 7)</td>
</tr>
<tr>
<td>Colorado</td>
<td>Tribal courts</td>
<td></td>
<td>(2, 0)</td>
</tr>
<tr>
<td>Florida</td>
<td>State courts, optional state</td>
<td></td>
<td>(0, 2)</td>
</tr>
<tr>
<td>Iowa</td>
<td>State courts, optional state</td>
<td></td>
<td>(0, 1)</td>
</tr>
<tr>
<td>Idaho</td>
<td>Tribal courts</td>
<td></td>
<td>(3, 0)</td>
</tr>
<tr>
<td>Kansas</td>
<td>Tribal courts</td>
<td></td>
<td>(2, 0)</td>
</tr>
<tr>
<td>Maine</td>
<td>Tribal courts</td>
<td></td>
<td>(2, 0)</td>
</tr>
<tr>
<td>Michigan</td>
<td>Tribal courts</td>
<td></td>
<td>(4, 0)</td>
</tr>
<tr>
<td>Minnesota</td>
<td>State courts, mandatory state</td>
<td>Red Lake, Boise</td>
<td>(2, 5)</td>
</tr>
<tr>
<td>Mississippi</td>
<td>Tribal courts</td>
<td></td>
<td>(1, 0)</td>
</tr>
<tr>
<td>Montana</td>
<td>Tribal courts, PL280 conflicts with state constitution</td>
<td></td>
<td>(7, 0)</td>
</tr>
<tr>
<td>North Carolina</td>
<td>Tribal courts</td>
<td></td>
<td>(1, 0)</td>
</tr>
<tr>
<td>North Dakota</td>
<td>Tribal courts, PL280 conflicts with state constitution</td>
<td></td>
<td>(2, 0)</td>
</tr>
<tr>
<td>Nebraska</td>
<td>State courts, mandatory state</td>
<td>Winnebago, Omaha</td>
<td>(1, 1)</td>
</tr>
<tr>
<td>New Mexico</td>
<td>Tribal courts</td>
<td></td>
<td>(10, 0)</td>
</tr>
<tr>
<td>Nevada</td>
<td>Tribal courts</td>
<td></td>
<td>(4, 0)</td>
</tr>
<tr>
<td>New York</td>
<td>State courts under public law 785 in 1950</td>
<td></td>
<td>(0, 4)</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>Tribal Courts</td>
<td></td>
<td>(1, 0)</td>
</tr>
<tr>
<td>Oregon</td>
<td>Tribal Courts, mandatory state but retroceded for all</td>
<td>Warm Springs, Umatilla</td>
<td>(2, 0)</td>
</tr>
<tr>
<td>South Carolina</td>
<td>Tribal courts</td>
<td></td>
<td>(1, 0)</td>
</tr>
<tr>
<td>South Dakota</td>
<td>Tribal courts, except for highways</td>
<td></td>
<td>(8, 0)</td>
</tr>
<tr>
<td>Utah</td>
<td>Tribal courts</td>
<td></td>
<td>(1, 0)</td>
</tr>
<tr>
<td>Washington</td>
<td>Tribal courts</td>
<td>Port Madison, Quinault</td>
<td>(13, 0)</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>State courts, mandatory state</td>
<td>Menominee</td>
<td>(1, 7)</td>
</tr>
<tr>
<td>Wyoming</td>
<td>Tribal courts</td>
<td></td>
<td>(1, 0)</td>
</tr>
</tbody>
</table>

**Note:** This table presents the nature of the variation in state versus tribal court jurisdiction in our sample, the names of exempted or retroced reservations, and detail on the nature of civil jurisdiction over contracts through PL280 or other regulations. Details from this table were taken from Cookson (2010), which assembled information from an assortment of legal resources (Johnson and Paschal 1992; Getches et al. 1998; Jimenez and Song 1998; Melton and Gardiner 2006).
Two aspects of court assignment under PL280 are important for our tests. First, reservations were not systematically assigned to state courts based on their potential for economic development. Parker (2012) and Brown et al. (2016) report that PL280 and non-PL280 reservations had similar levels of credit market activity, human capital, and per capita income around the time PL280 was passed. Rather, most commentators argue that the objective of PL280 was to ultimately assimilate Native Americans into the broader culture. Second, a key difference between state and tribal courts is that state courts appear to provide more predictable (less uncertain) contract enforcement (e.g., Mudd 1972; Anderson and Parker 2008). As a consequence, reservations with state courts see significantly more financial and long-term contracting (e.g., Parker 2012; Cookson 2014; Brown et al. 2016).

II. Results

PL280 and Heterogeneity in Tribal Courts

Although reservations were similar at the time PL280 was passed, there are now stark differences between reservations subject to PL280 and those that are not (table 2). Consistent with prior work on PL280, we verify that incomes are higher under state court jurisdiction by approximately $2000, both for the mean and median reservation (Anderson and Parker 2008; Cookson 2014; Brown et al. 2016). Indeed, a two-sample $t$-test indicates that this difference is statistically significant at the 5 percent level ($t = 2.14$).

The difference in average incomes across jurisdictions can be attributed to the substantial heterogeneity among non-PL280 reservations, and in particular, the longer lower tail of outcomes. Table 2a highlights the greater heterogeneity among non-PL280 reservations, which have both a greater standard deviation in per capita incomes (4523.2 > 3593.0) and a greater interquartile range (6618 > 4156) than PL280 reservations. Examining the distribution more directly, figure 1 shows the greater heterogeneity is due to a longer lower tail in the tribal court income distribution. Although the income distribution for PL280 reservations first order stochastically dominates the income distribution for non-PL280 reservations, there is little difference between incomes at the 75th percentile or above.

Turning to differences in court activity, data from the 1985 National American Indian Court Judges Association (NAICJA 1985) provides evidence that PL280 had a pronounced effect on caseload and court personnel across legal jurisdictions.
There is remarkably little activity in tribal courts on PL280 reservations (an average of 1.5 civil cases and zero criminal cases per 10,000 residents) and almost no judges or court personnel are assigned to these tribal courts. Tribal courts on non-PL280 reservations are much busier. On average, they witness over 50 civil cases and over 200 criminal cases per 10,000 residents. Autonomous tribal courts average over one judge and four court workers per 10,000 residents.

The difference in tribal court activity under PL280 confirms the intuition of prior work that the law meaningfully transferred jurisdiction over disputes to state courts (Anderson and Parker 2008). The transfer of jurisdiction from tribal courts to state courts is natural because state courts are better equipped to handle more complicated disputes and also offer clearer enforcement of contracts with appeal to relevant precedent (Parker 2012).

6. Even though tribal courts had an opportunity to hear cases when sharing concurrent jurisdiction with state courts under PL280, they did not.
Active Tribal Courts and Enforcement Uncertainty

We now use variation in tribal court characteristics to explain the heterogeneity in incomes within the set of reservations unaffected by PL280. Our results help explain why differences in legal jurisdiction have had pronounced effects on development of Native American reservations.

We show that the degree of uncertainty over the enforcement of civil disputes explains much of the variation in incomes across non-PL280 reservations. We investigate this enforcement uncertainty mechanism using the number of civil cases per capita as a proxy for the amount of enforcement uncertainty. This measure is motivated by Coasean bargaining, which has been a longstanding feature of the law and economics literature (Coase 1960). When the outcome of a potential case is more certain, it is more likely to be settled out of court, reducing aggregate caseload.

In support of the hypothesis that greater enforcement uncertainty hinders economic development, figure 2 indicates a strong negative relationship between the number of civil cases per capita and per capita income among reservations with autonomous tribal courts. A line of best fit has a slope of $-27.4$, which implies an increase of $62.1$ cases (approximately one standard deviation) is associated with a $1700$ decrease in per capita income.\(^7\) Furthermore, the R-squared in the

\[ \% \Delta = \exp(\beta) - 1. \]

---

**Notes:** These empirical cumulative distribution functions (CDFs) present the cross-reservation distributions of year 2000 per capita income by jurisdiction type (state court jurisdiction under PL280, and tribal court only jurisdiction for non-PL280 reservations). The one-sided Komolgorov-Smirnov p-value equals .0617, which indicates that the distribution of income for reservations with autonomous tribal courts is shifted to the left relative to the distribution of income for reservations subject to state courts under Public Law 280.

---

**Figure 1.** The Distribution of Per Capita Income Across Reservations by Jurisdiction Type: State Courts Versus Tribal Courts

\[ \text{Per Capita Income in Year 2000} \]

\[ \text{Quantile of Income Distribution} \]

- state courts
- tribal courts

---

\(^7\) This is calculated using the formula for the log-linear form given in Wooldridge (2003): \( \% \Delta = \exp(\beta) - 1. \)
associated simple linear regression is 0.141, which indicates that civil caseload alone explains a sizable fraction of the variation in reservation incomes.

Differences in enforcement uncertainty help us understand why PL280 had polarized effects on reservation development. Panel (a) of table 3 separates non-PL280 tribes into active and inactive tribal courts (above or below the median of civil cases per capita, respectively) and compares these outcomes to PL280 reservations. After accounting for uncertainty in contract enforcement, there is no difference in average per capita income between reservations under state-court jurisdiction and reservations with autonomous tribal courts. Both reservations under state courts and reservations with inactive tribal courts have average per capita incomes equal to roughly $23,500, and the difference in their incomes is not statistically different from zero ($t = –0.167$). In contrast to inactive tribal courts, reservations with active tribal courts have dramatically lower incomes than reservations subject to state court jurisdiction (approximately $3000, t = 3.14$).  

Graphical evidence strengthens our argument that court enforcement uncertainty explains the divergence in reservation development, not just on average but across the entire distribution. In figure 3, we separately plot the distribution of incomes for reservations with state courts, reservations with inactive tribal courts, and reservations with active tribal courts. Not only are the means and medians statistically indistinguishable between state courts and inactive tribal courts, but the distributions are virtually identical. Using a Kolmogorov-Smirnov

8. Moreover, active tribal courts have more judges and more personnel than inactive courts (table 3b). This pattern suggests that the greater enforcement uncertainty does not arise from under-staffing of the court, but enforcement rules and clear adherence to precedent, which have been mechanisms discussed in connection to the gap in outcomes between PL280 and non-PL280 reservations (Parker 2012).
test, we are unable to reject the null hypothesis that incomes from PL280 reservations and incomes from inactive tribal courts are drawn from the same distribution ($p = .928$). In contrast, incomes on reservations with active tribal courts are less than the other reservation types at all points along the distribution. The downward shift in the distribution is statistically significant. A Kolmogorov-Smirnov test for the difference in distributions between active tribal court incomes and PL280 (inactive tribal courts) incomes yields a $p$-value of .030 (.034).$^9$

Regression analysis supports our interpretation of these findings by showing that they are not due to scaling, outliers, or population differences. In particular, table 4 reports the results from OLS regressions of logged per capita income on

<table>
<thead>
<tr>
<th>Table 3: Differences Between Public Law 280 Reservations and Reservations with Active and Inactive Tribal Courts</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Income and Demographic Characteristics in 2000</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Number of reservations</td>
</tr>
<tr>
<td>Year 2000 per capita personal income in headquarters county</td>
</tr>
<tr>
<td>. . . Mean ($)</td>
</tr>
<tr>
<td>. . . Median ($)</td>
</tr>
<tr>
<td>. . . Standard deviation ($)</td>
</tr>
<tr>
<td>. . . Interquartile range ($)</td>
</tr>
<tr>
<td>Logged year 2000 population</td>
</tr>
<tr>
<td>. . . Mean</td>
</tr>
<tr>
<td>. . . Median</td>
</tr>
<tr>
<td>. . . Standard deviation</td>
</tr>
<tr>
<td>. . . Interquartile range</td>
</tr>
<tr>
<td>(b) Tribal Court Activity in 1985</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>. . . Civil cases (per 10,000 residents)</td>
</tr>
<tr>
<td>. . . Criminal cases (per 10,000 residents)</td>
</tr>
<tr>
<td>. . . Tribal court judges (per 10,000 residents)</td>
</tr>
<tr>
<td>. . . Tribal court personnel (per 10,000 residents)</td>
</tr>
</tbody>
</table>

Note: This table presents summary statistics on development and court activity by whether the reservation is subject to concurrent jurisdiction by state courts under Public Law 280, or if the reservation has a fully autonomous tribal court (separately split out by active tribal courts and inactive tribal courts; civil cases per capita greater / below the median). Per capita personal income and population come from the Bureau of Economic Analysis regional information system tables, while the information on tribal court activity comes from a 1985 survey of tribal court staffing and activity (NAICJA 1985).

9. Jointly, the distribution of income for non-PL280 reservations with active tribal courts is statistically different other reservations (pooling PL280 reservations with inactive tribal courts on non-PL280 reservations) with a Komogorov-Smirnov test $p$-value of 0.009.
Figure 3. The Distribution of Per Capita Income Across Reservations by Jurisdiction Type: State Courts Versus Tribal Courts

Notes: These empirical cumulative distribution functions (CDFs) present the cross-reservation distributions of year 2000 per capita income by jurisdiction type, similar to figure 1 except that active tribal courts (i.e., those hearing greater than the median of civil cases per capita in 1985) are split into a separate category from non-active tribal courts. Kolmogorov-Smirnov tests indicate that active tribal courts are shifted to the left relative to state courts ($p = .030$) as well as inactive tribal courts ($p = .034$), but that there is no statistical difference between the income distributions of reservations with state courts in comparison to reservations with inactive tribal courts.

Table 4. Legal Enforcement and Per Capita Income in Year 2000

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>State court indicator</td>
<td>0.093***</td>
<td>0.073*</td>
<td>0.145***</td>
<td>0.110***</td>
<td>0.145***</td>
<td>0.110***</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(0.040)</td>
<td>(0.046)</td>
<td>(0.042)</td>
<td>(0.046)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>Tribal court indicator (not active)</td>
<td>0.149***</td>
<td>0.102**</td>
<td>0.080***</td>
<td>0.055***</td>
<td>0.080***</td>
<td>0.055***</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
<td>(0.042)</td>
<td>(0.023)</td>
<td>(0.021)</td>
<td>(0.023)</td>
<td>(0.021)</td>
</tr>
</tbody>
</table>

Sample Control for population | Full | Full | Non-PL280 | Non-PL280 | Full | Full |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>0.039</td>
<td>0.265</td>
<td>0.136</td>
<td>0.352</td>
<td>0.127</td>
<td>0.304</td>
</tr>
<tr>
<td>N</td>
<td>105</td>
<td>105</td>
<td>78</td>
<td>78</td>
<td>105</td>
<td>105</td>
</tr>
</tbody>
</table>

Note: This table presents OLS results for the specification

$$\log(\text{income}_{\text{per}_\text{capita}}_i) = \beta_1 \text{stjur}_i + \beta_2 \text{court_activity}_i + \beta_3 \log(\text{population}_i) + \epsilon_i$$

where each observation is a reservation headquarters county, $\text{stjur}_i$ equals one if the reservation is under PL280 state jurisdiction, and zero otherwise, and $\text{court_activity}_i$ is a reservation-specific measure of court activity taken from a 1985 survey of tribal courts. The dependent variable is the log of $\text{income}_{\text{capita}}_i$, which is per capita income in the county in year 2000. Variables denoted with a (Z) are standardized to have a mean of zero and a standard deviation of 1 for ease of interpretation. ***, **, and * indicate statistical significance at the one, five, and ten percent levels.
an indicator for state jurisdiction through PL280, measures of tribal court activity, and in some specifications, logged population. First, we verify that reservations under PL280 average 7 to 9 percent higher incomes than non-PL280 reservations, and that the difference in income is not due to population differences (columns 1 and 2). Second, we show that greater activity in tribal courts is associated with lower levels of development among non-PL280 reservations. A standard deviation increase in tribal court caseload per capita leads to a 5.3 to 7.7 percent decline in per capita income (columns 3 and 4). Consistent with our graphical evidence, the magnitude of this decline is just as large as the effect of PL280 jurisdiction on incomes, even after controlling for population differences. Finally, we combine aspects of these tests to compare incomes across reservations with state courts, inactive tribal courts, and active tribal courts. In specifications that include indicator variables for state courts and inactive tribal courts, we find significantly higher incomes (approximately 10 to 15 percent) relative to reservations with active tribal courts. Consistent with our summary statistics and graphical evidence, the difference between state jurisdiction and tribal jurisdiction under an inactive tribal court is negligible (columns 5 and 6). Indeed, as we showed graphically, active tribal courts can explain the entire gap between PL280 and non-PL280 reservations.

III. DISCUSSION

We show that judicial uncertainty is an important constraint on economic development. Our evidence comes from studying the distribution of long run development outcomes in the wake of congressional legislation, Public Law 280, which gave states judicial authority on a subset of Native American reservations. Previous studies have documented higher per capita incomes on reservations subject to PL280. Our new findings show that the difference in averages across legal jurisdiction is due to a longer lower tail of economic activity among non-PL280 reservations. The dispersion in outcomes among non-PL280 reservations is explained by variation in court enforcement uncertainty, providing novel insight into how legal institutions affect growth.

Our findings point to an important question: What aspects of the real economy are affected by judicial uncertainty? Although we expect future research will clarify the specific channels, some have already begun to investigate these mechanisms. Among the most promising angles, uncertain contract enforcement can discourage sunk cost investment and stymie credit market development (Cookson 2014; Brown et al. 2016). Indeed, there is emerging evidence that greater credit provision arising from better contracting regimes can lead to better economic performance (Ponticelli 2013). This conjecture is supported anecdotally, “[Reservations] have a lack of access to capital that has stunted economic growth” (Chairman of the Senate Committee on Reservation Unemployment, Byron Dorgan [2010]). Further understanding these channels will refine efforts to ignite development on reservations and beyond.
REFERENCES


US Senate. 2010. “Unemployment on Indian Reservations at 50 percent: The Urgent Need to Create Jobs in Indian Country.” *Hearings before the Committee on Indian Affairs United States Senate*.
